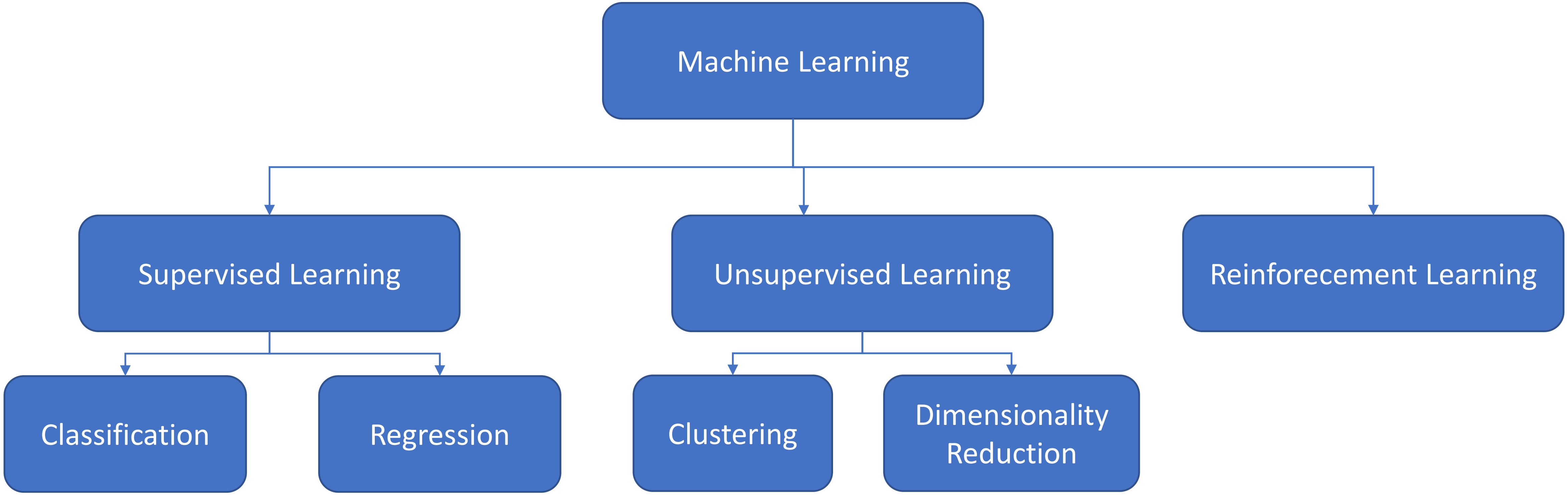
Modelamiento predictivo

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

Tipos de aprendizaje

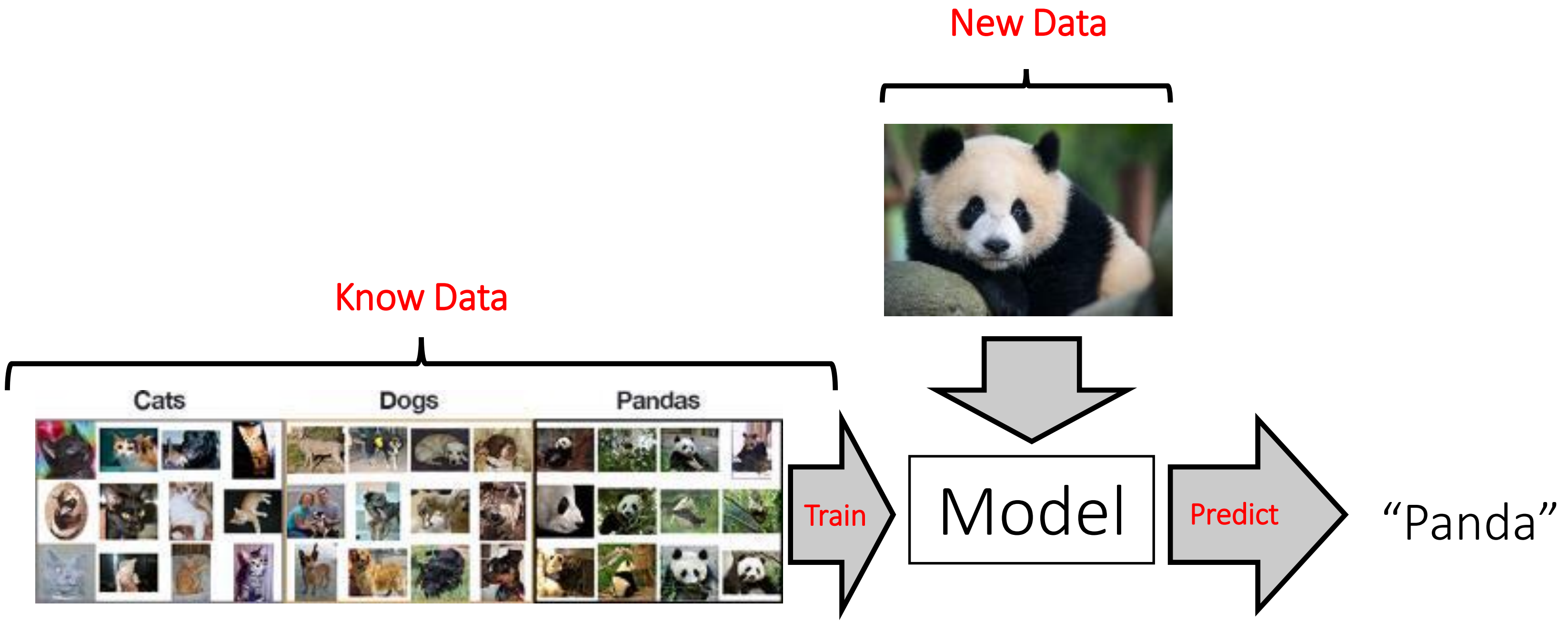


