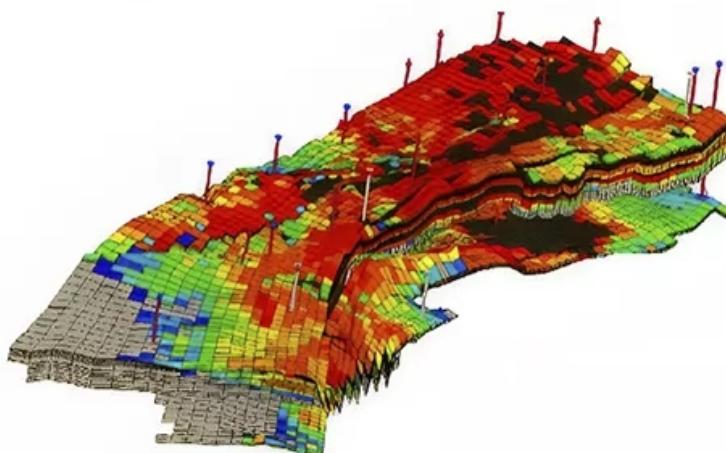
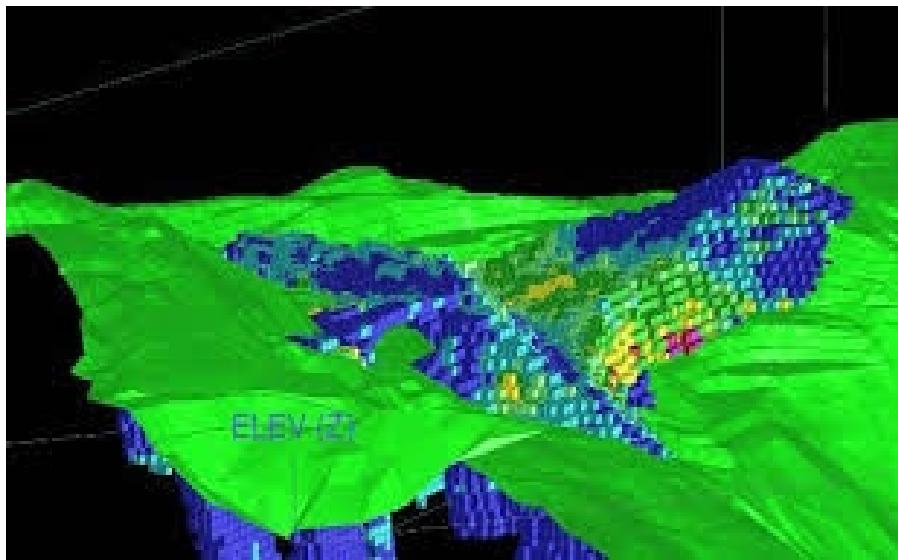


Grado en Inteligencia Artificial

Geoestadística

Prof. Dante Conti
Prof. Sergi Ramirez

Introducción



Aplicaciones:

Geología y ciencias de la tierra:

Estimar la distribución de recursos naturales, como minerales o petróleo.

Ciencias ambientales: Modelar la distribución de contaminantes o la dinámica de ecosistemas.

Ecología: Estudiar la distribución de especies y los patrones de hábitat.

Geomarketing: Analizar la distribución de clientes o mercados potenciales.

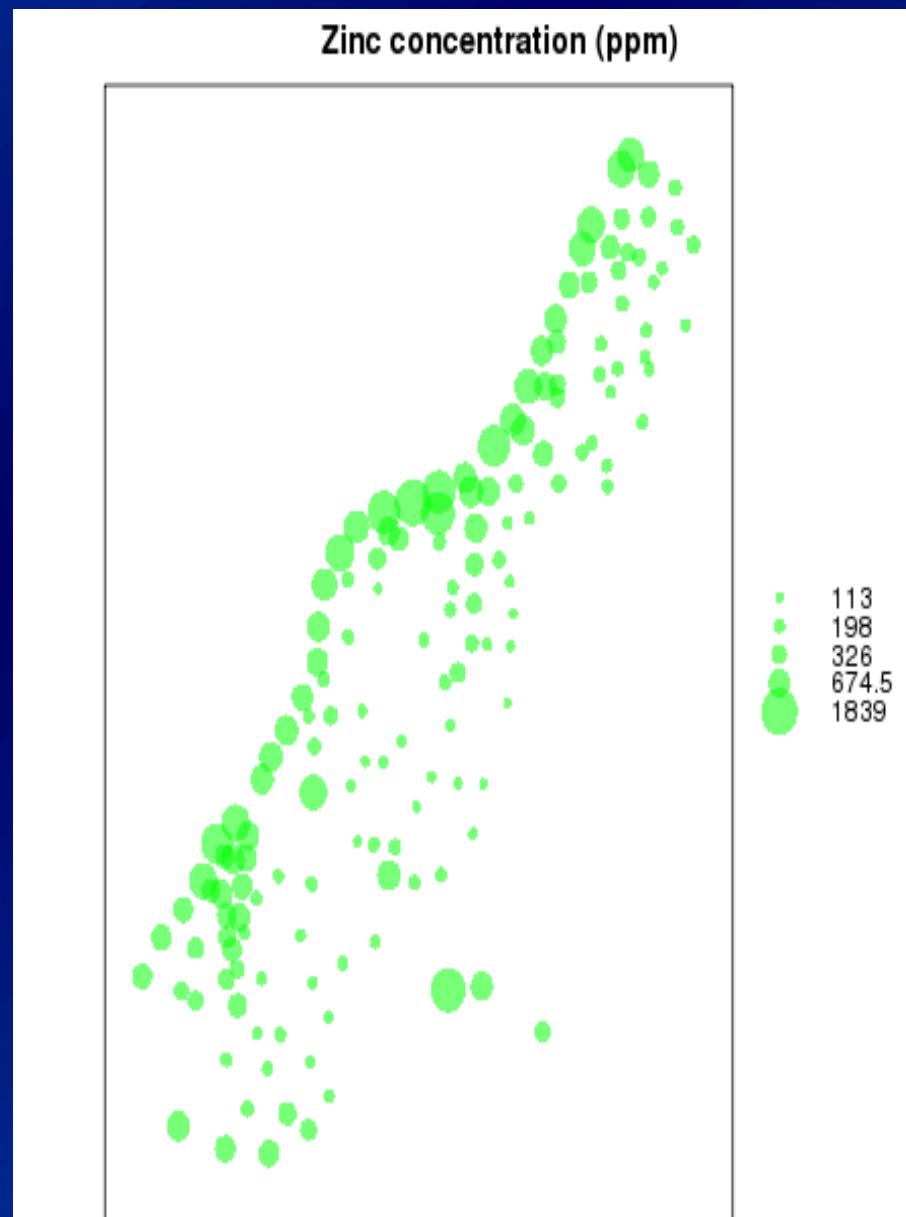
Ciencias de la salud: Modelar la distribución de enfermedades o factores de riesgo.

What is spatial statistics?

Geostatistics
Spatial analysis

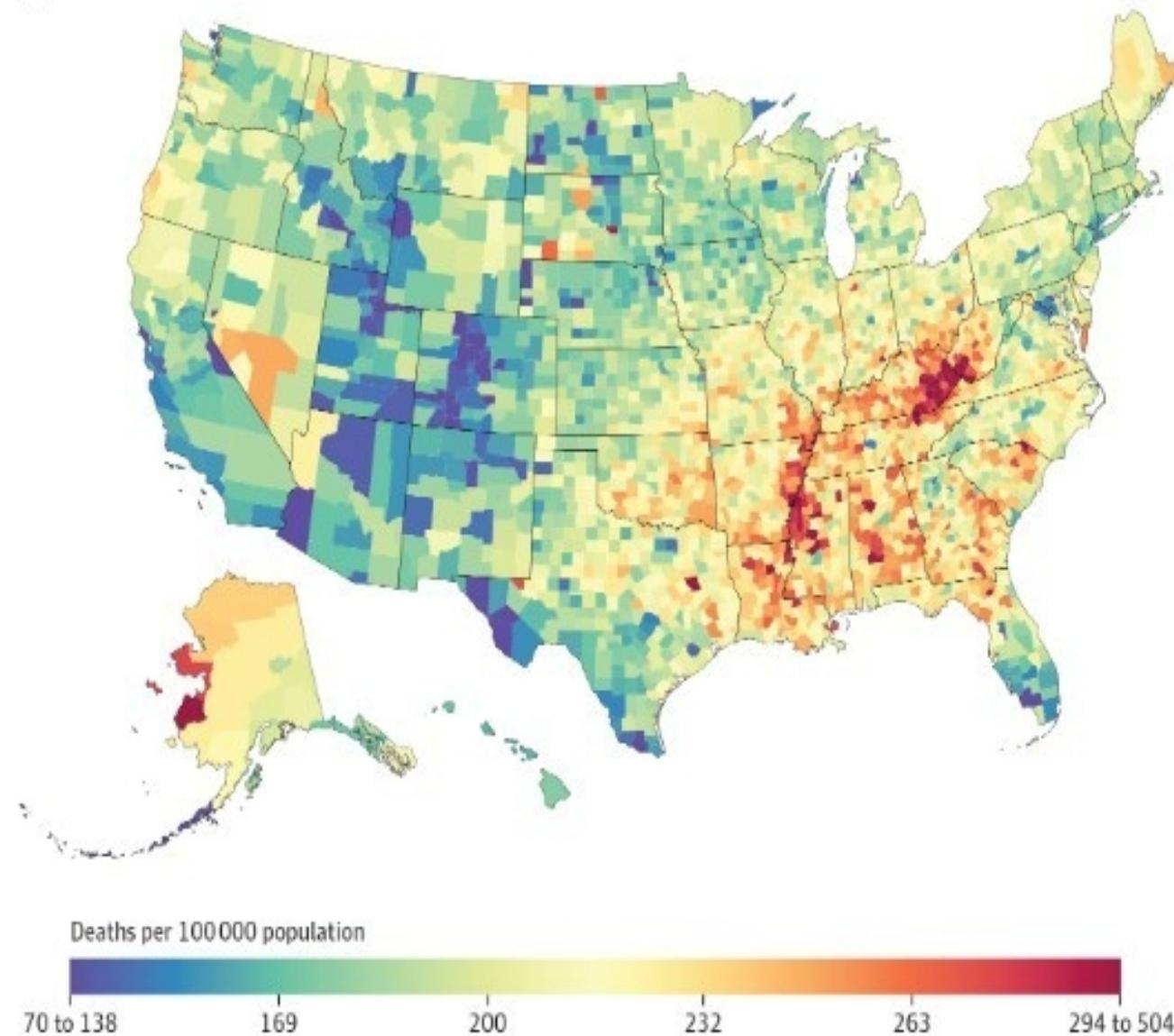
Example:

Zinc concentration
Parts per million (ppm)
River Meuse, France
zone 15m × 15m

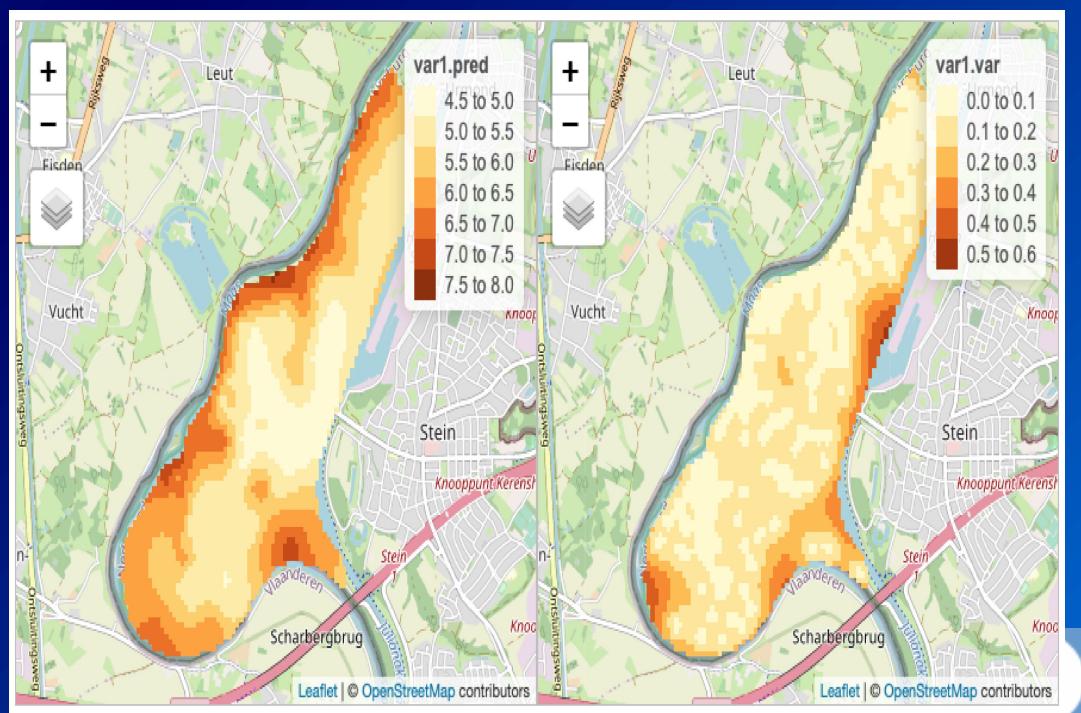
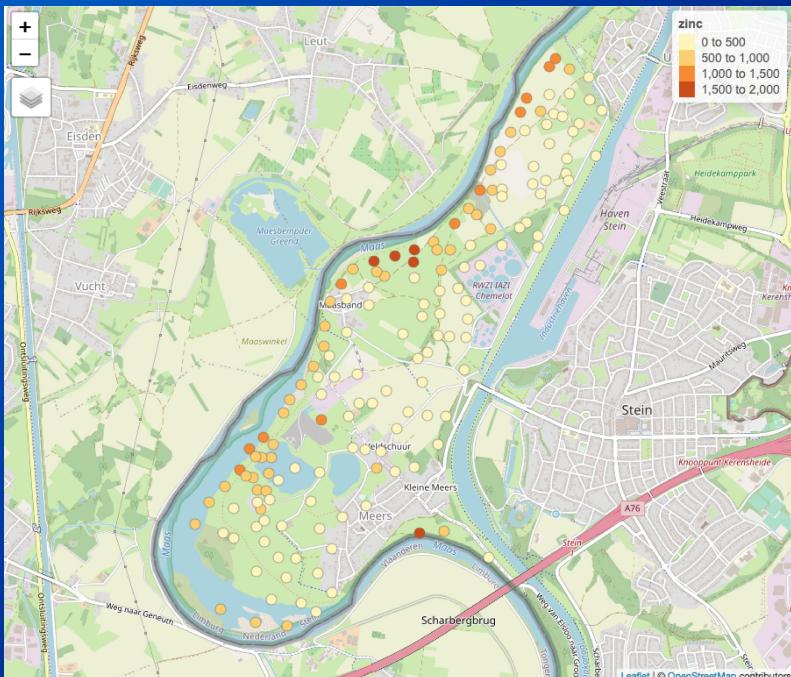


What is spatial statistics?