$$\hat{Y} = f(X, Y)$$

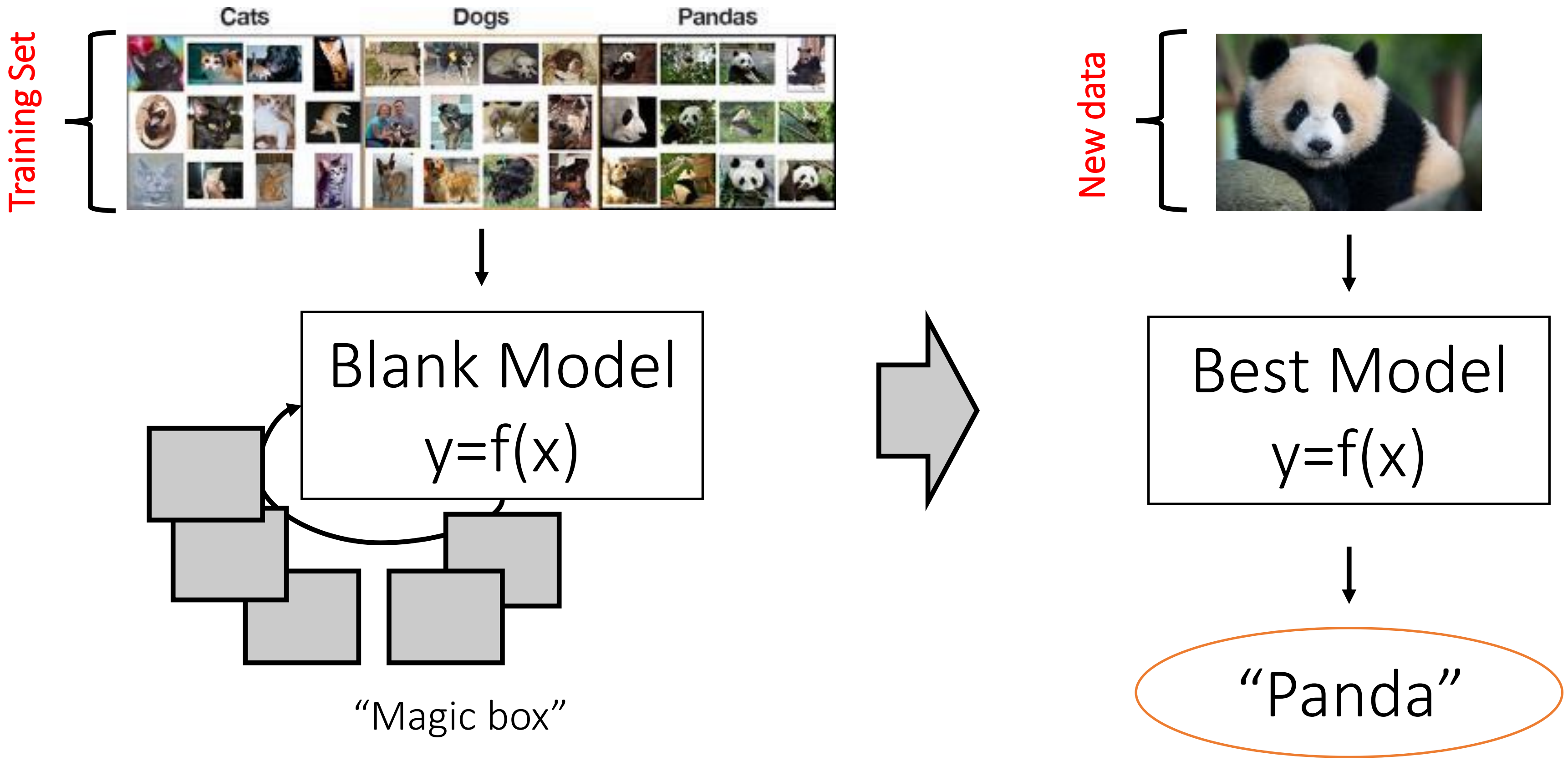
$$\hat{Z} = f(X)$$

$$\hat{X}_t = f(\hat{X}_{t-n})$$

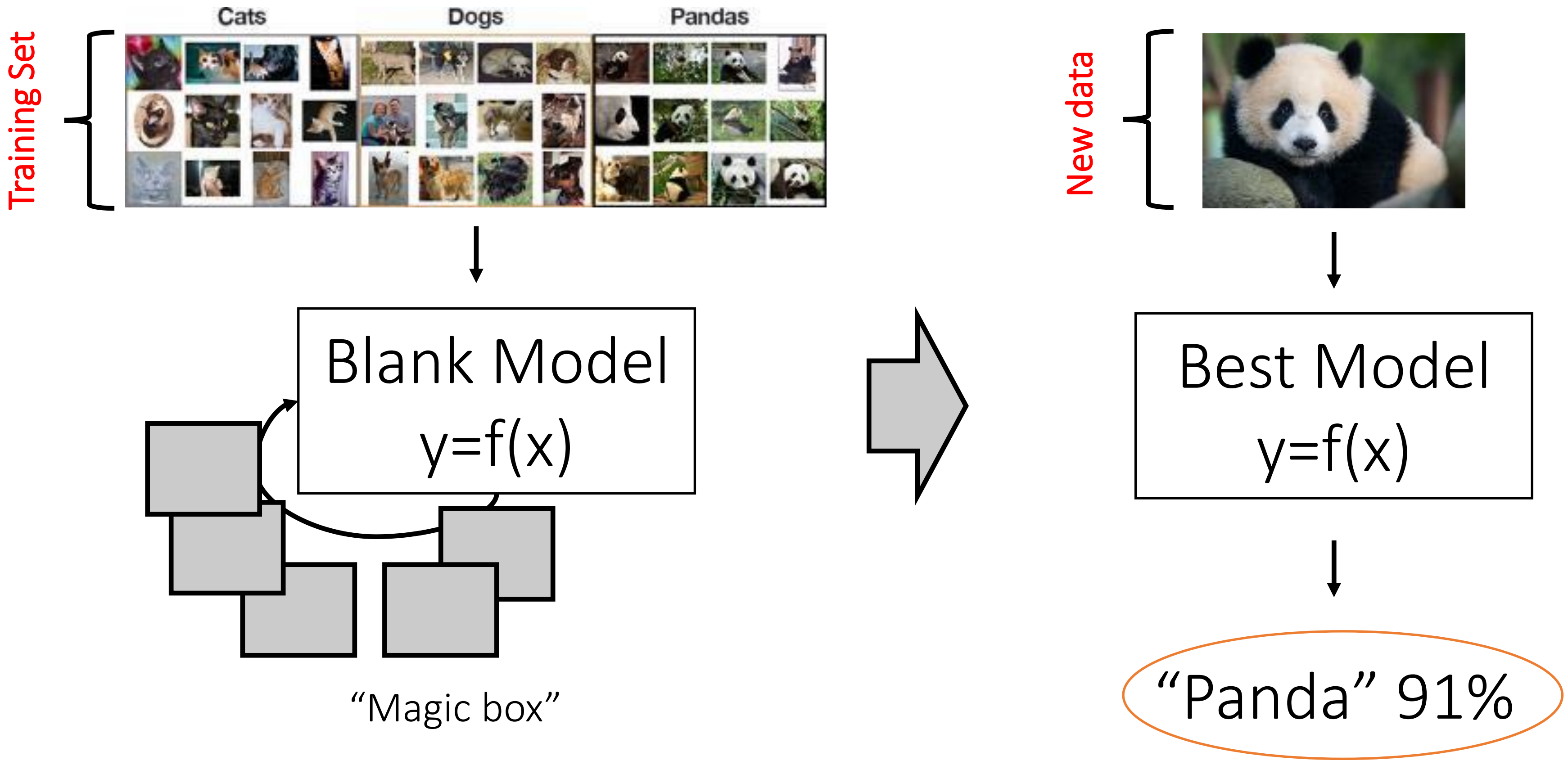
Idea general.