A Age-standardized mortality rate from neoplasms, both sexes, 2014



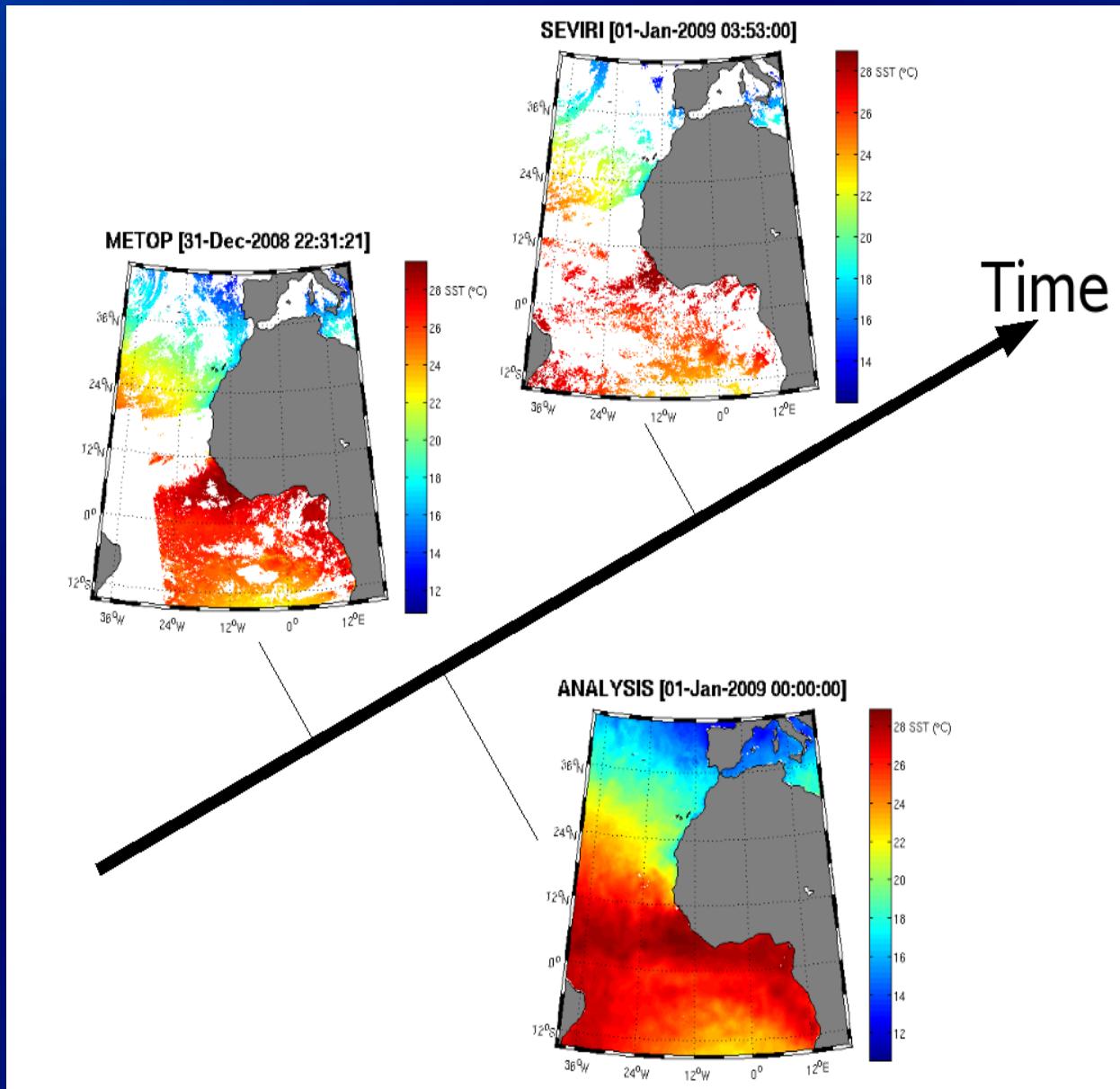
What is spatial statistics? → MODELS



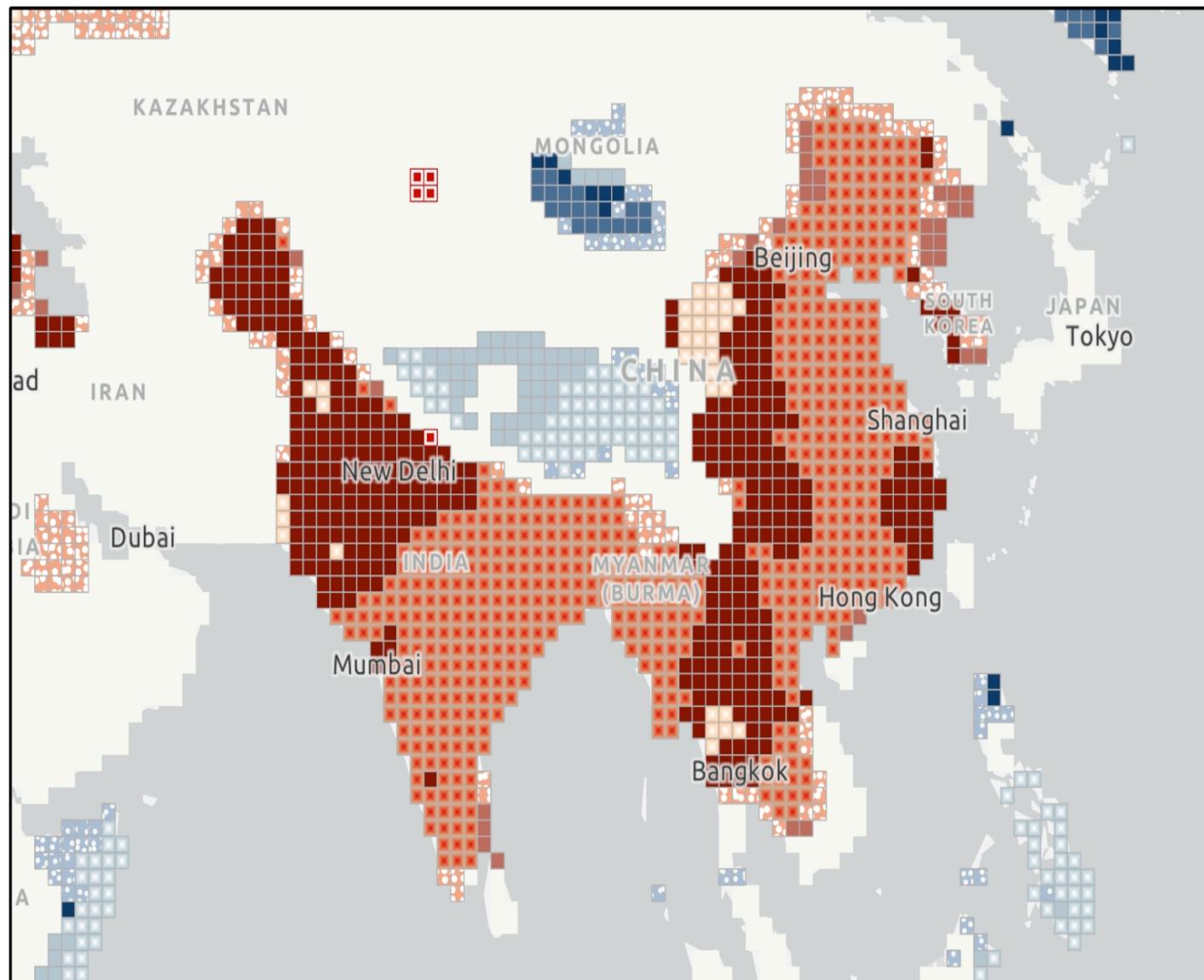
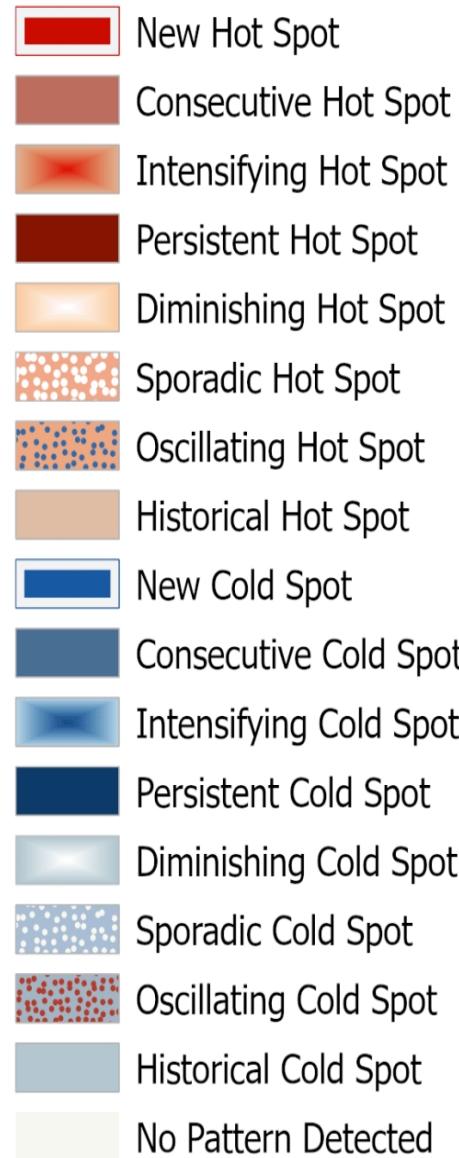
What is spatio-temporal statistics?

Indexing:
place and date

In nature:
Hardly separable
as geography and
history



What is spatial Statistics + time influences?



What is spatial statistics? → Typology

- TYPE I -Simple and general cases (Geostatistics): The simplest geostatistical models consider that the variation of the variable of interest will only depend on the distance between two points and, possibly, on their relative position.
- Models with Barriers: certain spatial processes may be affected by barriers within the region of study. This is the case when studying the noise level in a city, for example. A simple geostatistics model will only consider the distance to the sampling points when estimate the noise level, without ignoring that there are architectural barriers that affect the propagation of noise as buildings.

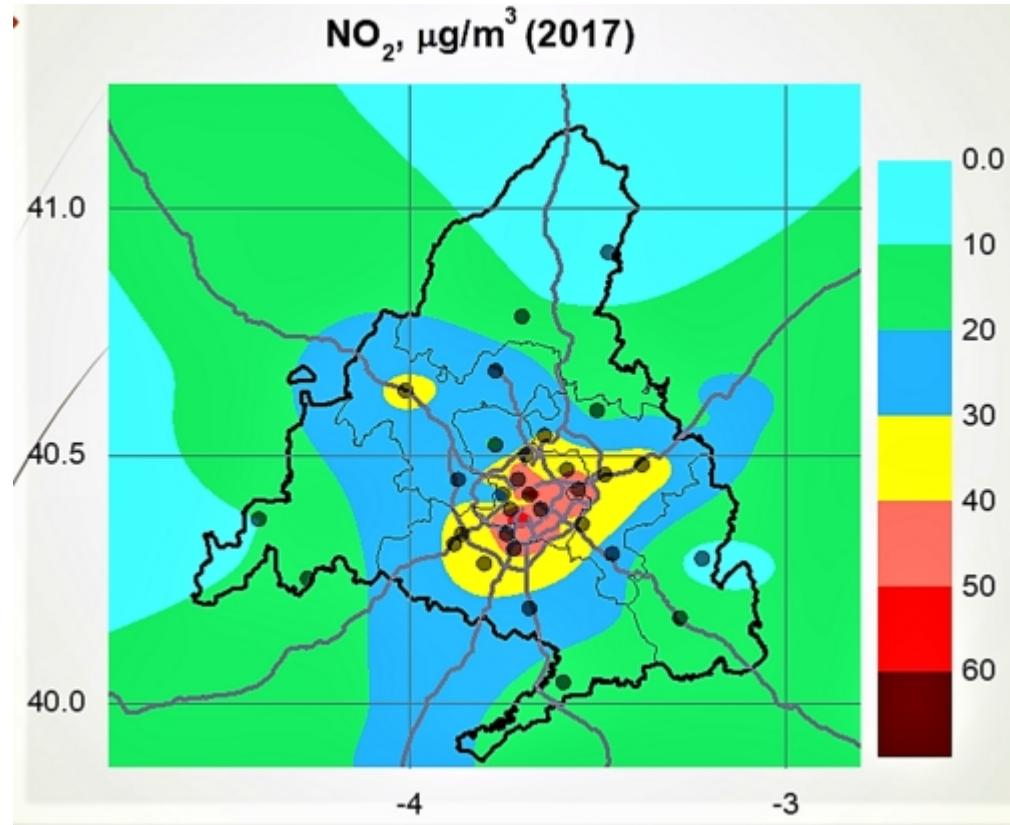


What is spatial statistics? → Typology

- TYPE II - Specific (Puntual Processes): The analysis of point processes focuses on the study of spatial distribution of the occurrence of certain events. Typical examples include the location of trees in a forest or cases of Cancer in a city. In general, the questions that arise when analyzing a specific process focus on whether the appearance of events is uniform in space or if there are areas where events tend to occur more or less frequent.



Introducción – Geoestadística (Tipo I)



Estimación, mediante interpolación kriging, de la distribución de NO₂ en Madrid y sus alrededores (año 2017). Los puntos negros indican los lugares de muestreo.

Ejemplo:

En Madrid, la geoestadística podría utilizarse para modelar la distribución de la contaminación del aire, basándose en datos de estaciones de medición, y luego predecir la calidad del aire en áreas no muestreadas.

What is spatial statistics? → Summary

“all things are related between themselves, but things that are closer, are more correlated than distant ones”

Geostatistics focuses on the study of variables that change in space continuously. such as, for example, environmental pollution in a city or the temperature on the surface of the sea. In this case, observations are available at a few measurement points, and the models are made to estimate the variable of interest in other parts of the study region.



Conceptos básicos

El término geoestadística apareció por primera vez en Matheron (1962). El objetivo es el uso de métodos probabilísticos-inferenciales con la incorporación del componente geográfico.

La geoestadística estudia los fenómenos regionalizados, que son aquellos que:

- (1) Se extienden en el espacio, siendo el dominio espacial, D, continuo (se puede observar en cualquiera de sus puntos) y fijo (las ubicaciones observadas, si $i=1,2,3,\dots,n$, no son estocásticas; se seleccionan, por el procedimiento que sea, a juicio del investigador).
- (2) Presentan una organización o estructura debida a la dependencia espacial existente.

El objetivo fundamental de la geoestadística es sacar provecho de la dependencia espacial existente para llevar a cabo predicciones (interpolaciones) óptimas en ubicaciones o áreas de interés donde no disponemos de mediciones.

Las dos partes del análisis geoestadístico son: el análisis estructural de la dependencia espacial y la predicción (que se suele acompañar del calificativo “krigeada”).

Conceptos básicos

(1) **Variable regionalizada (v.r.)** o regionalización, definida en un espacio geográfico, y que se supone que mide y representa correctamente dicho fenómeno.

El conjunto $Z(s)$ con sus valores $Z(s_i)$ $i = 1, 2, 3, n$ representan una colección de valores regionalizados o v.r.'s y conforman una **función aleatoria geoespacial (f.a.)**.

En cada localización s , $Z(s)$ es una variable aleatoria.

Para un conjunto de puntos dado, s_1, s_2, \dots, s_n , las variables aleatorias $Z(s_1), Z(s_2), \dots, Z(s_n)$ están ligadas por una red de correlaciones espaciales que son las responsables de la similitud en los valores que toman.

Las f.a. serán estrictamente estacionarias si mantienen su ley de probabilidades en su traslación, es decir que la ley de probabilidades se mantiene en $Z(s_1)$ y $Z(s_1+h)$
(Condición muy ideal en problemas reales)

Las f.a serán estacionarias de segundo orden si la media (Esperanza) y la varianza en el traslado h es la misma. (Condición que se asume para la mayoría de los problemas reales)

Conceptos básicos (IMPORTANTE)

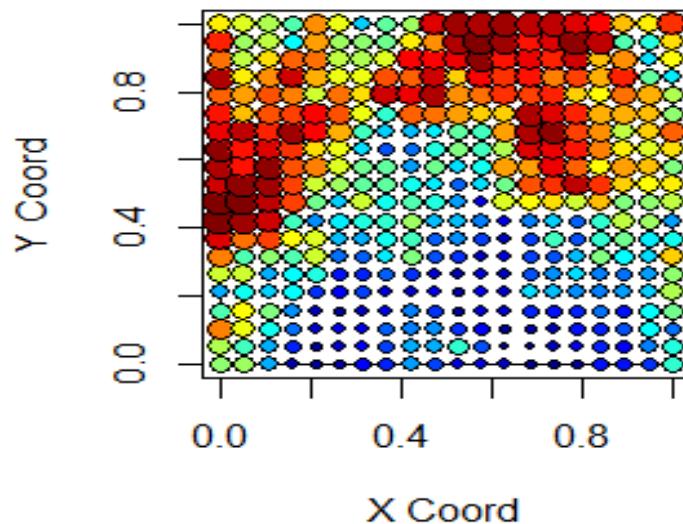
(*) Por tanto, la hipótesis de estacionariedad significa que la ley espacial que gobierna f.a., o parte de ella, es invariante a traslaciones.

(**) La hipótesis de estacionariedad permitirá actuar como si todas las v.a. que conforman la f.a. tuvieran la misma distribución de probabilidad (o los mismos momentos, aquí esperanza y varianza), haciendo posible el proceso inferencial. Por eso se le da tanta importancia a que la f.a. sea estacionaria, del tipo que sea.

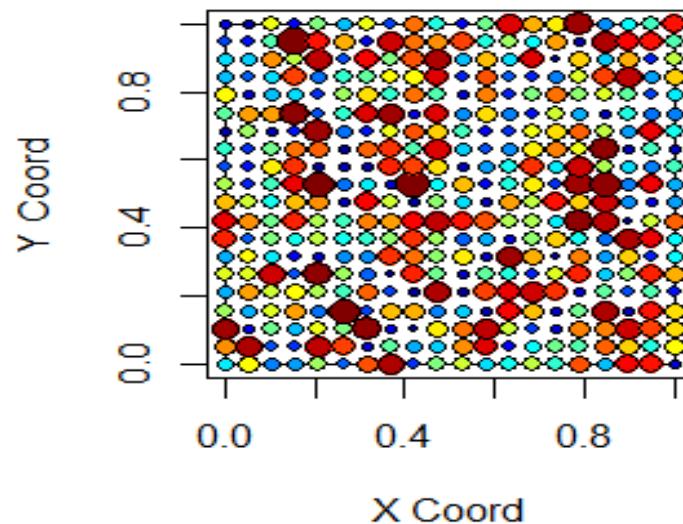
MODELADO GEOESPACIAL – ANÁLISIS ESTRUCTURAL DE LA DEPENDENCIA ESPACIAL y PREDICCIÓN

La estadística espacial se basa en la suposición de que las unidades georreferenciadas cercanas están relacionadas (son dependientes) de alguna manera.

Dependencia espacial



Aleatoriedad



MODELADO GEOESPACIAL – ANÁLISIS ESTRUCTURAL DE LA DEPENDENCIA ESPACIAL

Ahora bien, para poder llevar a cabo predicciones geoestadísticas es necesario representar previamente los patrones de dependencia espacial observados mediante funciones que indiquen cuál es la estructura de dicha dependencia espacial. Dichas funciones **son los semivariogramas**. Dado que la identificación de la estructura de la dependencia espacial existente en el fenómeno de interés es la clave del éxito del proceso predictivo, al semivariograma se le considera la piedra angular de la predicción geoestadística.

Models and Semivariograms

If we consider a stochastic process $Z(x)$, where x indicates the location of the point in space D in which the variable is observed, it is common to assume that the process is stationary of second order, i.e., that the study region D has constant mean μ and variance σ^2 . In addition, The process is also often considered to be isotropic in which the covariance between two any observations depend only on the distance separating them (and not on the direction or their relative position).

$$\hat{Z}(x) = \sum_{i=1}^n \lambda_i Z(x_i).$$

The weights λ_i are such that their sum is one and are obtained in such a way to minimize the variance of the error of estimation. These weights depend on what is known as a (semi)variogram, which measures the dependence between any two x_1 and x_2 points in space:

$$\gamma(x_1, x_2) = \frac{1}{2}E[(Z(x_1) - Z(x_2))^2]$$

where $E[\cdot]$ denotes the expected value. In a stationary and isotropic process the variogram only depends of the distance between two points.

BASIC CONCEPTS

Random variable:
denoted Y

Indexing in space:

$s(x, y) \in D_s$

with (x, y) coordinates
 $D_s = \{s_1, s_2, \dots, s_n\}$

Geostatistical process:

Random variable succession

in function of space

denoted by $\{Y(s), s(x, y) \in D_s\}$

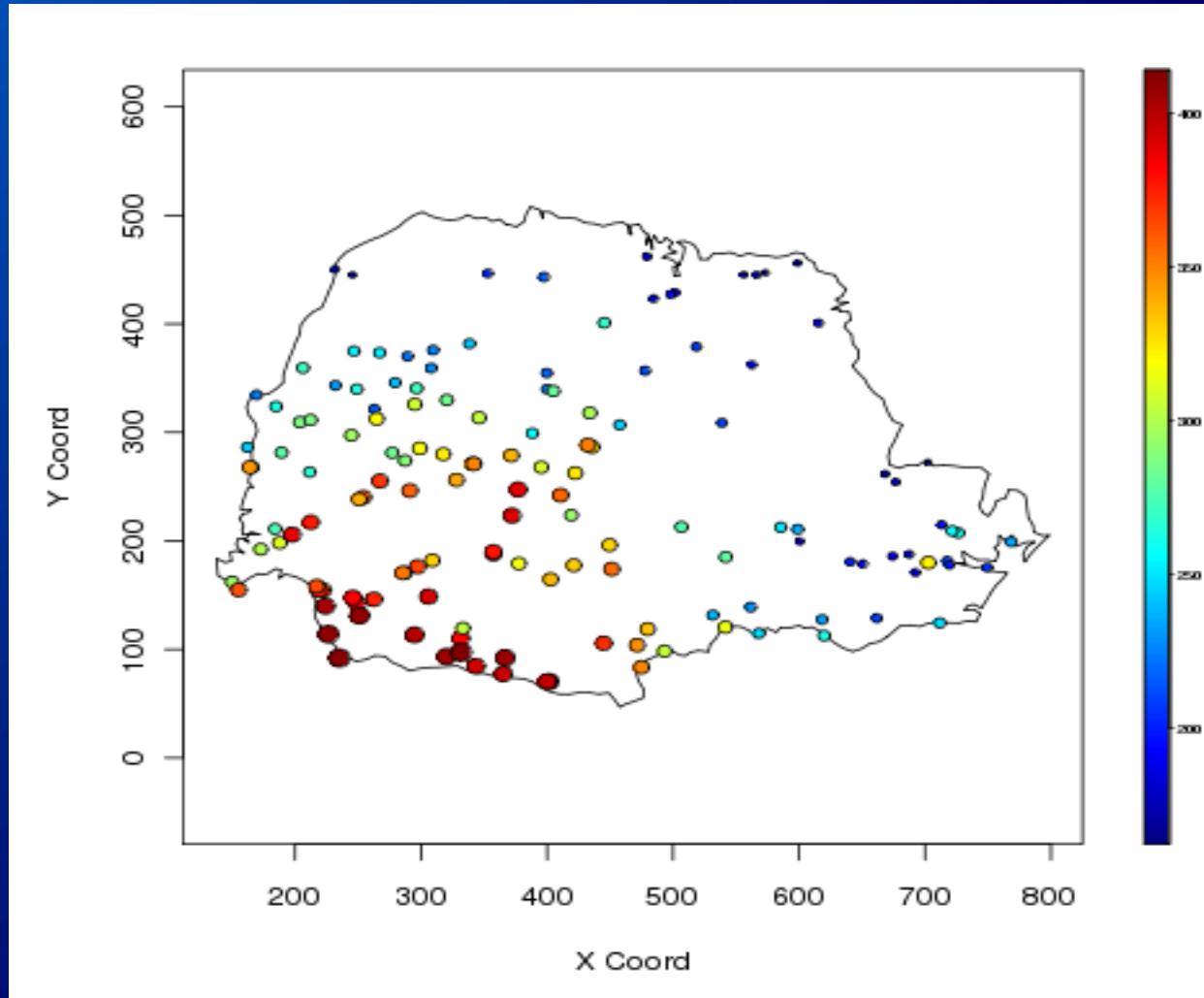


Data:

Rainfall measurement

in mm 143 sites in the state Parana in Brazil

in May 2000



GOALS:

Have a model: simple
explainable and
parametric.

Make predictions:
Interpolation

Have information:
Validation and
prediction(Mean and
Variance on the data)

Models in Spatial statistics

Esquematización de la modelización espacial

$$Y(s) = T(s) + Z(s), \quad s(x, y) \in D_s$$

- Tendencia espacial $\{T(s), \quad s(x, y) \in D_s\}$
- Componente estacionaria espacial $\{Z(s), \quad s(x, y) \in D_s\}$

Modelización de la tendencia espacial

$$\begin{aligned}T(s) = & \alpha_0 + \\& \alpha_1 x + \alpha_2 y + \alpha_3 xy + \\& \alpha_4 x^2 + \alpha_5 y^2 + \alpha_6 x^2 y + \alpha_7 xy^2 + \alpha_8 x^2 y^2 + \\& \varepsilon(s), s(x, y) \in D_s\end{aligned}$$

- Estimación de $\alpha_0, \dots, \alpha_8$:
 - modelo de regresión múltiple
 - mínimos cuadrados
- Criterio:
 - modelo simple
 - modelo explicable
 - R_{adj}^2

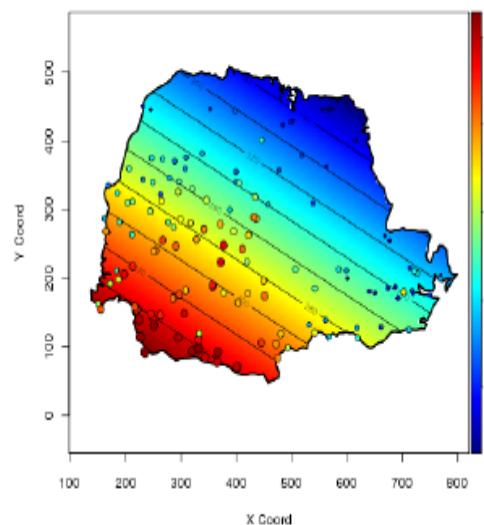


Figura: Tendencia polinomial de grado 1
($R_{adj}^2 = 0,73$)

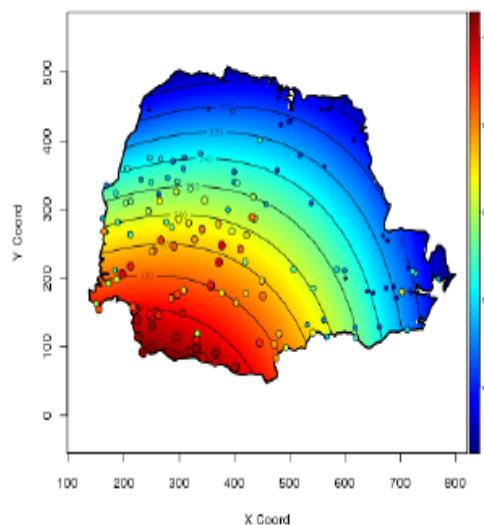


Figura: Tendencia polinomial de grado 2
($R_{adj}^2 = 0,78$)

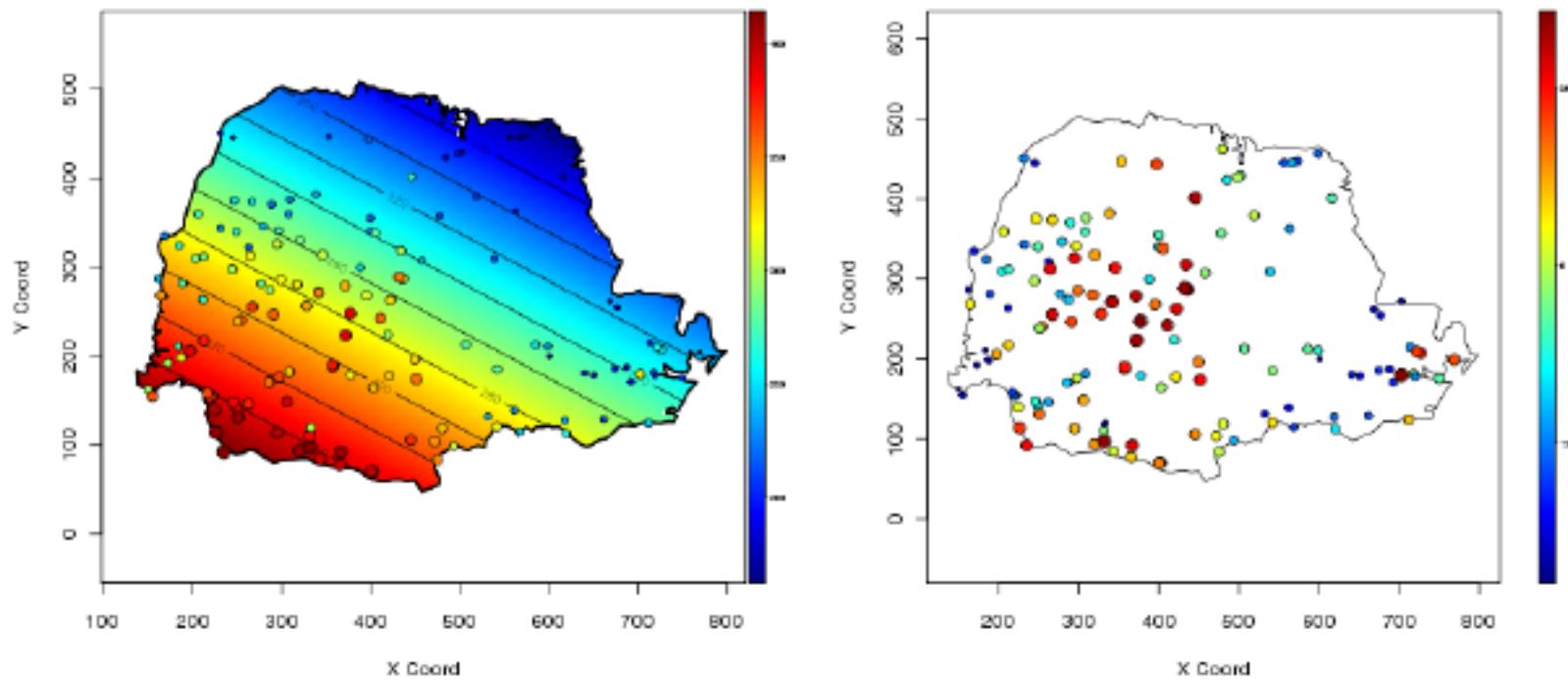


Figura: Tendencia retenida y datos $\{Y(s) - T(s), s(x, y) \in D_s\}$ sin tendencia

Modelización espacial aditiva

$$Y(s) = T(s) + Z(s), \quad s(x, y) \in D_s$$

- Lo que queda luego de quitar:
 - la tendencia espacial $\{T(s), s(x, y) \in D_s\}$
- Se denota $\{Z(s), s(x, y) \in D_s\}$
- Olvidamos la hipótesis iid
- Dependencia posible entre valores cercanos

- Proceso estacionario de:

- sentido amplio
- segundo orden

- Orden 1:

$$E(Z(s)) = 0, \forall s(x, y) \in D_s$$

- Orden 2:

$$\begin{aligned} \text{Cov}(Z(s), Z(s')) &= E(Z(s)Z(s')) \\ &= C_Z(\Delta_s) \end{aligned}$$

- Con:

- $C_Z(\Delta_s)$ la función de covarianza espacial
- $\Delta_s = \sqrt{\Delta_x^2 + \Delta_y^2}$ la distancia euclídea entre s y s'

- En la práctica:

- no trabajamos con $C_Z(\Delta_s)$
- trabajamos con $\gamma_Z(\Delta_s)$

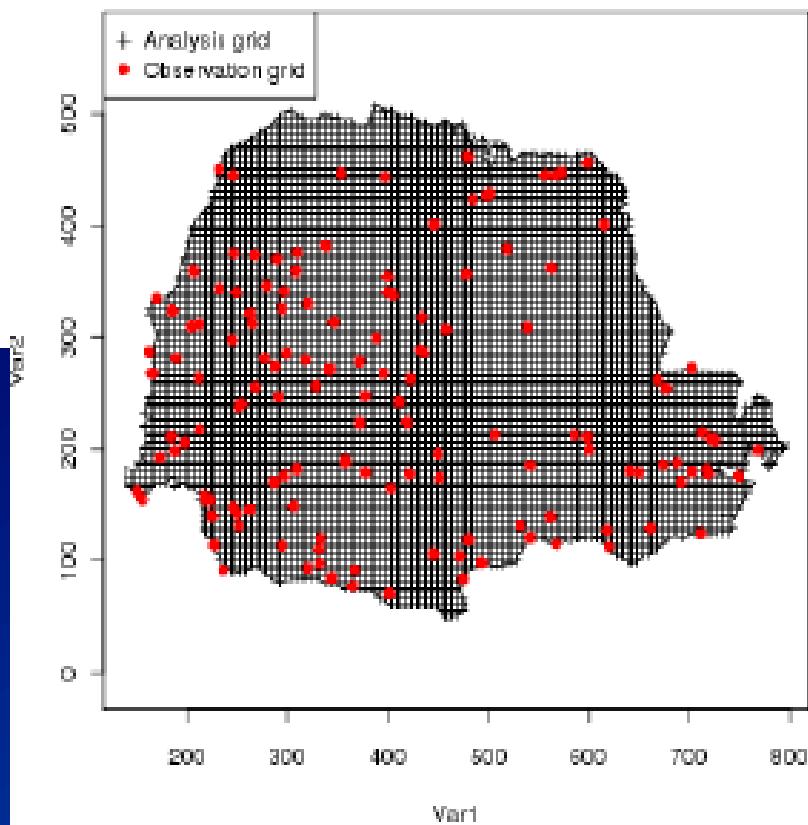
- Definición:

$$\gamma_Z(\Delta_s) = \frac{1}{2} \text{Var}(Z(s) - Z(s'))$$

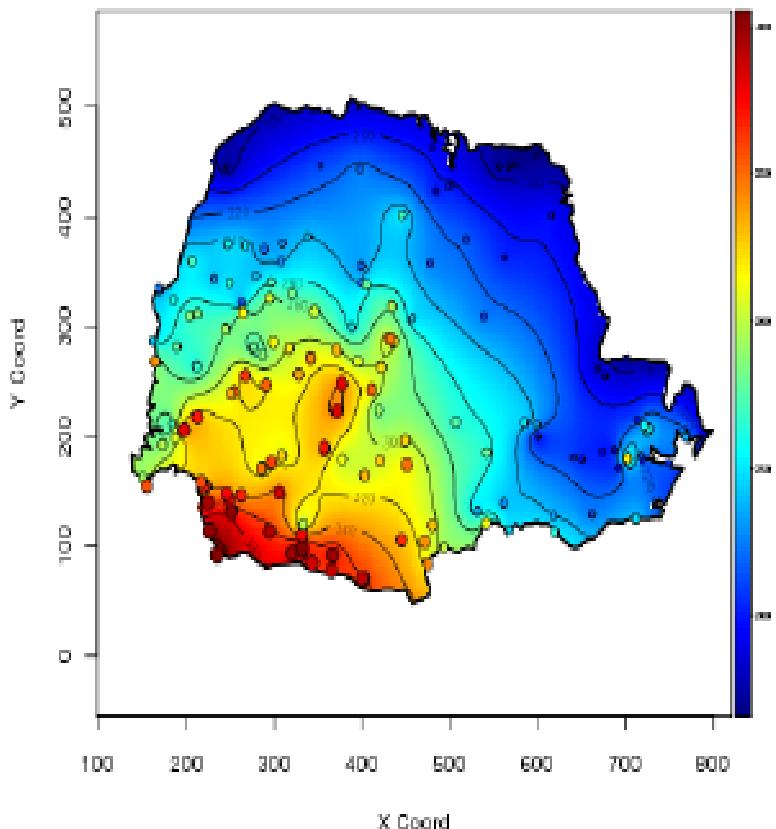
- Link con la función de covarianza espacial:

$$\gamma_Z(\Delta_s) = C_Z(0) - C_Z(\Delta_s)$$

- Grilla de observación:
 - donde hay las observaciones
 - se denota $s(x, y) \in D_s$
 - n puntos
- Grilla de análisis:
 - o grilla de interpolación
 - donde queremos estimar Y
 - se denota $s_0(x_0, y_0) \in D_{s_0}$
 - n_0 puntos



Kriging mean estimates



Kriging variance estimates

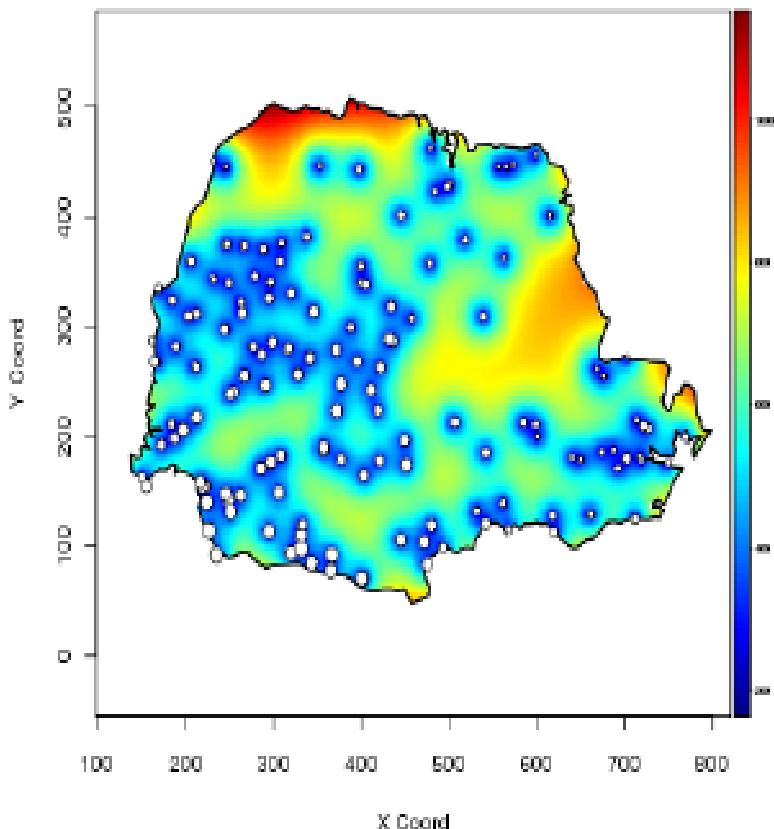
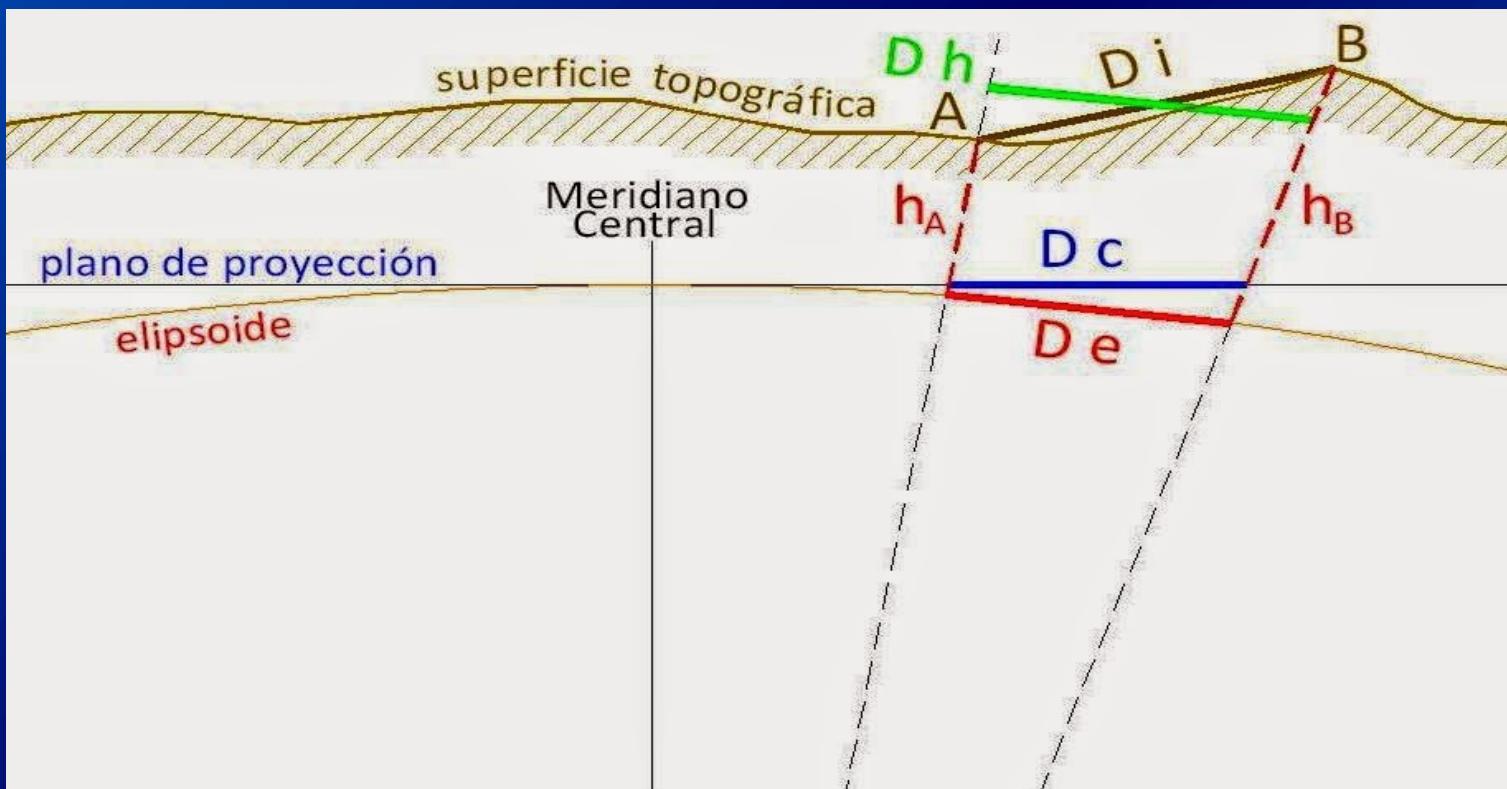


Figura: Estimación de la media
(krigeado universal)

Figura: Estimación de la varianza
(krigeado universal)

MOVING TO SECOND PART → GEODESIC DISTANCE

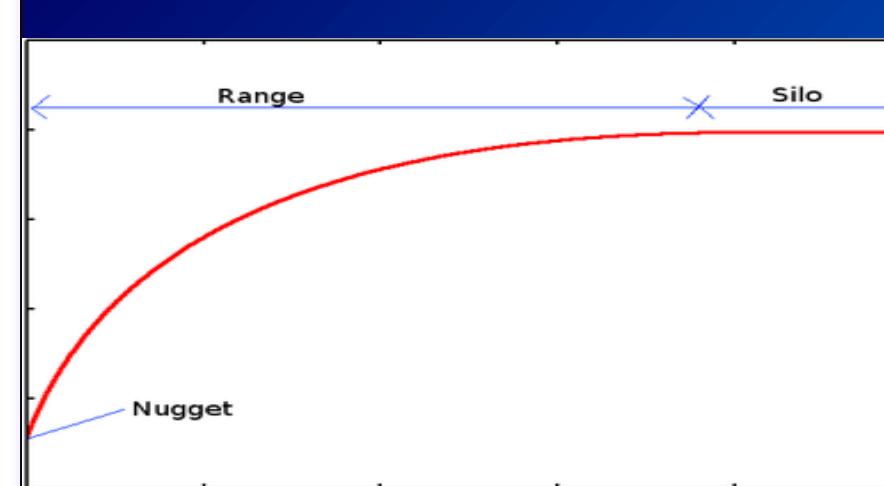
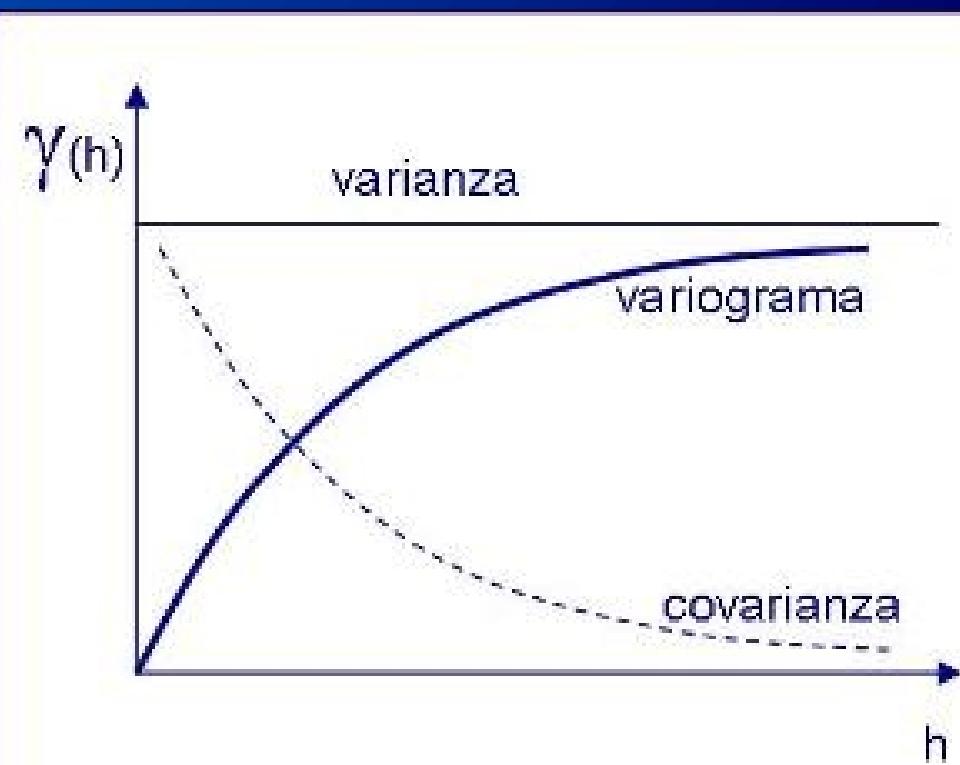
A geodesic line is the shortest path between two points on a curved surface, such as the Earth. They are the analogue of a straight line on a flat surface or whose plane of section at all points along the line remains normal to the surface. It is a way of showing the distance on an ellipsoid while that distance is projected onto a flat surface.



MOVING TO SECOND PART → VARIOGRAM

The variogram or semivariogram is a tool that allows analyzing the spatial behavior of a variable over a defined area, resulting in an experimental variogram that reflects the maximum distance and the way in which a point has influence on another point at different distances.

Once the experimental variogram has been calculated, a mathematical model must be adjusted to the experimental variogram in the best possible way, which is known as a theoretical variogram.



Model the spatial trend $\{T(s), s(x,y) \text{ in } D_s\}$:

1) Create the vectors X and Y

- Perform the regression with polynomial:

order 0 (constant)

order 1

order 2

2) For each model:

-calculate the $R^2(\text{adj})$

-Plot predictive values

3) Choose a spatial trend model

4) plot the residuals $\{Y(s) - T(s), s(x,y) \text{ in } D_s\}$



*Modeling the stationary spatial component

$\{Z(s), s(x,y) \text{ in } D_s\}$:

(a) Create the empirical semi-variogram:

(b) Modelling the semi-variogram:

Use exponential, Gaussian and Spherical models.

(c) Choose a model

(d) Interpret spatial variability parameters

(e) Write the spatial model

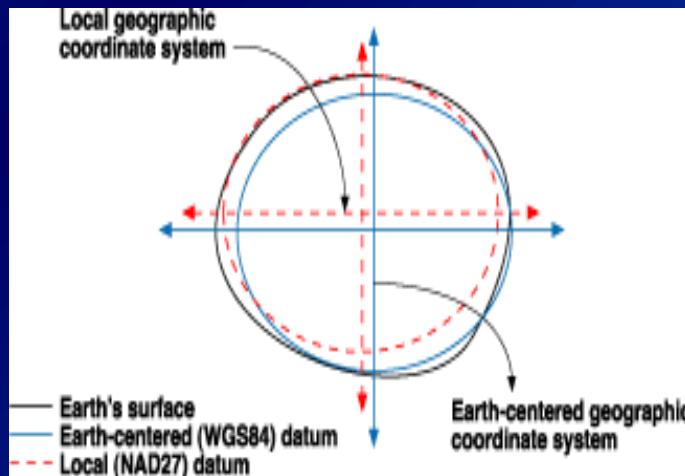
How to proceed?

(1) What is spatial-temporal statistics? - Reminder

The types of spatial data are commonly: geostatistical data, and point patterns. The first type refers to continuous domain data, that is, it assumes that between two points, infinite data can exist. It refers to continuity in the spatial structure of the underlying spatial random process from which the observations have been generated. Because of the continuity of the spatial domain, geostatistical data is also called "spatial data with continuous variation". Continuity is then associated with the underlying random process and not with the measured attribute (whether the variable is continuous or discrete in nature does not determine whether the data is geostatistical or not). Regional or area data are those where the domain is fixed and made up of a discrete set of areas, surfaces, or polygons. Finally, Point patterns of analysis deals with patterns of points that come from a random point process conformed by the points or sites where the events occur.

Some points before modelling (2)

To locate coordinates with which a spatial data is associated, a reference system is needed. There are two types of coordinates, Cartesian and geographical. Cartesian coordinates are measured from the center of the earth, while geographic coordinates are measured from a reference surface or datum. For Europa & South America the commonly used datum is WGS84 (World Geodetic System 84). This is the default standard datum for coordinates on commercial GPS devices. To combine layers of information or to perform other spatial data processing it is necessary to know the datum and often transform or convert the coordinates. Transforming involves moving from one reference frame to another (changing the datum), whereas converting coordinates does not change datum.



Some points before modelling (3)

As a matter of practicality, it is usual to project the geographic coordinate system (expressed in degrees, minutes and seconds) to a Cartesian coordinate system, such as the UTM (Universal Transverse Mercator) projection system. This operation allows distances between sites from which data is read to be expressed as absolute distances (meters) rather than relative distances (degrees). Therefore, an initial step in spatial data analysis is to convert geographic coordinates into Cartesian coordinates (UTM).