Enfoque determinístico

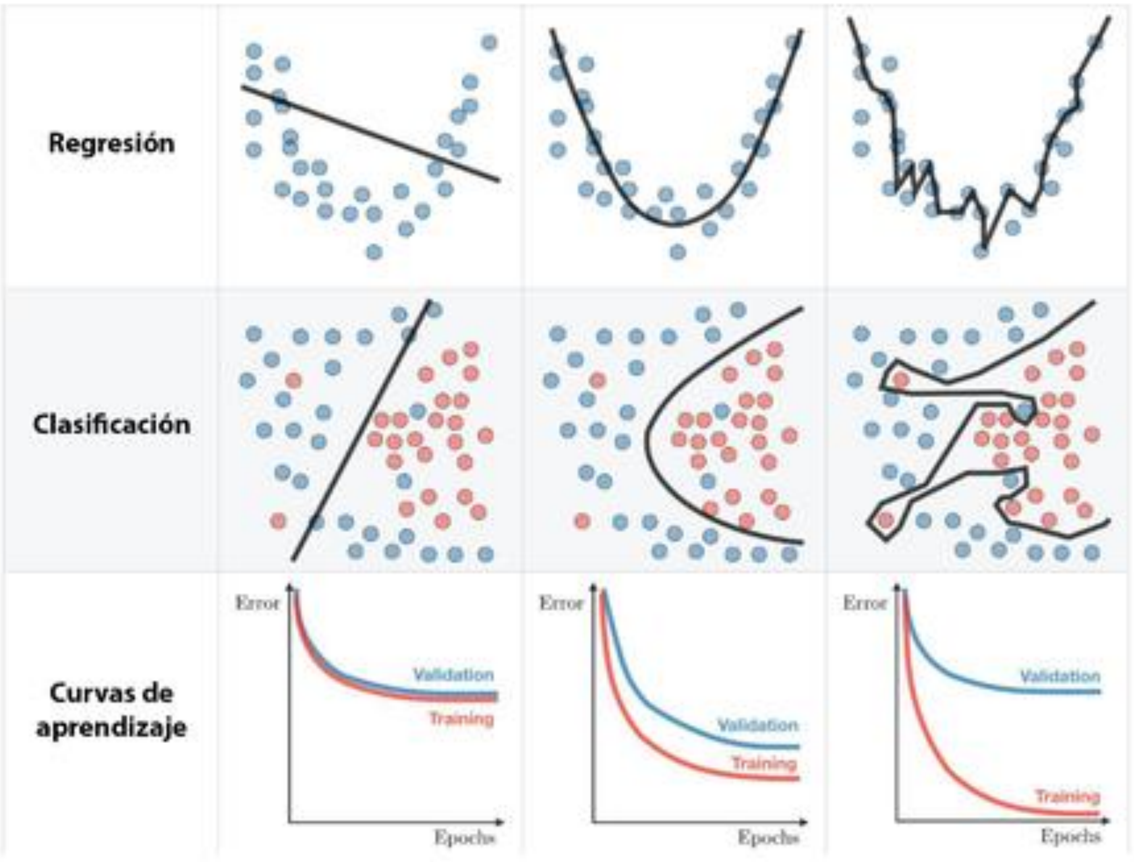


Enfoque probabilístico



Modelos de regresión versus clasificación

- Por la naturaleza de ambos problemas, se suelen evaluar de manera diferente
- Clasificación es binaria, pertenece o no a la categoría predicha
- Regresión permite calculo de métricas relativas



Buscamos construir, a través de un proceso delicado, el mejor $y = f(x)$ posible.