Anselin, Luc. 1995. "Local Indicators of Spatial Association—Lisa." *Geographical Analysis* 27 (2). Wiley Online Library: 93–115.

[http://www.agro.unc.edu.ar/~estadisticaaplicada/GpADEAA/
manejo-de-datos-espaciales.html](http://www.agro.unc.edu.ar/~estadisticaaplicada/GpADEAA/manejo-de-datos-espaciales.html)

Semivariogram

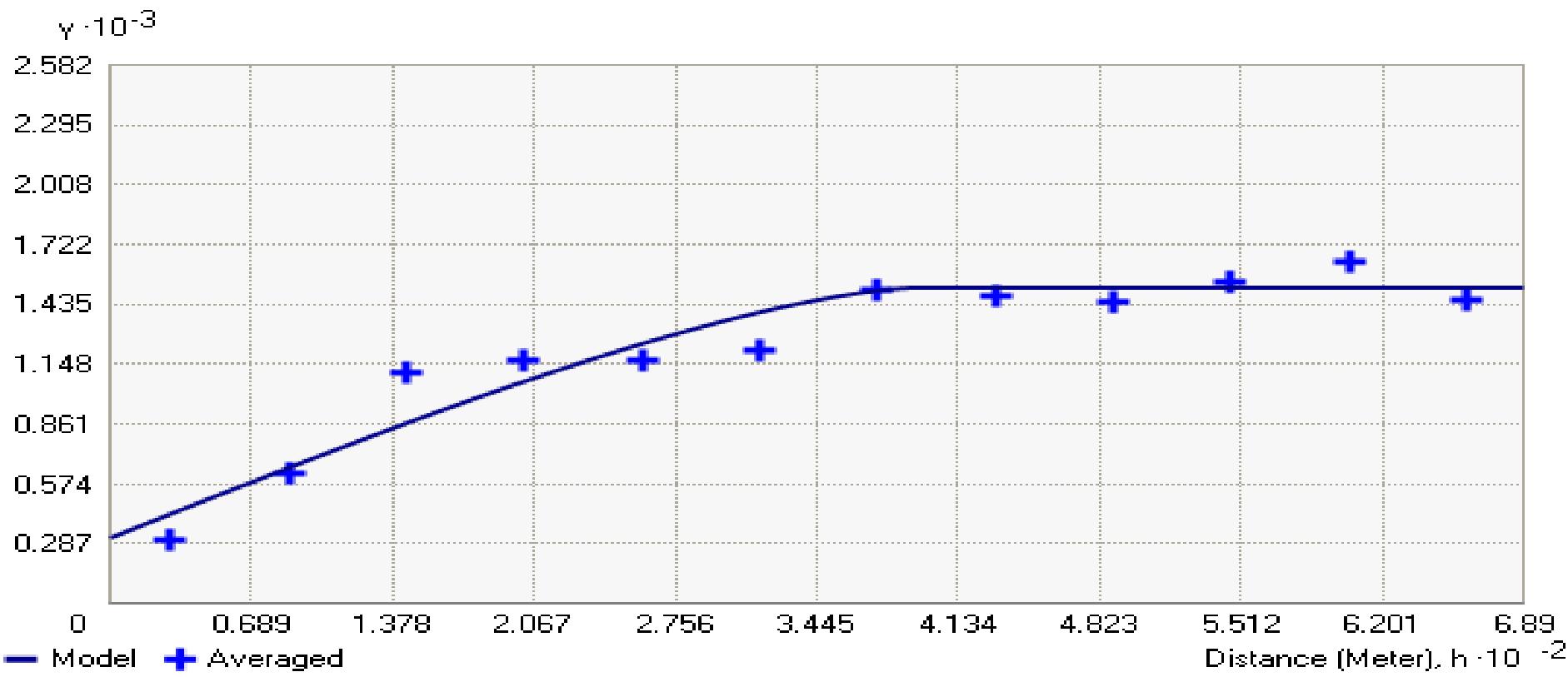
$$\text{if } E[Z(s)] = \mu$$

$$\text{then } 1/2\text{Var}[Z(s) - Z(s+h)] = \gamma(h).$$

The semivariogram can be interpreted as a function of the variance of the difference between the observations. If the semivariogram is only a function of the distance between observations, then it is known as an isotropic semivariogram, i.e. it does not depend on direction. The semivariogram and covariogram are parameters of the spatial process and play a critical role in geostatistical methods of spatial data analysis.

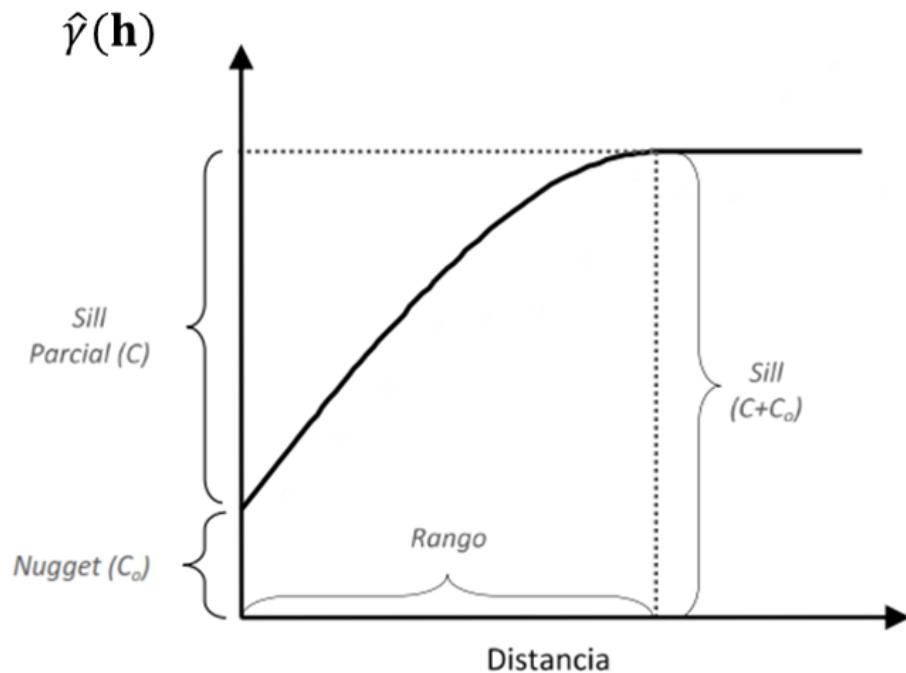
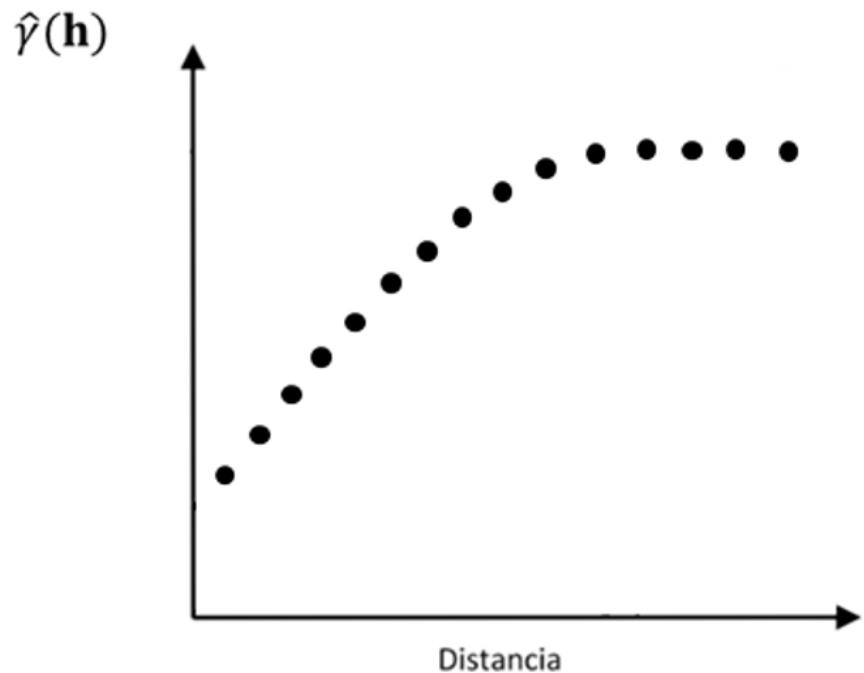
Semivariogram

For all pairs of locations separated by distance h . The formula involves calculating half the difference squared between the values of the paired locations. To plot all pairs quickly becomes unmanageable. Instead of plotting each pair, the pairs are grouped into lag bins. For example, compute the average semivariance for all pairs of points that are greater than 40 meters but less than 50 meters apart. The empirical semivariogram is a graph of the averaged semivariogram values on the y-axis and distance (or lag) on the x-axis.



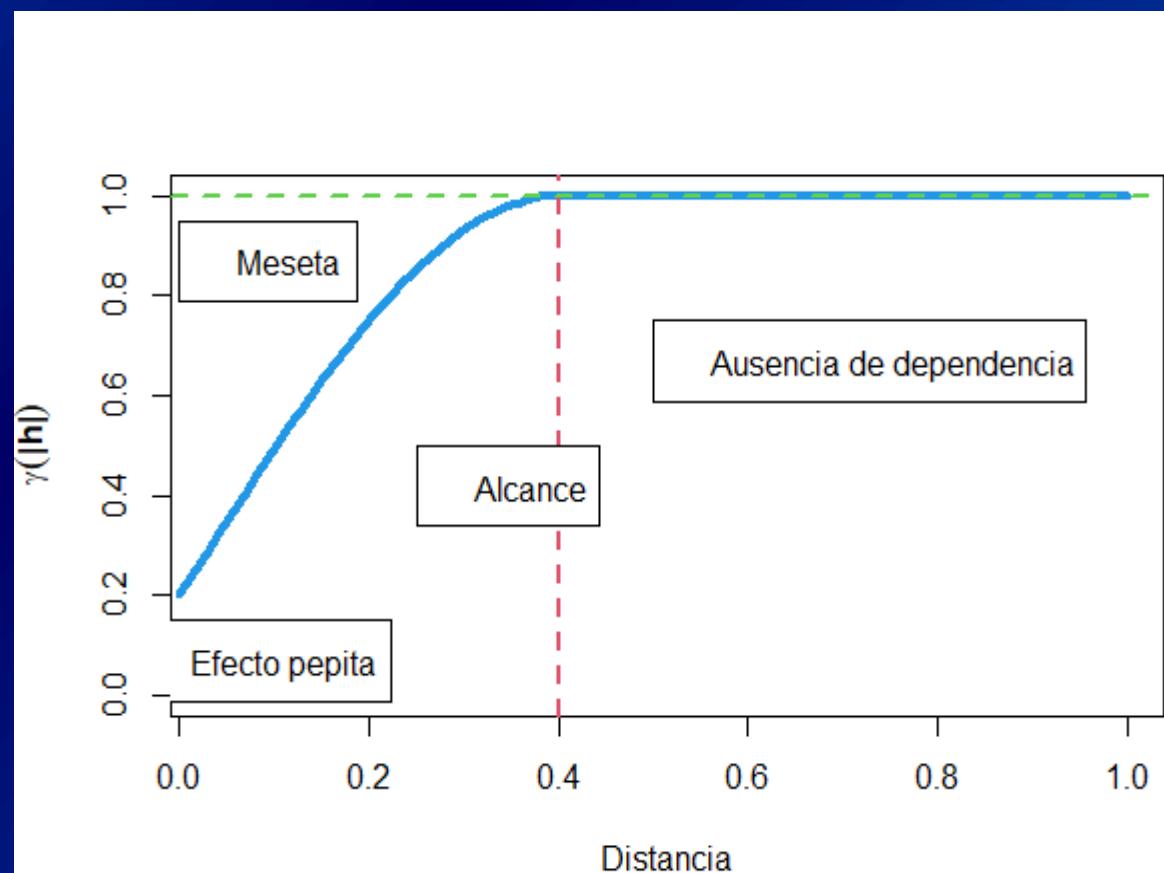
Semivariogram

The parameters of a semivariogram are: the variance nugget or simply nugget (C_0), the structural variance or partial sill (C) and the range (R). The asymptote is called the plateau of the semivariogram or C and the lag or distance h^* in which the plateau is reached is called R or range. Observations $Z(s_i)$ and $Z(s_j)$ for which $|Z(s_i) - Z(s_j)| \geq h^*$ are spatially independent.

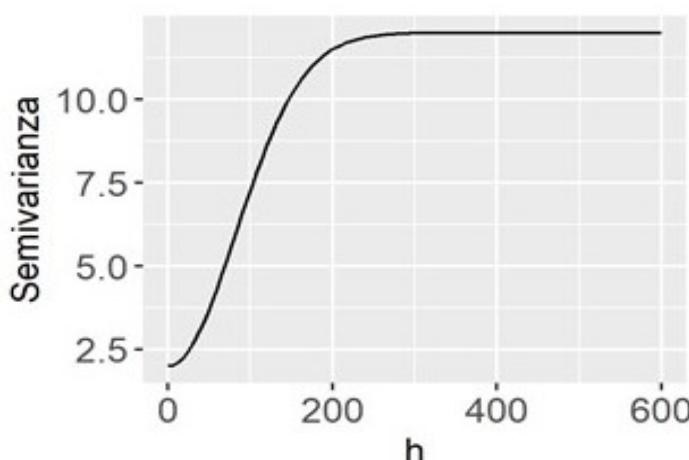


The ordinate at the origin of the semivariogram represents C_0 , therefore $C_0 = \lim_{h \rightarrow 0} g(h) \neq 0$. This parameter represents the sum of random or spatially unstructured errors, as well as errors associated with spatial variability. A high C_0 value indicates that most of the spatial variation is not explained by the semivariogram. The threshold variance or sill is obtained by adding $(C_0 + C)$ and is the variance of independent observations, that is, observations that were taken at a greater distance than R .

Semivariogram

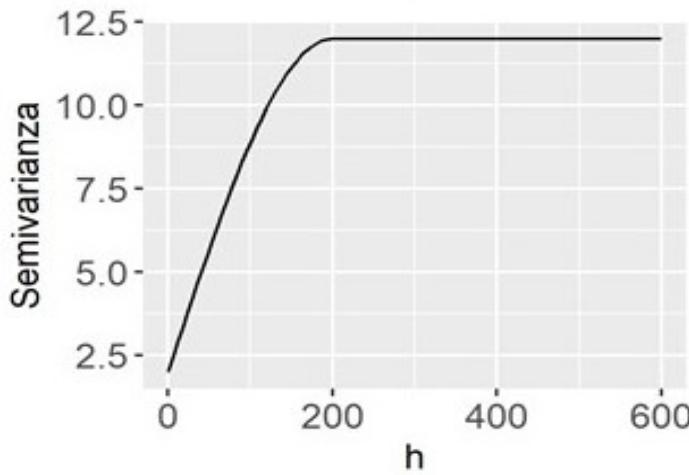


Modelo Exponencial



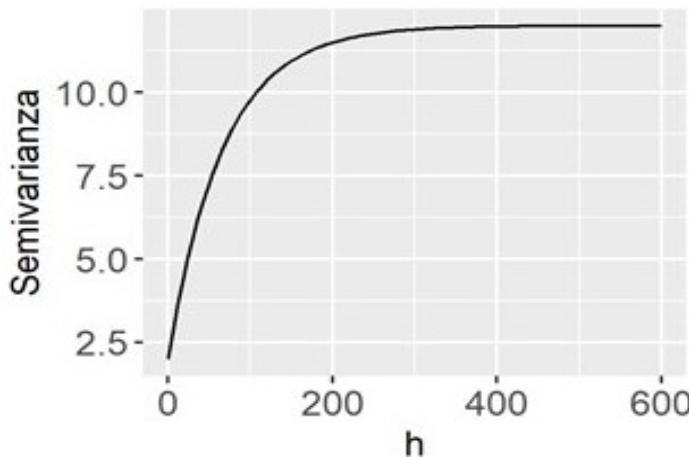
$$\gamma(h) = \begin{cases} C_0 & h=0 \\ C_0 + C \left\{ 1 - \exp \left\{ \frac{-3h}{R} \right\} \right\} & h \neq 0 \end{cases}$$

Modelo Esférico



$$\gamma(h) = \begin{cases} C_0 & h=0 \\ C_0 + C \left\{ \frac{3h}{2R} - \frac{1}{2} \left(\frac{h}{R} \right)^3 \right\} & 0 < h \leq R \\ C_0 + C & h > R \end{cases}$$

Modelo Gausiano



$$\gamma(h) = \begin{cases} C_0 & h=0 \\ C_0 + C \left\{ 1 - \exp \left\{ -3 \left(\frac{h}{R} \right)^2 \right\} \right\} & h \neq 0 \end{cases}$$

Modelling – Kriging Interpolation

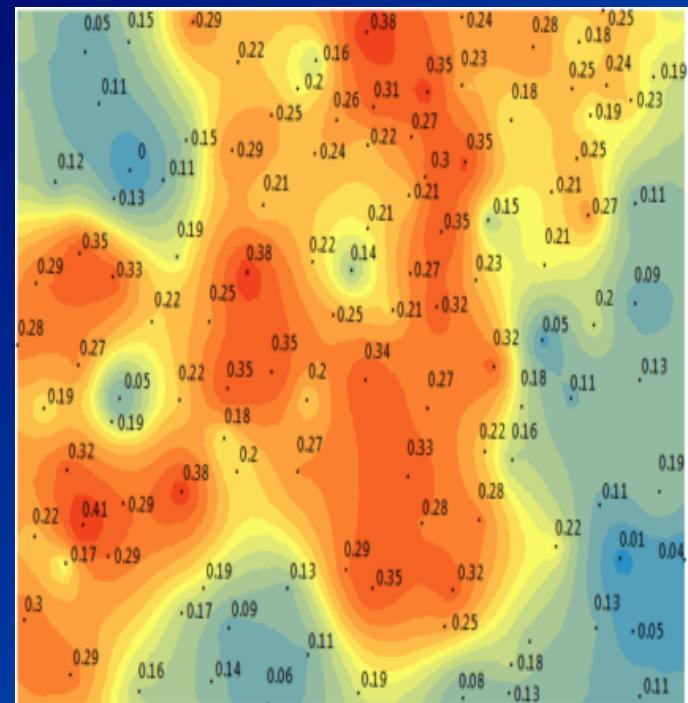
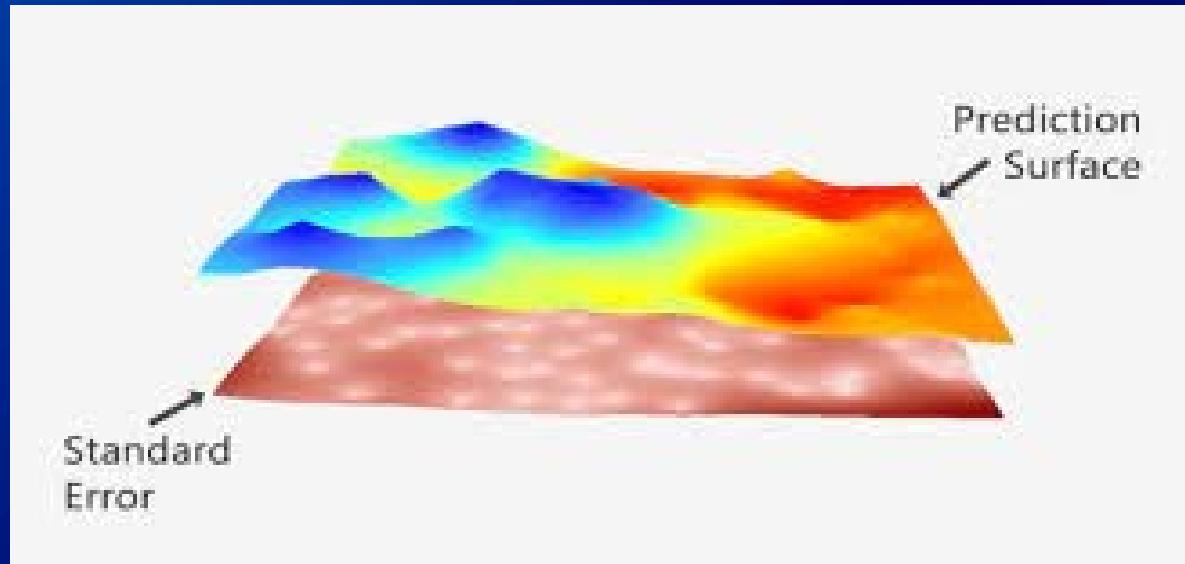
Spatial prediction, that is, the prediction of variable values where there are no observations, is usually done by the kriging method based on the adjusted semivariogram. Kriging provides the best unbiased linear estimator of the expected value for the point and an estimation error known as kriging variance. This variance depends on the fitted semivariogram model and the location in space of the observed data since it is the data observed at different sites that provide information to approximate the value at the point without data.

$$\hat{Z}(x) = \sum_{i=1}^n \lambda_i Z(x_i).$$

$$\gamma(x_1, x_2) = \frac{1}{2} E[(Z(x_1) - Z(x_2))^2]$$

Kriging Interpolation → Types

In ordinary kriging the mean of the variable is estimated locally. If the population mean of the variable is known, which rarely occurs, simple kriging is used. In universal kriging, the influence of a spatial trend of the data is also estimated. The prediction assigned to the unknown points can be done in a timely manner (point kriging) or by defining blocks (block kriging).



Kriging Interpolation → Universal Kriging

The assumption of intrinsic stationarity is not fulfilled when there are pronounced geographical trends of a systematic and non-random nature. The trend can be regional, that is, a systematic variation throughout the region of interest or local from one point to another within the region studied. The existence of trends can be explored by graphing the data of the variable analyzed according to the variable that is supposed to generate the spatial trend. The trend is also manifested in experimental semivariograms with an increase in semivariance with distance that has no limits. If there is a trend, then μ is no longer constant, but depends on s . In addition, the experimental semivariogram of the data no longer estimates the semivariogram of random errors, $\varepsilon(s)$. It is necessary to estimate the semivariogram of $\varepsilon(s) = Z(s) - \mu(s)$. When this variogram is the input of the kriging, the interpolation process is known as "universal kriging"

Kriging Interpolation → Validation

In the case of spatial data, it is not only necessary to have a measure of global prediction error, but also to evaluate the error of the prediction in each specific site, i.e. to size the point error of the spatial prediction.

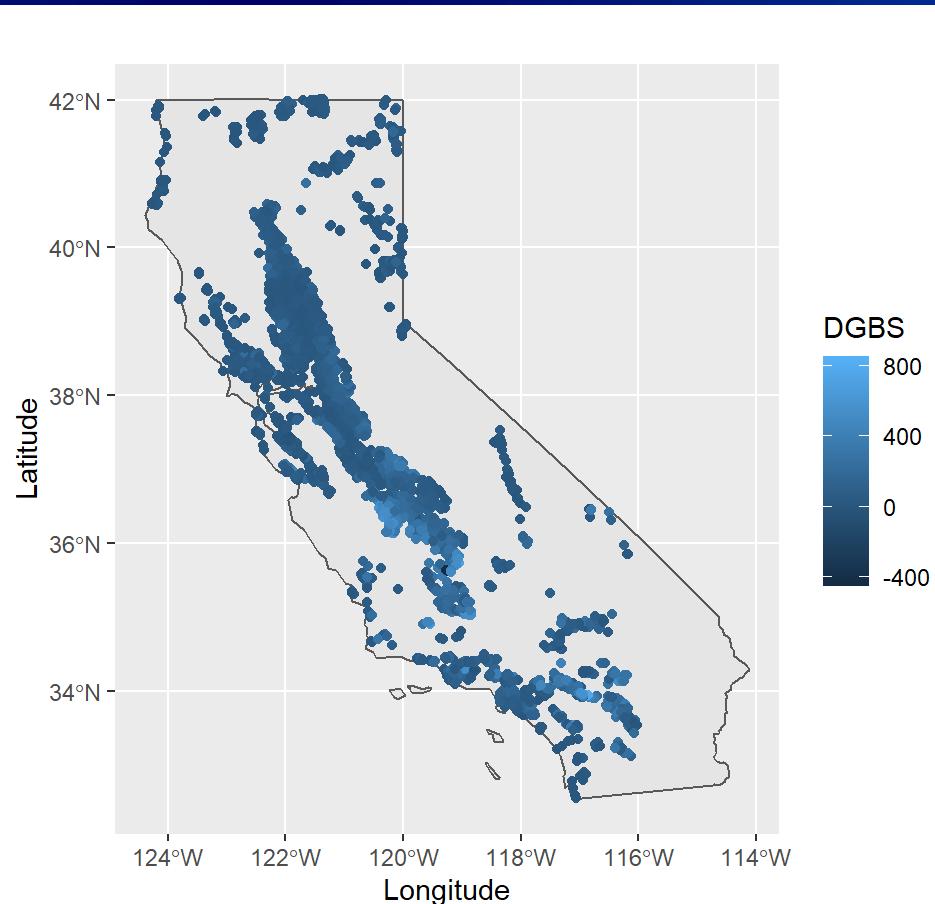
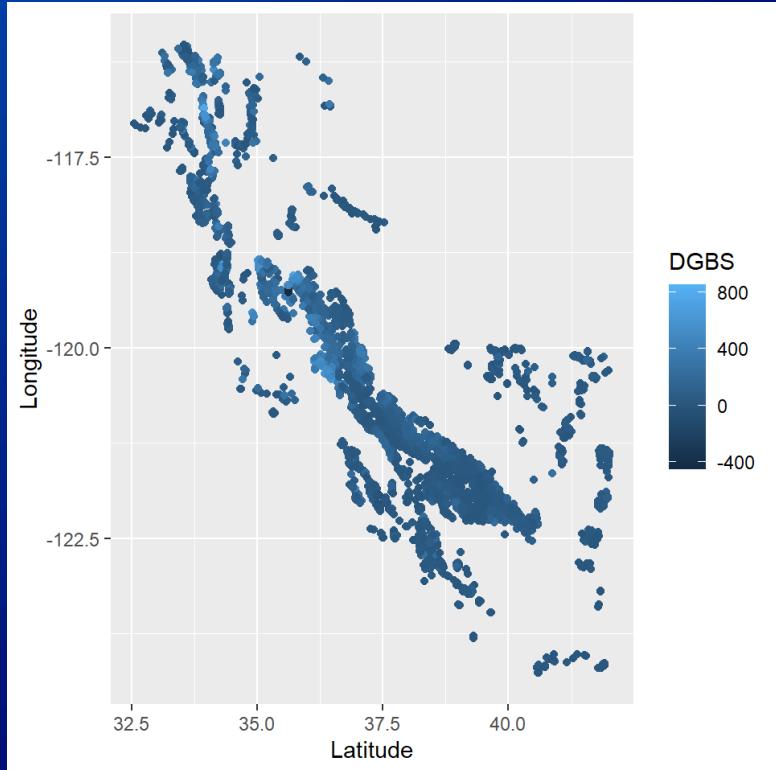


EXAMPLE

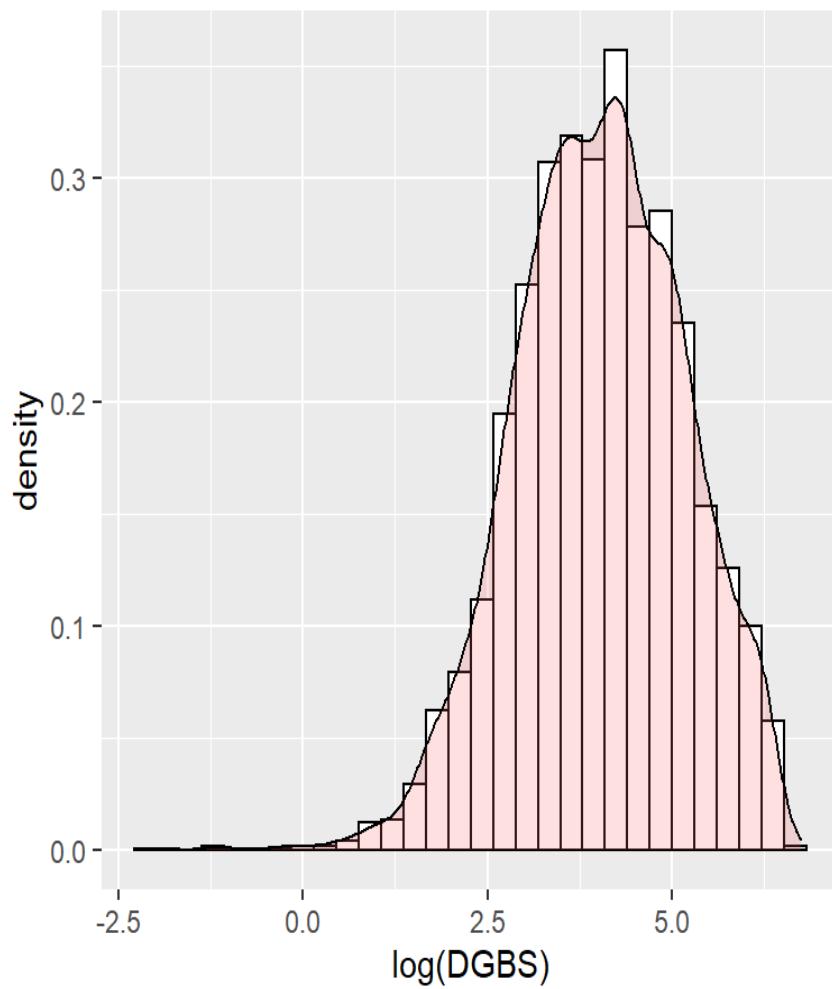
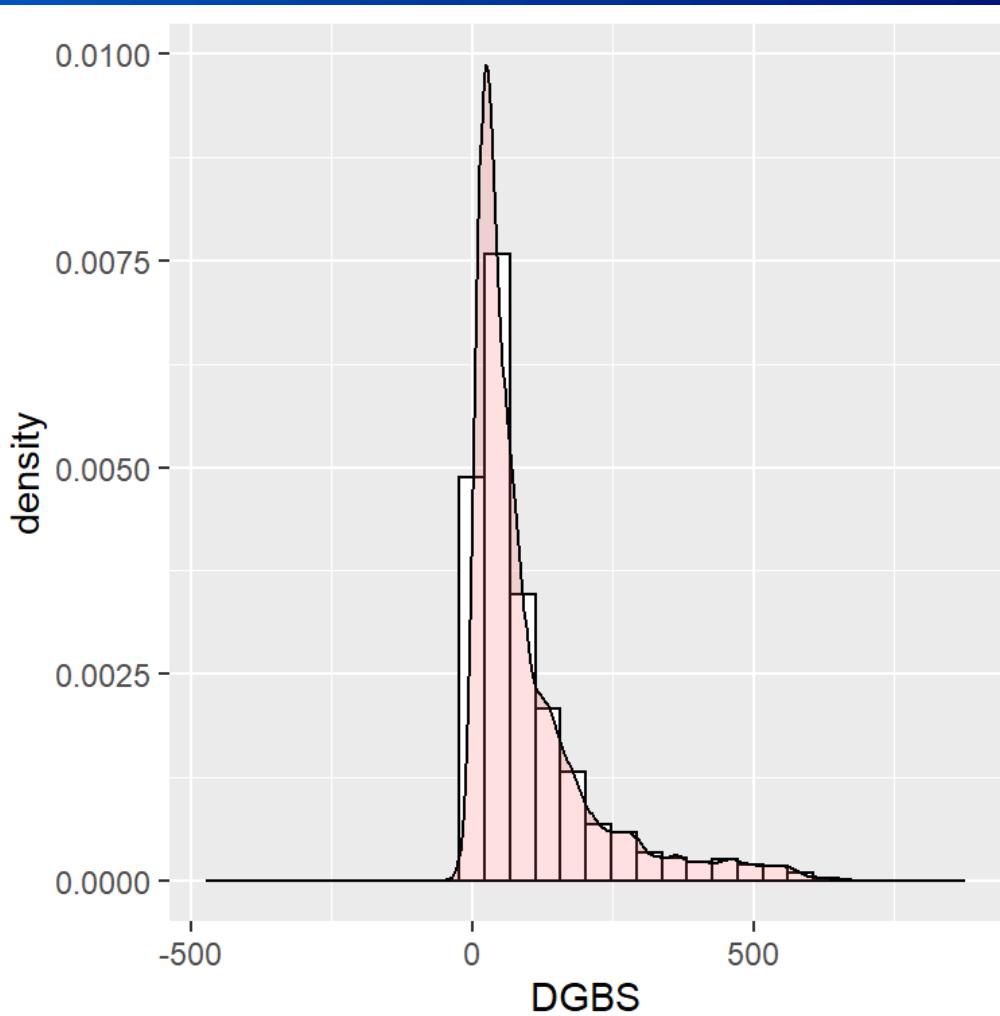
DGBS measurement variable

The DBGS is the depth at which the groundwater is located minus the distance to the reference point to the soil surface.