Aprendizaje

Minimizar una función de evaluación

$$S(M) = \sum_{i=1}^{N_{test}} d[f(x(i);M), y(i)]$$

Sum over examples

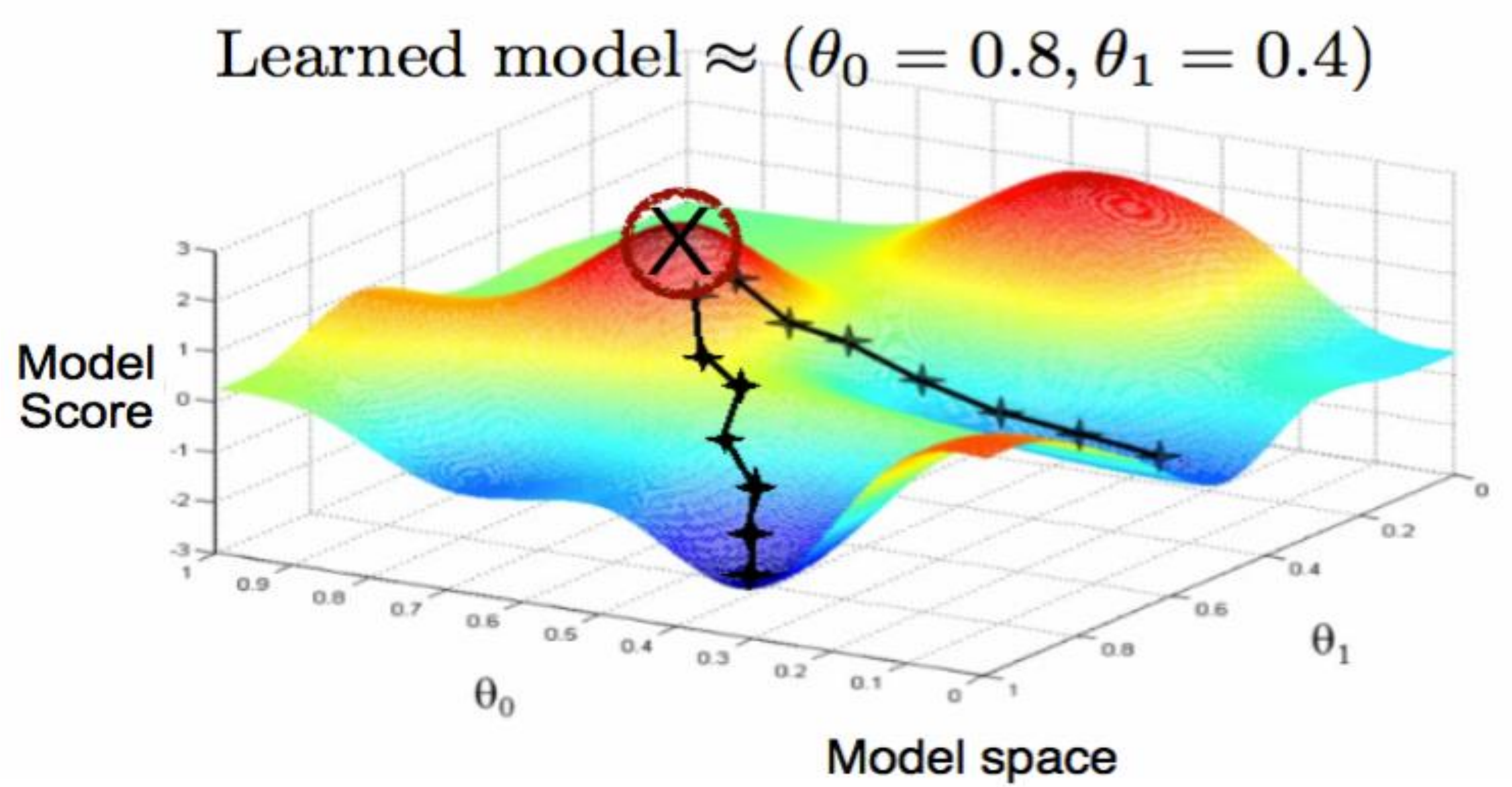
Distance between predicted and true

Predicted class label for item *i*

True class label for item *i*

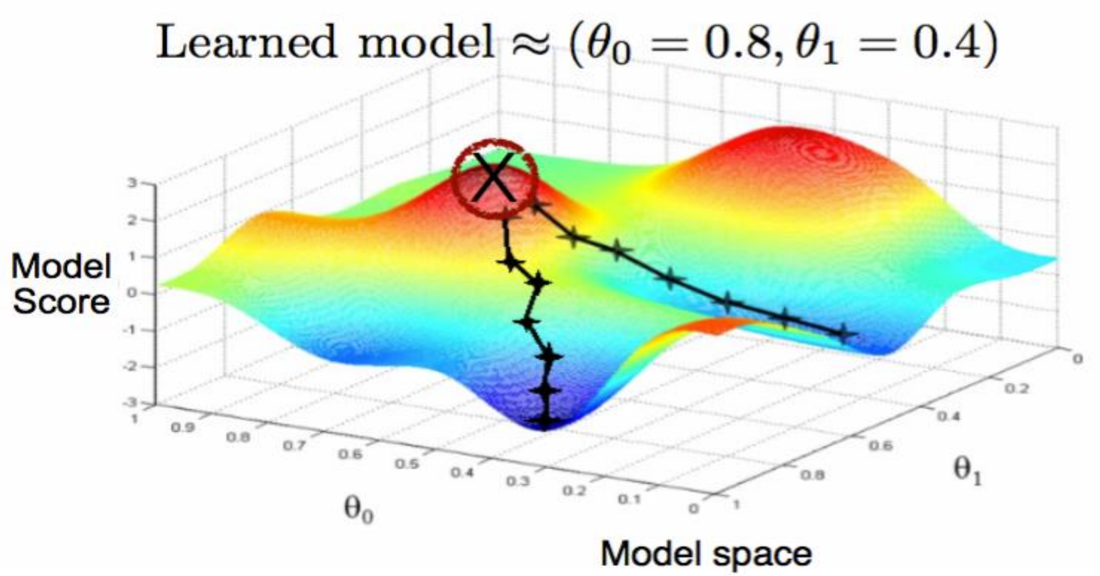
Algoritmo de búsqueda

Busca el conjunto de parámetros que maximicen la puntuación del modelo.



Algoritmos de regresión o clasificación

- Regresiones
- K nearest neighbors
- Clasificadores bayesianos
- Arboles de decisión
- Support Vector Machines
- Neural networks
- Deep neural networks



Modelamiento predictivo

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

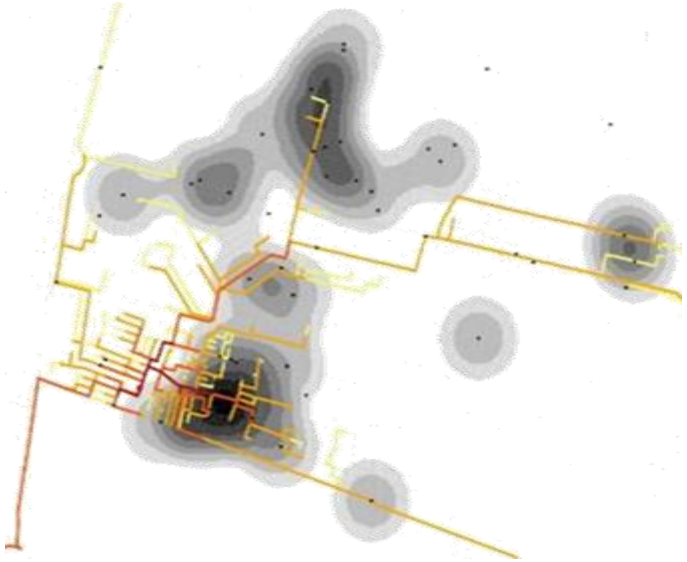
Inferencia Espacial

Dr. Raimundo Sánchez
raimundo.sanchez@uai.cl
@raimun2

Modelamiento espacial

La inferencia estadística convencional se basa en dos supuestos fundamentales

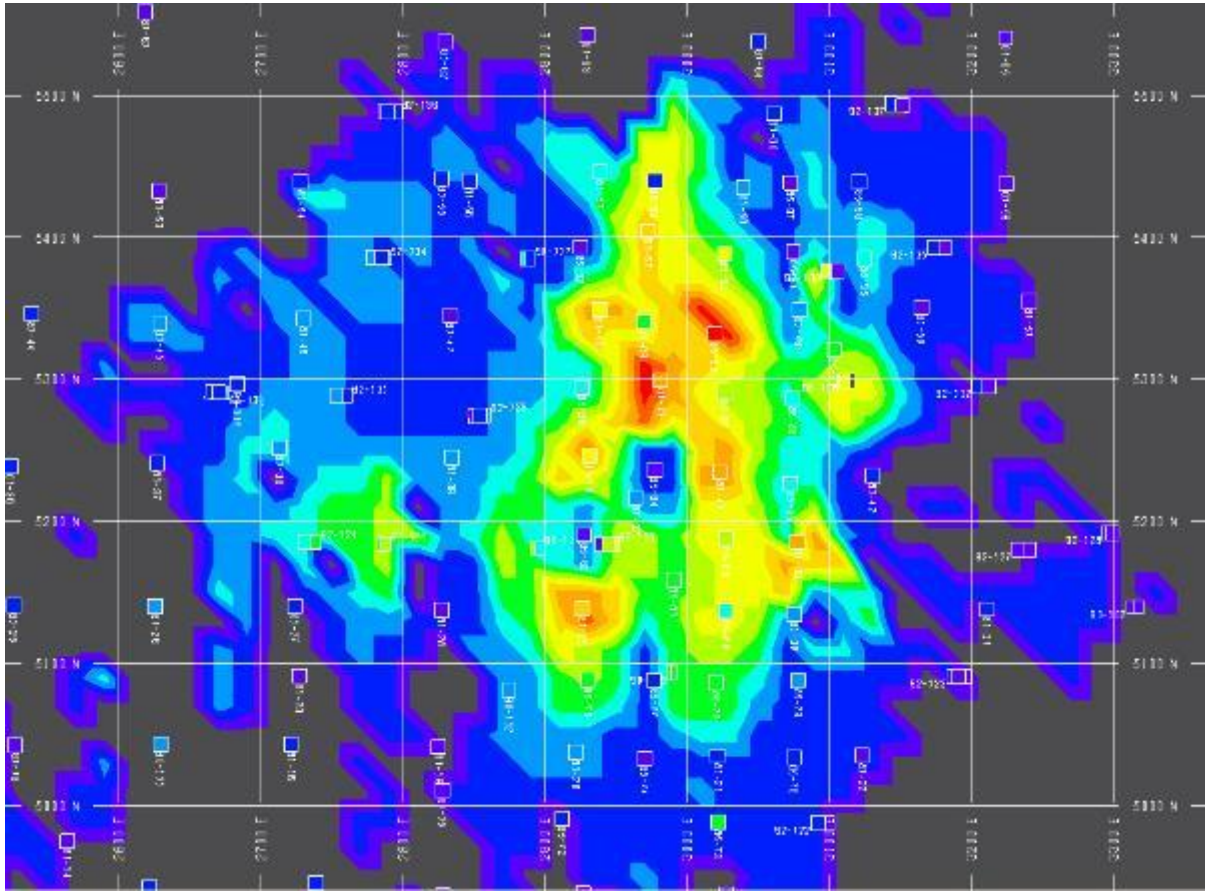
- Los valores de una variable son independientes unos de otros
- Los valores de una variable se distribuyen de forma aleatoria



Ley de Tobler!

Continuidad espacial

- Los datos geoespaciales no siempre se obtienen para todo el territorio
- En muchos casos se recurre a muestreo, o a medición en algunas ubicaciones especificas del espacio
- En caso de discontinuidad espacial, seria necesario interpolar valores en locaciones sin datos.

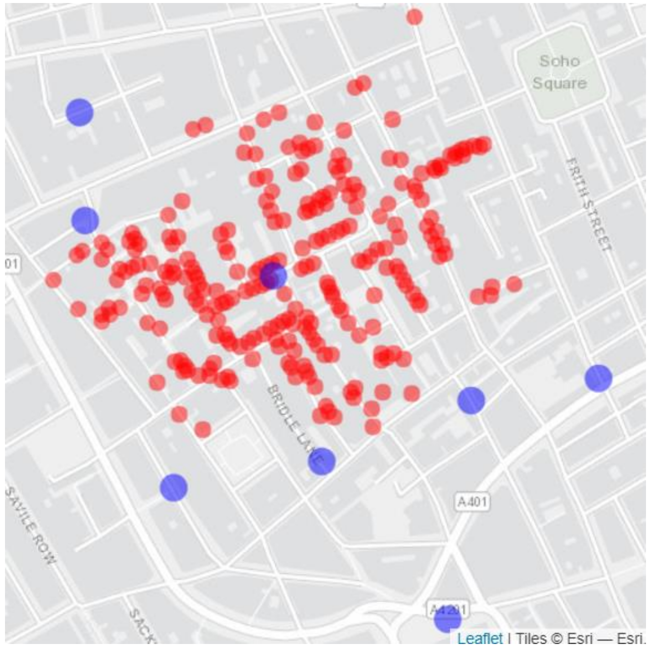


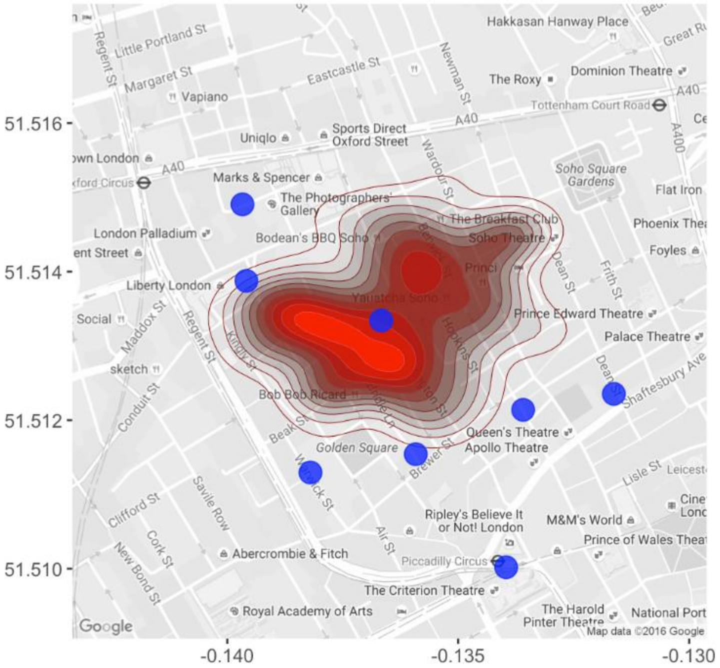
Enfoques de inferencia espacial

- Modelos de densidad
- Modelos de interpolación univariada
  - Base radial o ponderada por distancia
  - Kriging
- Métodos basados en covariables
  - Co-Kriging
  - Regresiones espaciales

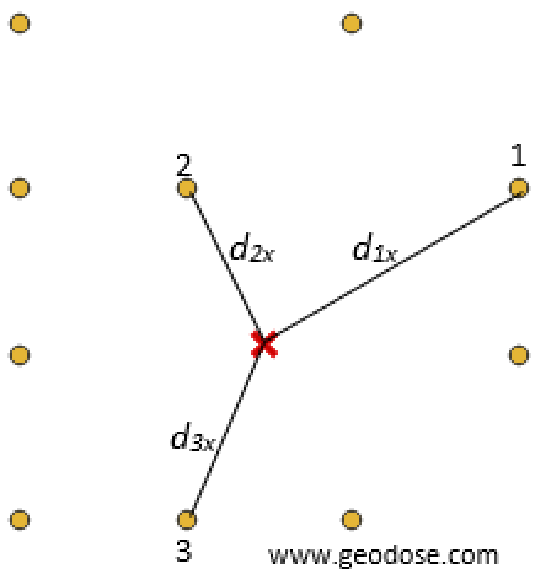
Kernel Density

- Construye distribuciones de probabilidad alrededor de los datos utilizando el método densidad de kernel
- Distribuciones 2D generan un campo de valores representando el espacio continuo
- Permite detectar Hotspots de fenómenos espaciales
- Apropiado cuando se busca inferir la densidad de un fenómeno en el espacio



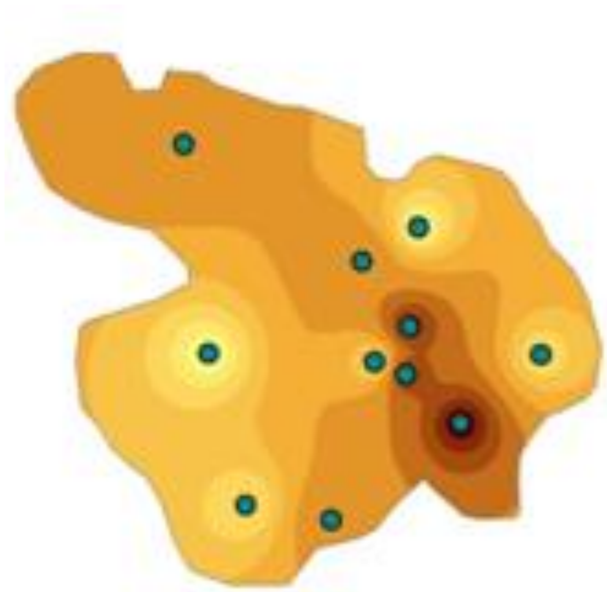


Interpolación ponderada por distancia



- Selecciona los puntos que se encuentren dentro de un radio
- Calcula la distancia
- Pondera los valores de los puntos cercanos inversamente por la distancia
- Parámetro Alpha indica el peso que se le da a la distancia

$$\hat{z}_S = \frac{\sum_{i=1}^{N} \frac{z_i}{d_i^{\alpha}}}{\sum_{i=1}^{N} \frac{1}{d_i^{\alpha}}}$$