The measurements were taken in the fall of 2016, between October and December for 4028 underground water wells in California.

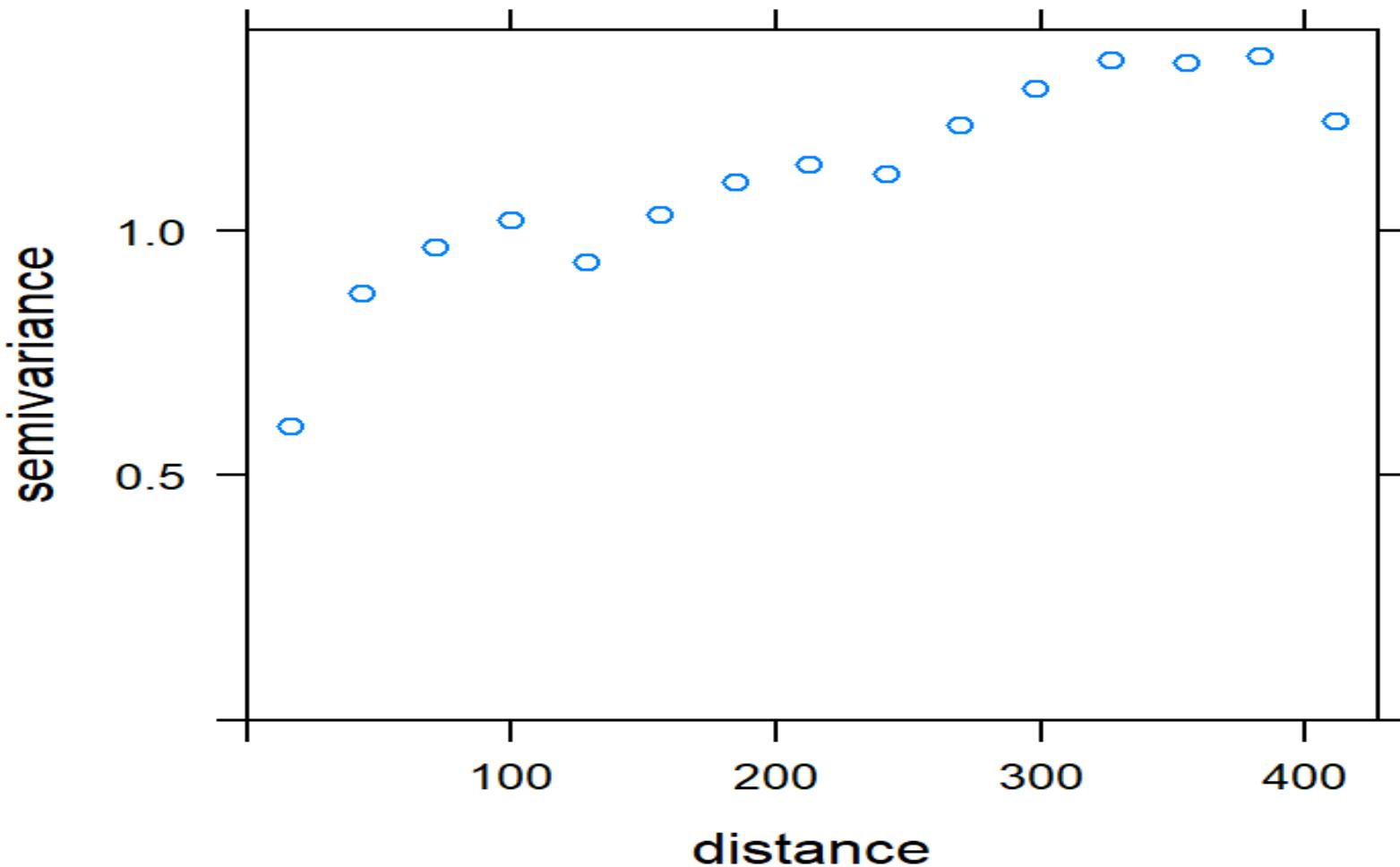


EXAMPLE



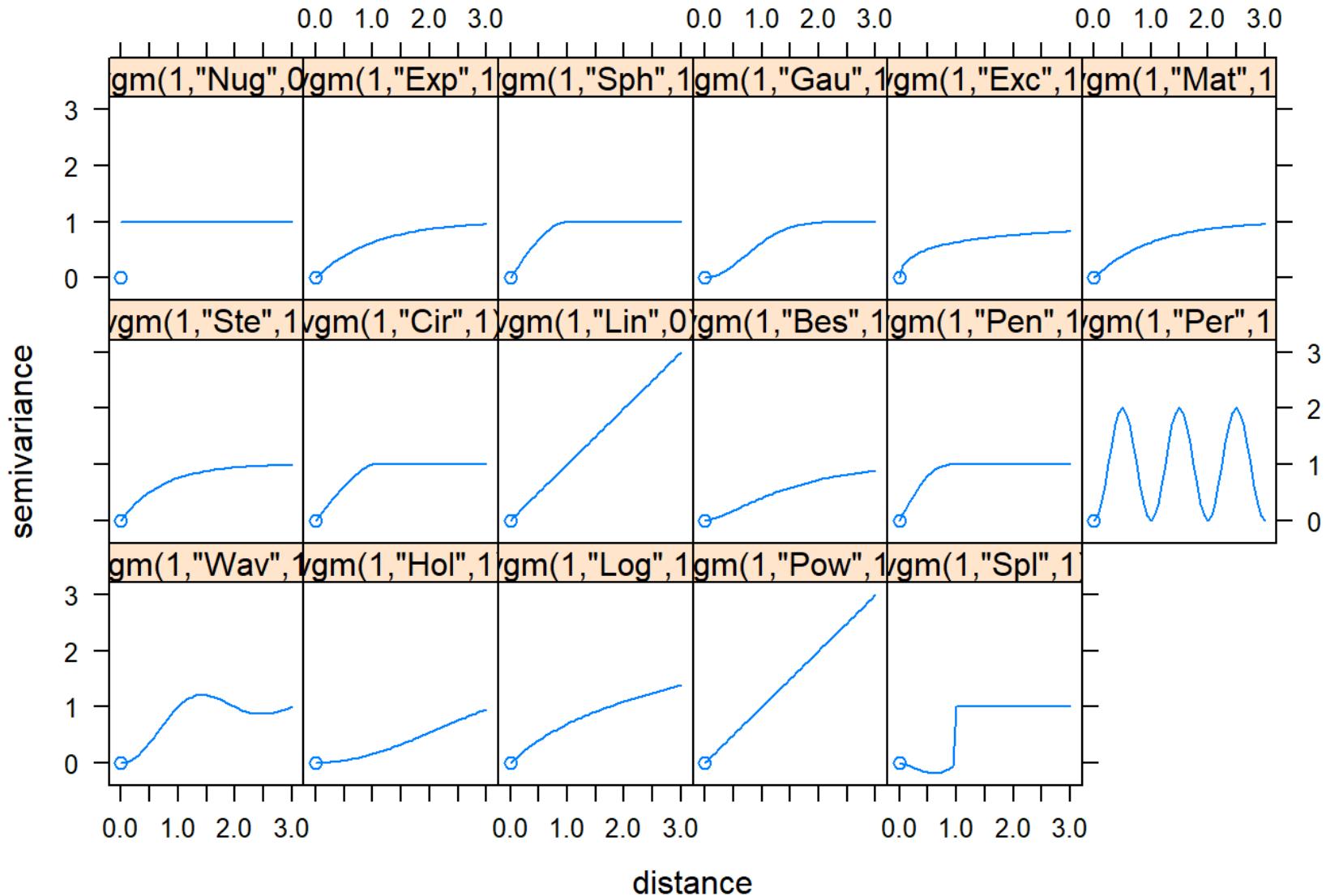
EXAMPLE

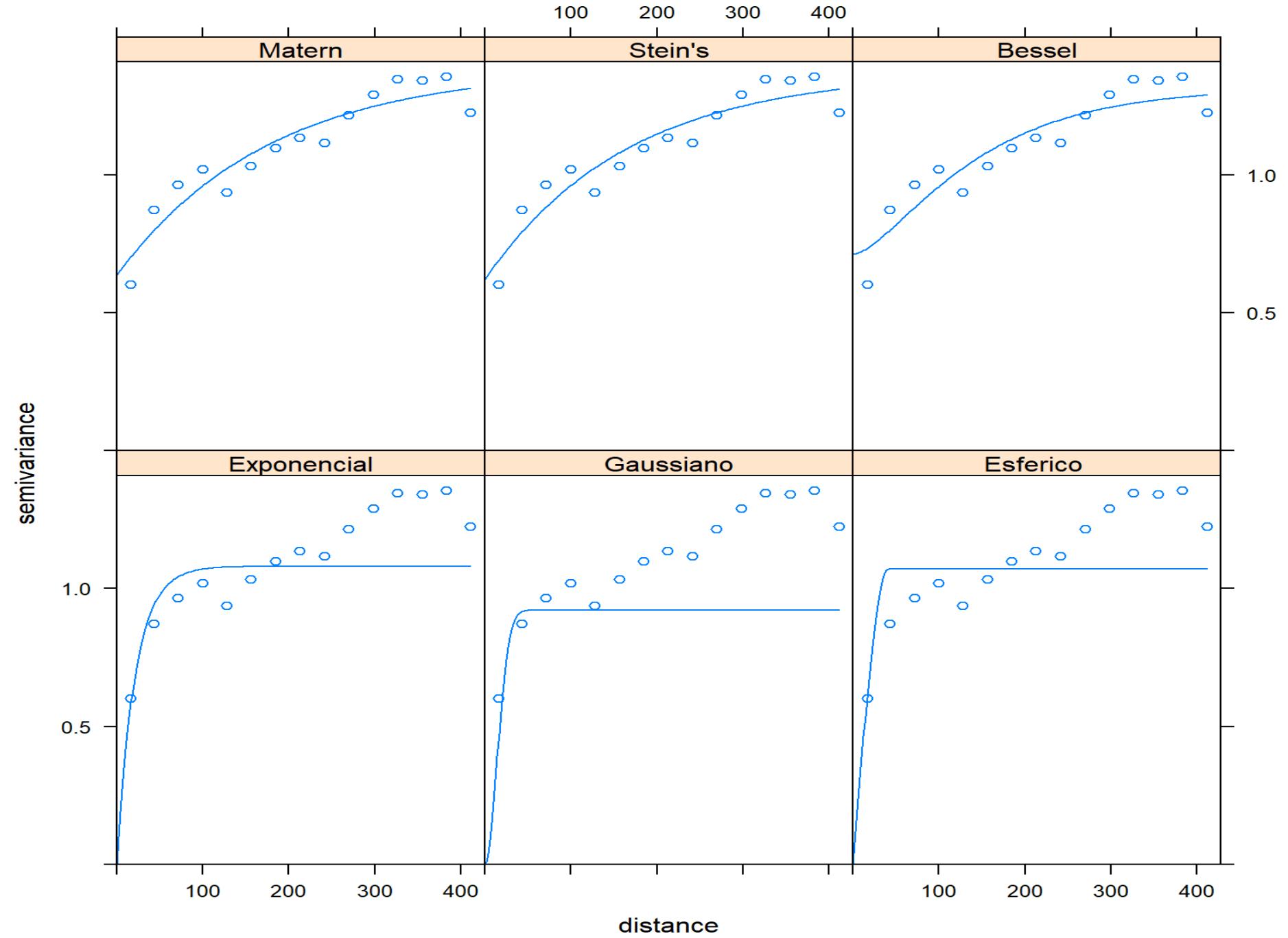
Semivariograma Empirico



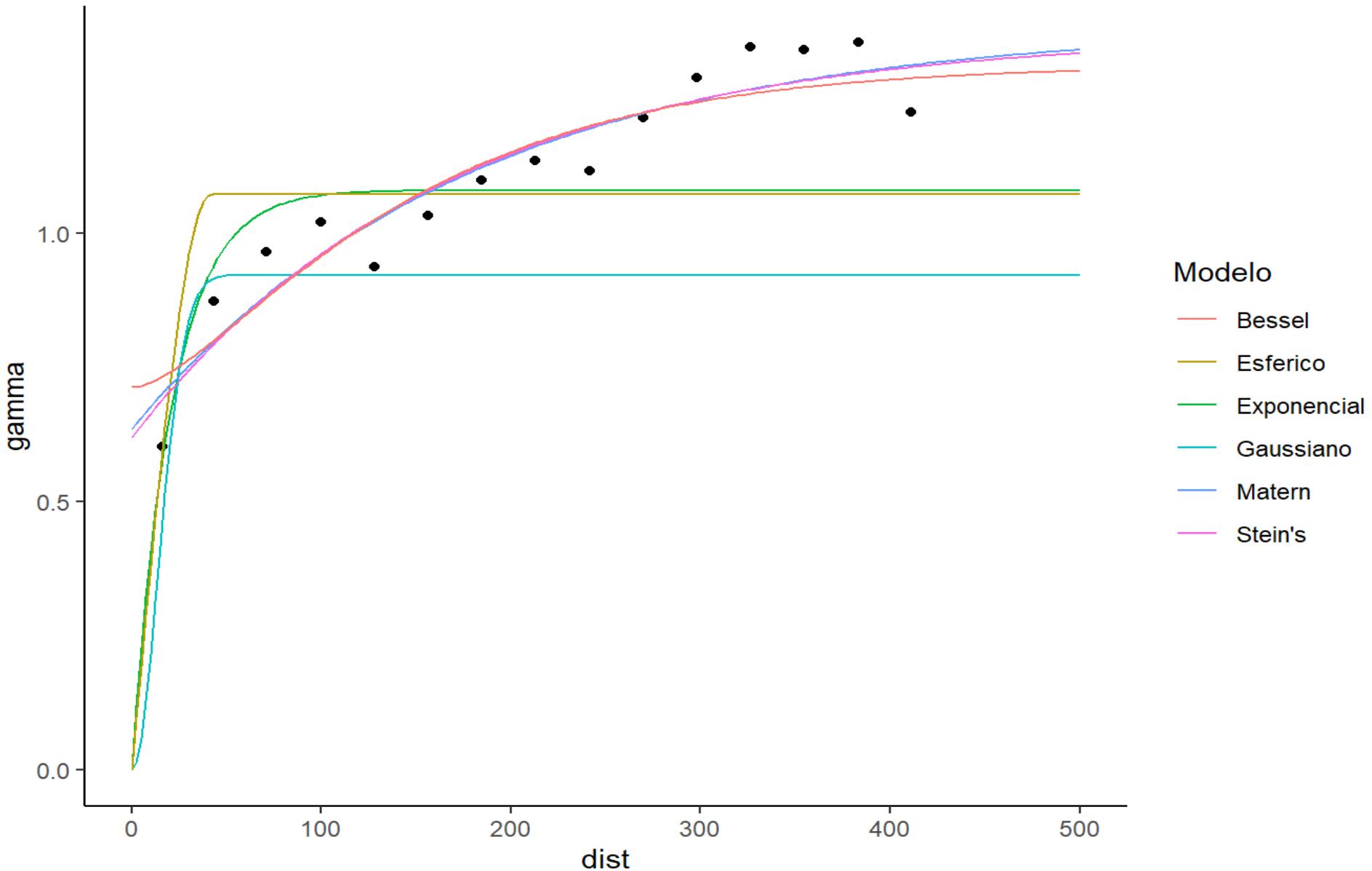
EXAMPLE

DGBS measurement variable - Variograms





EXAMPLE



REFERENCES

Cressie, N. "Statistics for spatial data", Wiley New York, 1992.