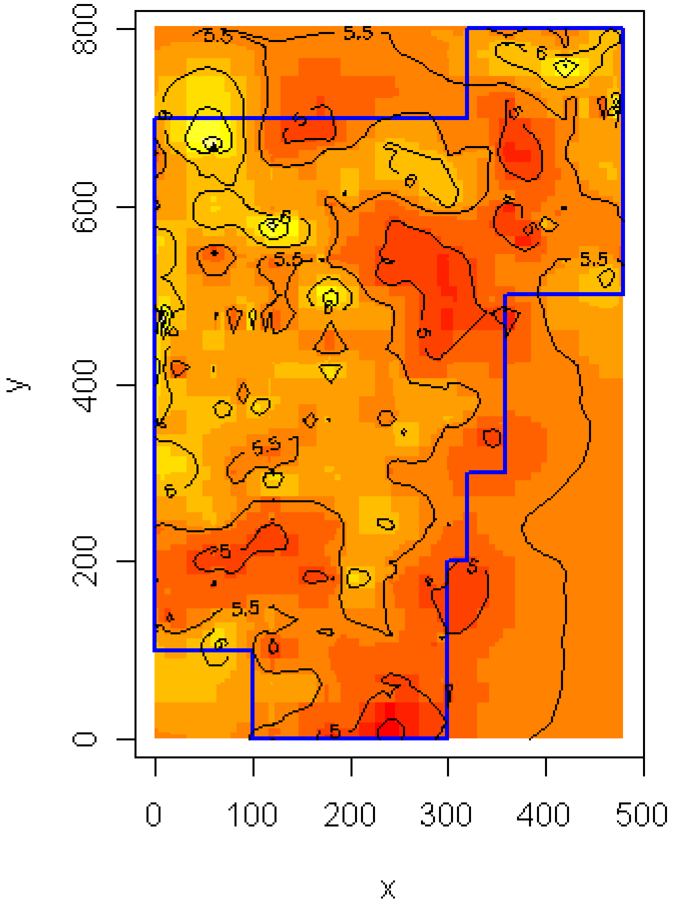


Kriging (y cokriging)

Pasos para ejecutar kriging

1. Calculo del variograma empírico (o correlograma)
2. Ajuste y selección de un modelo de variograma teórico
3. Optimizar los pesos w utilizando el variograma teórico ajustado, es decir, kriging (o cokriging).
4. Predicción de los valores en los lugares de interés

$$\hat{z} = \sum_{i=1}^{n} w_i z_i \qquad \sum w_i = 1.$$

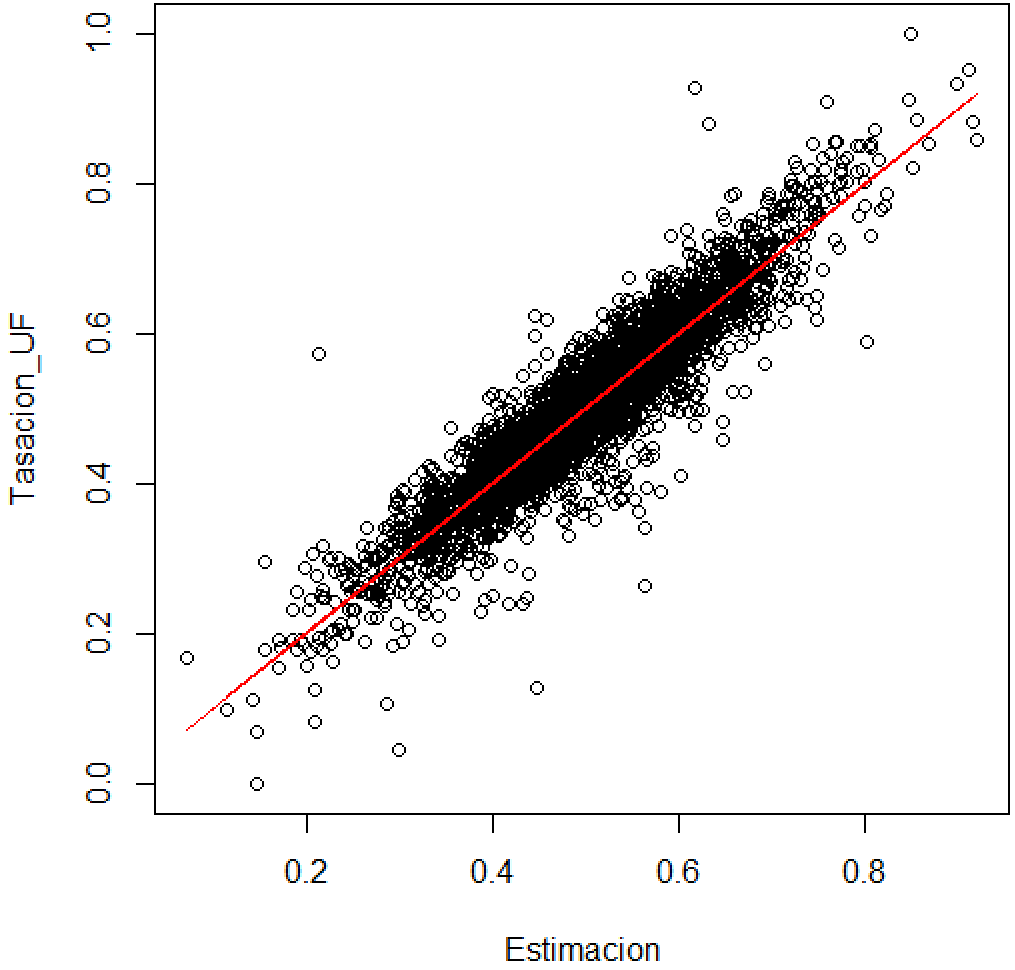


Regresiones

- Se calcula como una función de ajuste lineal
- Determinada por la minimización de los cuadrados entre cada observación y su estimación
- Coeficientes óptimos asignados a cada variable explicativa.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Valor propiedades vs estimación



Regresión espacial

Modelo de error espacial: incorpora efectos espaciales mediante un factor de error:

$$y = x\beta + \varepsilon$$
$$\varepsilon = \lambda W\varepsilon + \xi$$

Donde:

- ε es un vector de factores de error, ponderado con una matriz de pesos espaciales W
- λ es el coeficiente de error espacial
- ξ es un vector de errores sin autocorrelación

Modelo de lag espacial: incorpora efectos espaciales incorporando variables de lag espacial:

$$y = \rho Wy + x\beta + \varepsilon$$

Donde:

- Wy es la VD con lag espacial para matriz de pesos espaciales W
- ε es un vector de factores de error
- ρ es el coeficiente de lag espacial

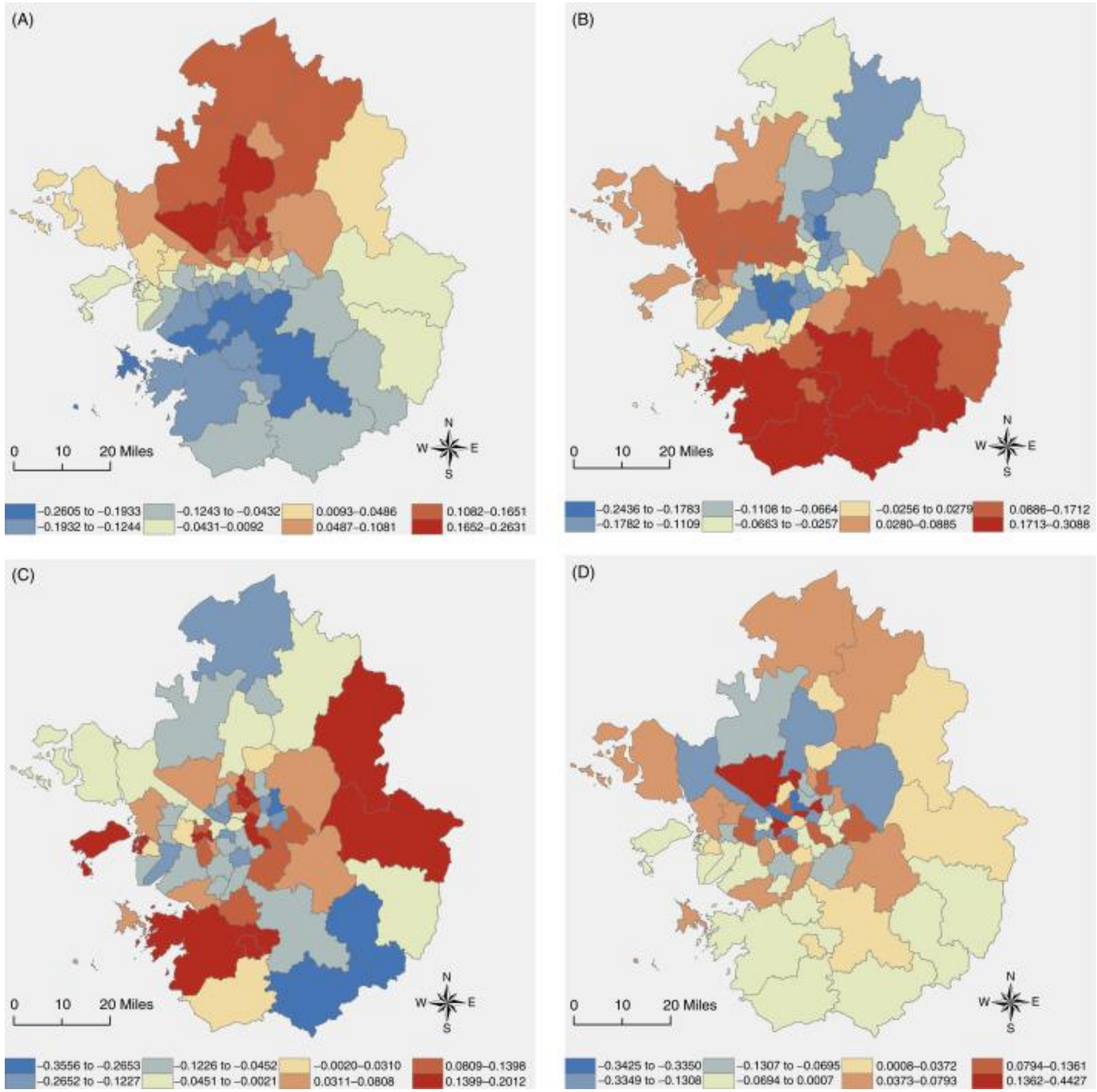
Pasos para regresión espacial

1. Definir un criterio de vecindad (contigüidad, distancia, etc) y asignar pesos a estas relaciones espaciales
2. Chequear autocorrelación espacial de la variable dependiente
3. Determinar tipos de error espacial y escoger el tipo de regresión espacial correspondiente
4. Estimar un modelo de regresión espacial y un OLS

Evaluación modelos espaciales

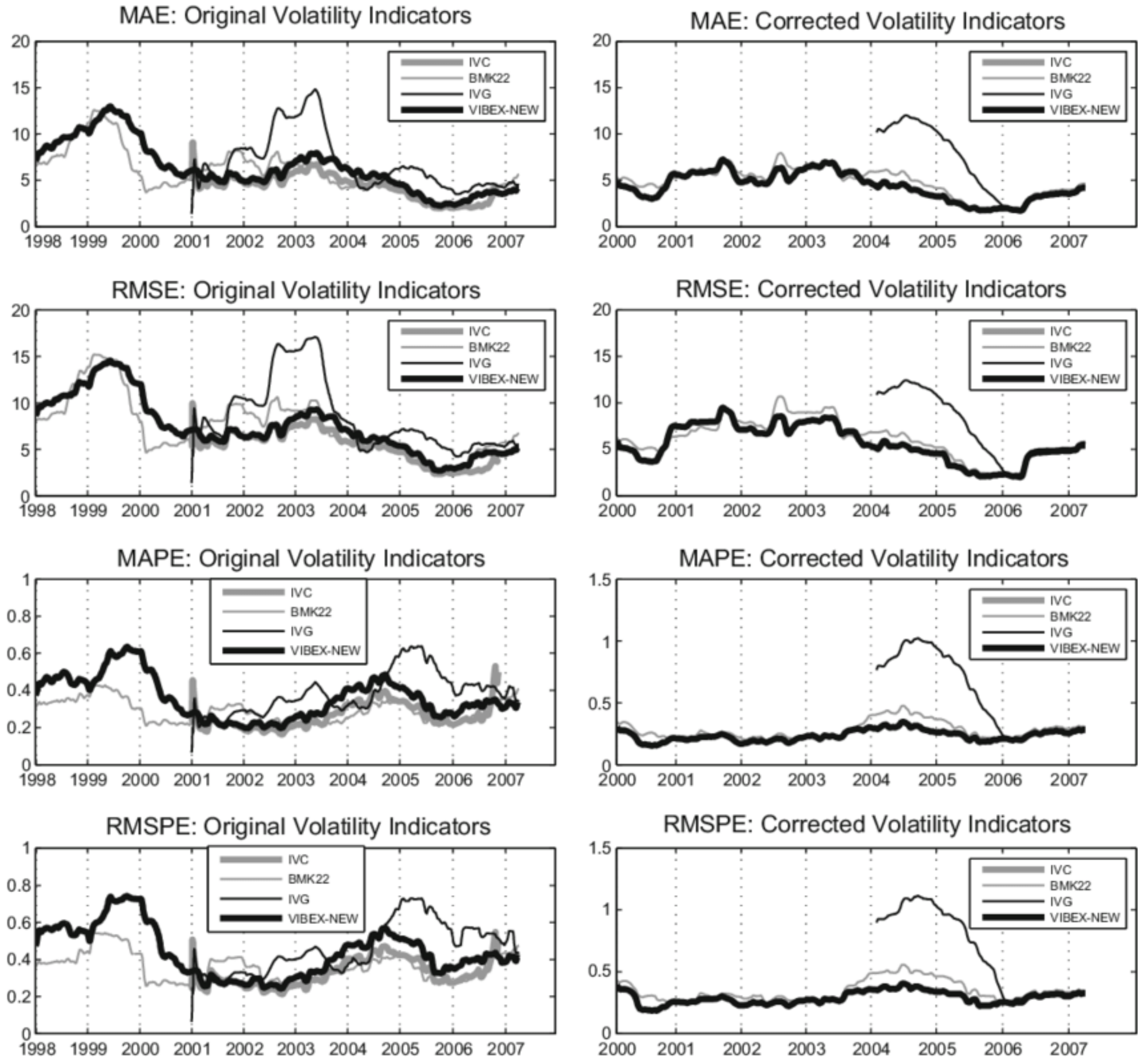
- Autocorrelación espacial de los errores
- Calculo del índice de moran global y su p-valor