Cressie, N. and Wikle, K. "Statistics for spatio-temporal data", Wiley Series in Probability and Statistics, 2011

https://spatialanalysis.github.io/lab_tutorials/1_R_Spatial_Data_Handling.html

https://spatialanalysis.github.io/lab_tutorials/4_R_Mapping.html

<https://rpubs.com/robertrespa/404095>

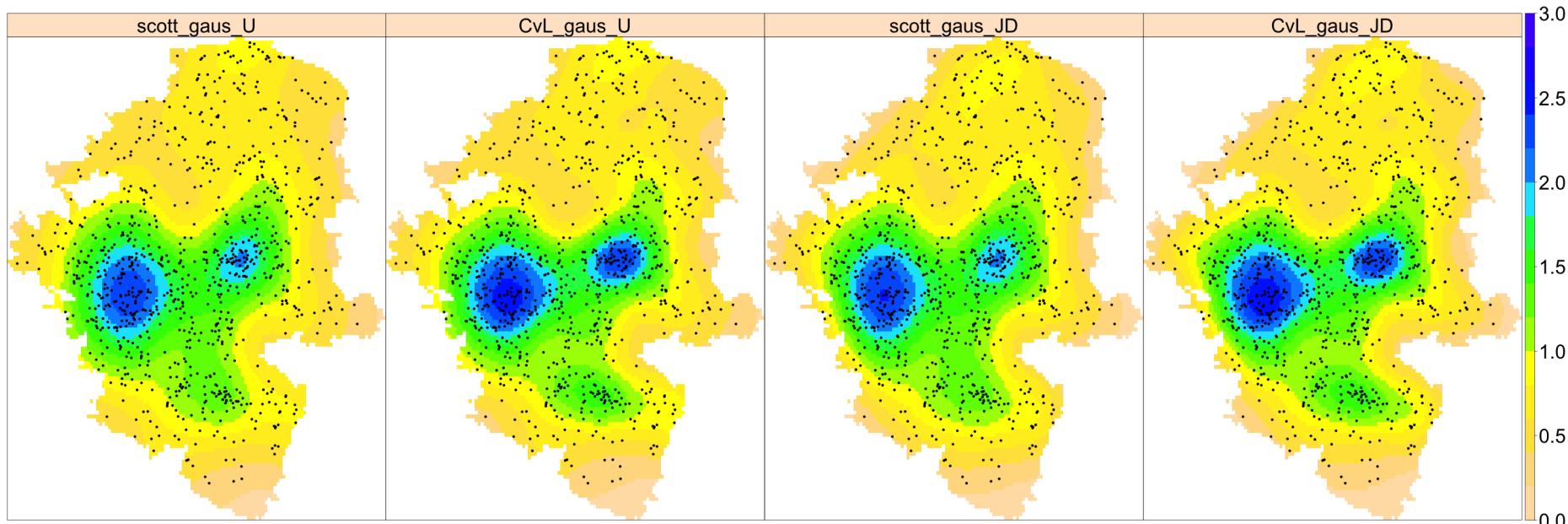


MODELOS PATRONES DE PUNTO (Procesos de Punto)

Los datos en forma de conjunto de puntos distribuidos irregularmente dentro de una región del espacio surgen en muchos contextos diferentes; por ejemplo, localizaciones de incendios forestales, delitos, árboles en un bosque, nidos en una colonia de cría de pájaros, ubicación de núcleos en una sección microscópica de tejido, depósitos de oro mapeados en un estudio geológico, estrellas en un cúmulo estelar, accidentes de tráfico, terremotos, llamadas de teléfonos móviles, avistamientos de animales o casos de una enfermedad rara.

Se llama patrón espacial de puntos a cualquier conjunto de datos de este tipo. La disposición espacial de los puntos es el principal foco de investigación. Son muchos los campos de la ciencia donde este tipo de estructuras son de interés; por ejemplo, en ecología, epidemiología, geociencia, astronomía, econometría e investigación criminal. El análisis estadístico de la disposición espacial de los puntos puede revelar características importantes, como que los yacimientos de oro suelen encontrarse cerca de una gran falla geológica o que los casos de una enfermedad son más frecuentes cerca de una fuente de contaminación.

MODELOS PATRONES DE PUNTO (*Procesos de Punto*)



Estimación de la intensidad basada en funciones núcleo para los datos de Medellín (puntos negros) durante 2005. Las etiquetas de los nombres comienzan con el método de suavizado, seguido del núcleo utilizado y de la corrección de borde. Los valores de la intensidad indican el número de crímenes por 100 km². Se usan JD y U para indicar los estimadores de Jones-Diggle y con corrección uniforme para los bordes, respectivamente.

MODELOS PATRONES DE PUNTO (*Procesos de Punto*)

Los puntos en un patrón de puntos pueden tener todo tipo de atributos. Un estudio forestal podría registrar cada ubicación, especie y diámetro del árbol; un catálogo de estrellas puede dar sus posiciones en el cielo, masas, formas y colores; las ubicaciones de los casos de enfermedades pueden estar vinculadas a registros clínicos detallados. Esta información auxiliar adjunta a cada punto en el patrón de puntos se llama marca y en ese caso se habla de un **patrón de puntos marcado**.

Un proceso puntual espacial es un proceso estocástico cuyas realizaciones consisten en un conjunto numerable de puntos en el plano (patrón puntual). Heurísticamente, se trata de un conjunto de datos que se encuentra en una región concreta (o área de estudio).

Procesos de Punto → MODELADO

La teoría de procesos puntuales espaciales constituye la base para el análisis de eventos observados geográficamente a través de sus coordenadas (longitud, latitud) en un espacio bi-dimensional. Esta rama de los procesos puntuales pertenece al campo de la estadística espacial en conjunción con la de procesos estocásticos. De hecho, un proceso puntual espacial es un proceso estocástico cuyas realizaciones consisten en un conjunto numerable de puntos en el plano (patrón puntual). Heurísticamente, se trata de un conjunto de datos que se encuentra en una región concreta (o área de estudio).

Sea $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, $0 \leq n < \infty$ una realización (patrón puntual) observada de un proceso puntual simple (*i. e.* sin múltiples eventos por localización) y finito X en \mathbb{R}^2 en la región $W \subset \mathbb{R}^2$ y con la métrica (distancia) asociada $d(\mathbf{u}, \mathbf{v})$, siendo \mathbf{u} y \mathbf{v} dos localizaciones espaciales en W . En general, las realizaciones consisten en un conjunto numerable de puntos (llamados en muchas ocasiones eventos). Las Figs. [43.1](#) y [43.2](#) permiten ver algunos ejemplos de patrones puntuales. Para cualquier conjunto arbitrario $A \subset \mathbb{R}^2$, el cardinal de X viene dado por la función de conteo:

$$N(X \cap A) = \sum_{\mathbf{x} \in X} \mathbf{1}\{\mathbf{x} \in A\} < \infty,$$

donde $\mathbf{1}$ es una función indicadora.

Además, como establece la fórmula de Campbell ([Baddeley et al., 2015](#)), para cualquier función medible $f : \mathbb{R}^2 \rightarrow [0, \infty)$ se cumple que:

$$\mathbb{E} \left[\sum_{\mathbf{x} \in X} f(\mathbf{x}) \right] = \int_{\mathbb{R}^2} f(\mathbf{u}) \lambda(\mathbf{u}) d\mathbf{u}, \quad (43.1)$$

Procesos de Punto → MODELADO

donde $\lambda(\cdot)$ determina la **función de intensidad** de X y gobierna su distribución espacial. De hecho, $\lambda(\mathbf{u})$ proporciona el valor esperado de eventos por unidad de área en un entorno de $\mathbf{u} \in \mathbb{R}^2$. Teniendo en cuenta que $f(\mathbf{x}) = 1\{\mathbf{x} \in A\}$, se puede observar fácilmente la relación entre la función de intensidad $\lambda(\cdot)$ y la de conteo N :

$$\mathbb{E}[N(X \cap A)] = \int_A \lambda(\mathbf{u}) d\mathbf{u}.$$

Si la función de intensidad $\lambda(\cdot)$ es constante, *i.e.* $\lambda(\cdot) = \lambda$, se dice que el proceso X es homogéneo, mientras que, en caso contrario, se dice que es inhomogéneo; en este último caso, la distribución espacial varía a lo largo de la región soporte

En la práctica se suele observar solo una única realización y, por ello, es importante disponer de una estimación de $\lambda(\cdot)$ que represente fielmente la distribución espacial del proceso subyacente que ha generado el patrón observado. A continuación se exponen diferentes **estimadores no paramétricos de la intensidad**.

Procesos de Punto → MODELADO (Núcleos simples)

Dos estimadores no paramétricos de la función de intensidad ampliamente utilizados en patrones puntuales en \mathbb{R}^2 , basados en **funciones núcleo**, vienen dados por:

$$\hat{\lambda}_\sigma^U(\mathbf{u}) = \frac{1}{c_{\sigma,W}(\mathbf{u})} \sum_{i=1}^n \kappa_\sigma(\mathbf{u} - \mathbf{x}_i), \quad \mathbf{u} \in W \quad (43.2)$$

y

$$\hat{\lambda}_\sigma^{JD}(\mathbf{u}) = \sum_{i=1}^n \frac{\kappa_\sigma(\mathbf{u} - \mathbf{x}_i)}{c_{\sigma,W}(\mathbf{x}_i)}, \quad \mathbf{u} \in W, \quad (43.3)$$

donde κ_σ es una función de densidad de probabilidad en \mathbb{R}^2 con parámetro de suavizado (ancho de banda) σ , y

$$c_{\sigma,W}(\mathbf{u}) = \int_W \kappa_\sigma(\mathbf{u} - \mathbf{v}) d\mathbf{v}, \quad \mathbf{u} \in W$$

Procesos de Punto → MODELADO

43.2.1.1 Selección del parámetro de suavizado

Scott (1992) propuso elegir este parámetro a través de un regla un tanto naïf (llamada *rule of thumb*) de la forma:

$$(s_x n^{-1/6}, s_y n^{-1/6}),$$

para cada coordenada cartesiana x, y , donde s_x, s_y son las desviaciones típicas de las coordenadas (x, y) de los eventos. Este procedimiento es útil para análisis exploratorios. La función `bw.scott()` de `spatstat.explore` proporciona este estimador. Nótese que, en el caso de Scott, el parámetro de suavizado es, por construcción, un vector de dos componentes para suavizar ambas coordenadas cartesianas.

Cronie & Van Lieshout (2018) propusieron encontrar el parámetro óptimo minimizando:

$$CvL(\sigma) = \left(|W| - \sum_{i=1}^n \frac{1}{\hat{\lambda}_\sigma^*(\mathbf{x}_i)} \right)^2,$$

donde $|W|$ es el tamaño de la región W y $\hat{\lambda}_\sigma^*(\mathbf{x}_i)$ es un estimador de la intensidad sin corregir (bien sea la expresión (43.2) o la expresión (43.3), pero sin el término de corrección) evaluado en \mathbf{x}_i y con parámetro de suavizado σ . La idea de este estimador proviene de la fórmula de Campbell, ya que:

$$\mathbb{E} \left[\sum_{x \in X} 1/\lambda(x) \right] = \int_W (1/\lambda(x)) \lambda(x) du = |W|.$$

Para un patrón puntual \mathbf{x} , la función `bw.CvL()` de `spatstat.explore` calcula el parámetro de suavizado mediante el método de Cronie y van Lieshout (se denotará por Cronie–van Lieshout).

Procesos de Punto → MODELADO (Núcleos irregulares y Estimación por Voronoi) y MODELADO en Redes Lineales

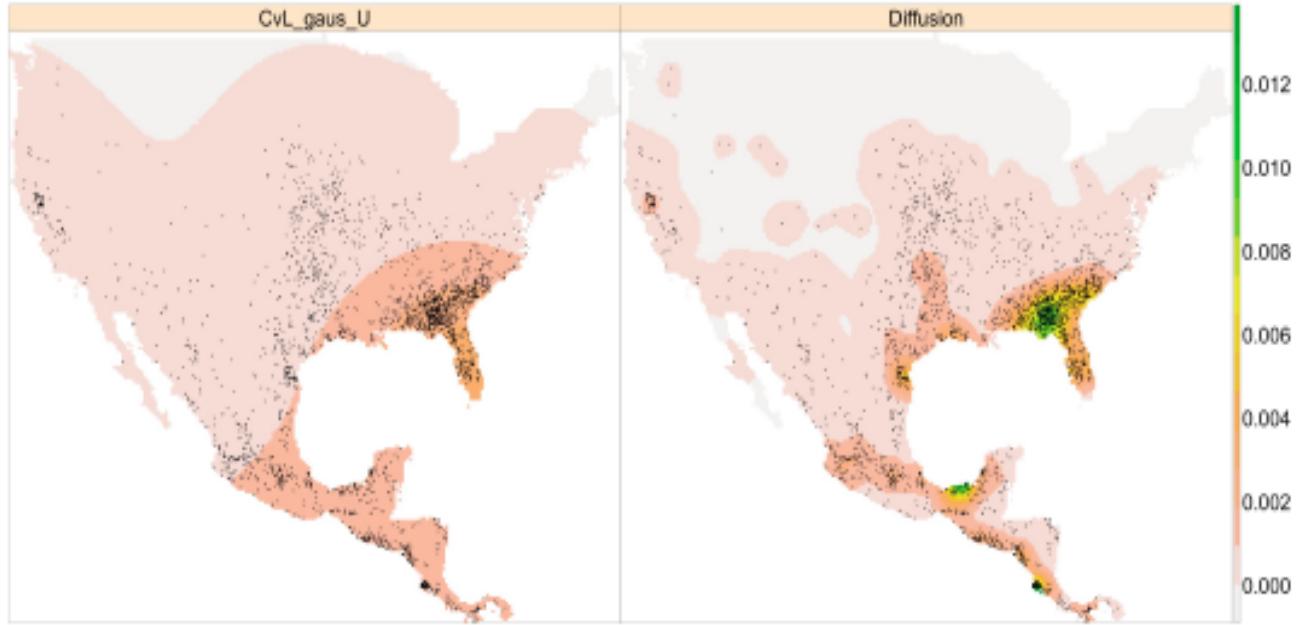


Figura 43.3: Estimación basada en función núcleo para incendios (puntos negros) en EE. UU. y Centroamérica (sin las islas) desde el 24 de febrero hasta el 3 de marzo de 2022. Izquierda: estimador con corrección uniforme con núcleo gaussiano. Derecha: estimador de difusión. El parámetro de suavizado fue obtenido con el criterio de Cronie–van Lieshout. Los valores de la intensidad son fuegos por 1.000 km².

Procesos de Punto → MODELADO (Núcleos irregulares y Estimación por Voronoi) y MODELADO en Redes Lineales

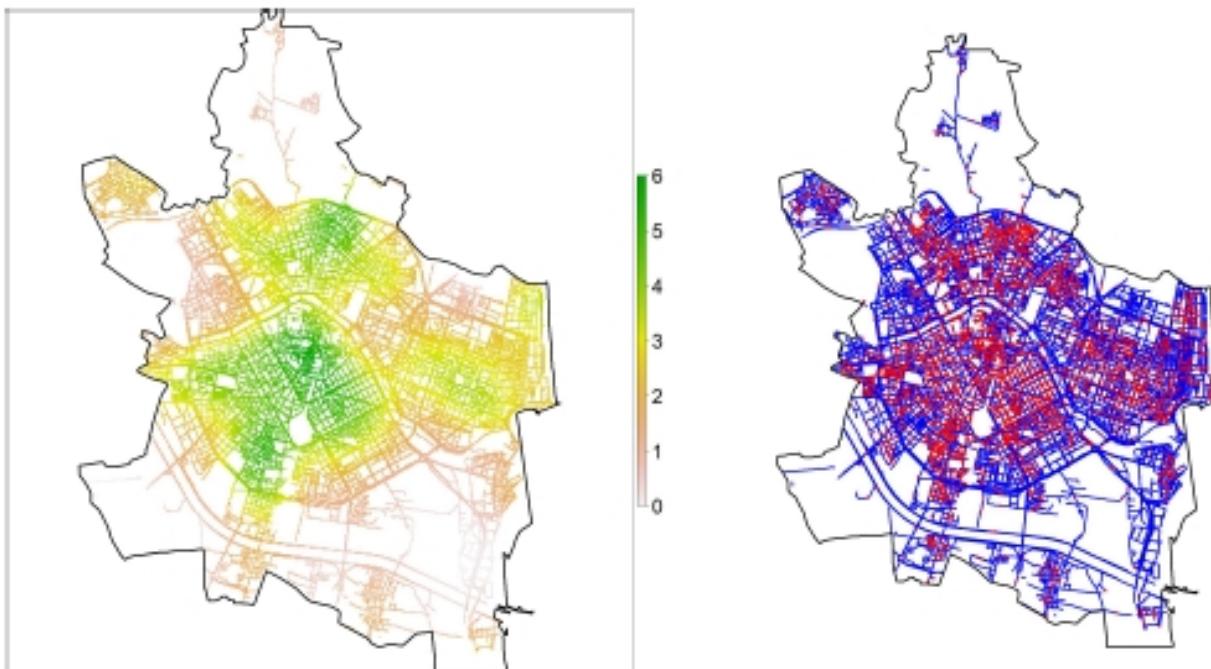


Figura 43.6: Intensidad estimada por función núcleo, usando el estimador de borde uniforme corregido (izquierda), para los datos de delitos (puntos rojos) en Valencia durante 2020 (derecha). Los valores de intensidad muestran el número de crímenes por km lineal.