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\right)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$



Evaluación modelos de regresión

- R-cuadrado $$\frac{\sum_{t=1}^{T}(\hat{Y}_t-\overline{Y})^2}{\sum_{t=1}^{T}(Y_t-\overline{Y})^2}$$
- RMSE $$\sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i-y_i)^2}{n}}$$
- Sesgo $$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)}{n}$$
- MAE $$\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i|$$
- MAPE $$\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{|y_i|} * 100$$



Métricas comunes de desempeño de clasificación

- Zero-one loss:

$$S_{0/1}(M) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} I[f(x(i);M), y(i)]$$

$$\text{where } I(a,b) = \begin{cases} 1 & a \neq b \\ 0 & \text{otherwise} \end{cases}$$

- Squared loss:

$$S_{sq}(M) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} [f(x(i);M) - y(i)]^2$$

Matriz de confusión

Se concentra en la capacidad predictiva de un modelo, en lugar de la rapidez con la que se tarda en clasificar o crear modelos, escalabilidad, etc.

Confusion matrix		**Predicted Class**	
		No	Yes
Actual Class	No	True Negative	False Positive
	Yes	False Negative	True Positive

Métricas asociadas para la matriz de confusión:

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

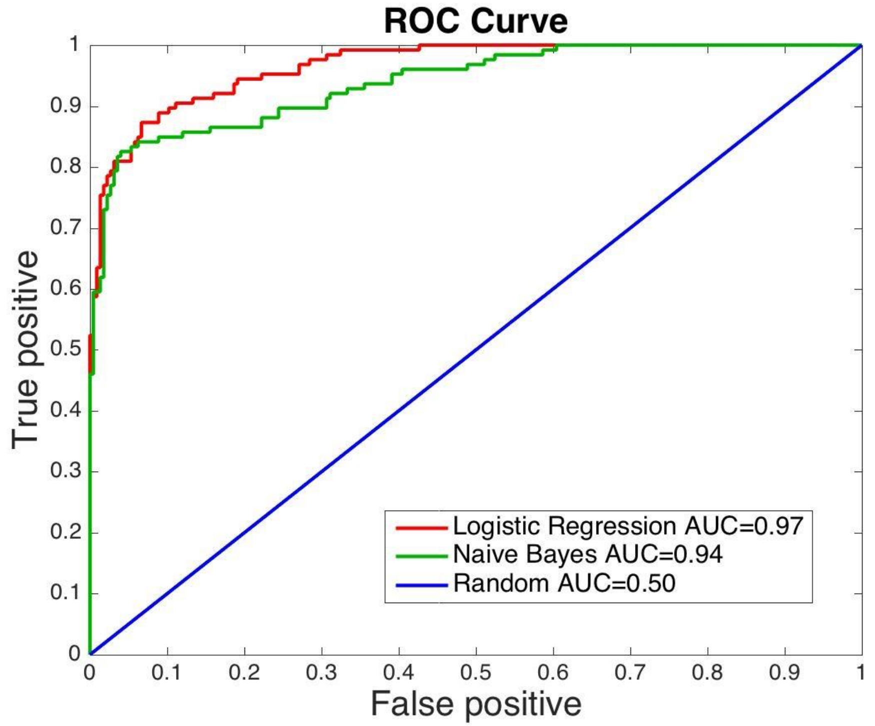
$$\text{F1-score} = \frac{2 * TP}{2 * TP + FP + FN}$$

ROC curve

Desarrollada en la década de 1950 en teoría de detección de señales para analizar señales ruidosas
Caracteriza la compensación entre golpes positivos y falsas alarmas
La curva ROC traza la tasa de Verdaderos positivos (TP) en el eje y contra la tasa Falsos Positivos en el eje x para diferentes valores.

- Desempeño de cada clasificador es representado como un punto en la curva ROC.

- Al cambiar el umbral de algoritmo, distribución de muestras o matriz de costes, cambia la ubicación del punto, que genera la curva final



Inferencia Espacial

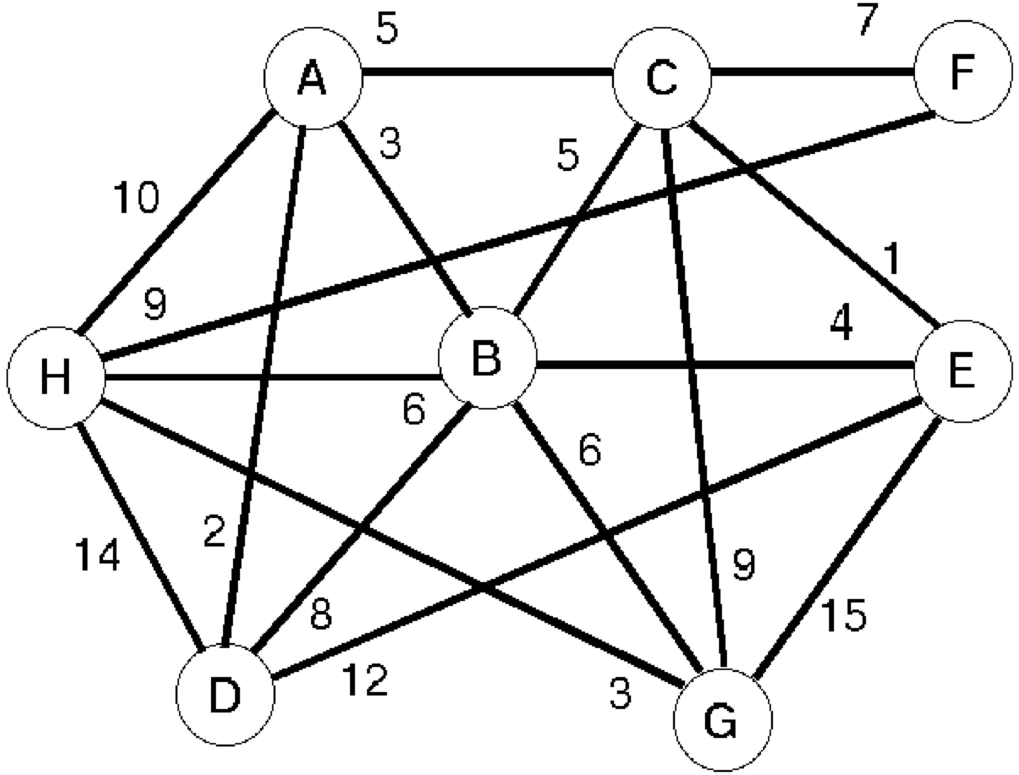
Dr. Raimundo Sánchez
raimundo.sanchez@uai.cl
@raimun2

Analisis de redes

Dr. Raimundo Sánchez
raimundo.sanchez@uai.cl
@raimun2

Análisis de redes

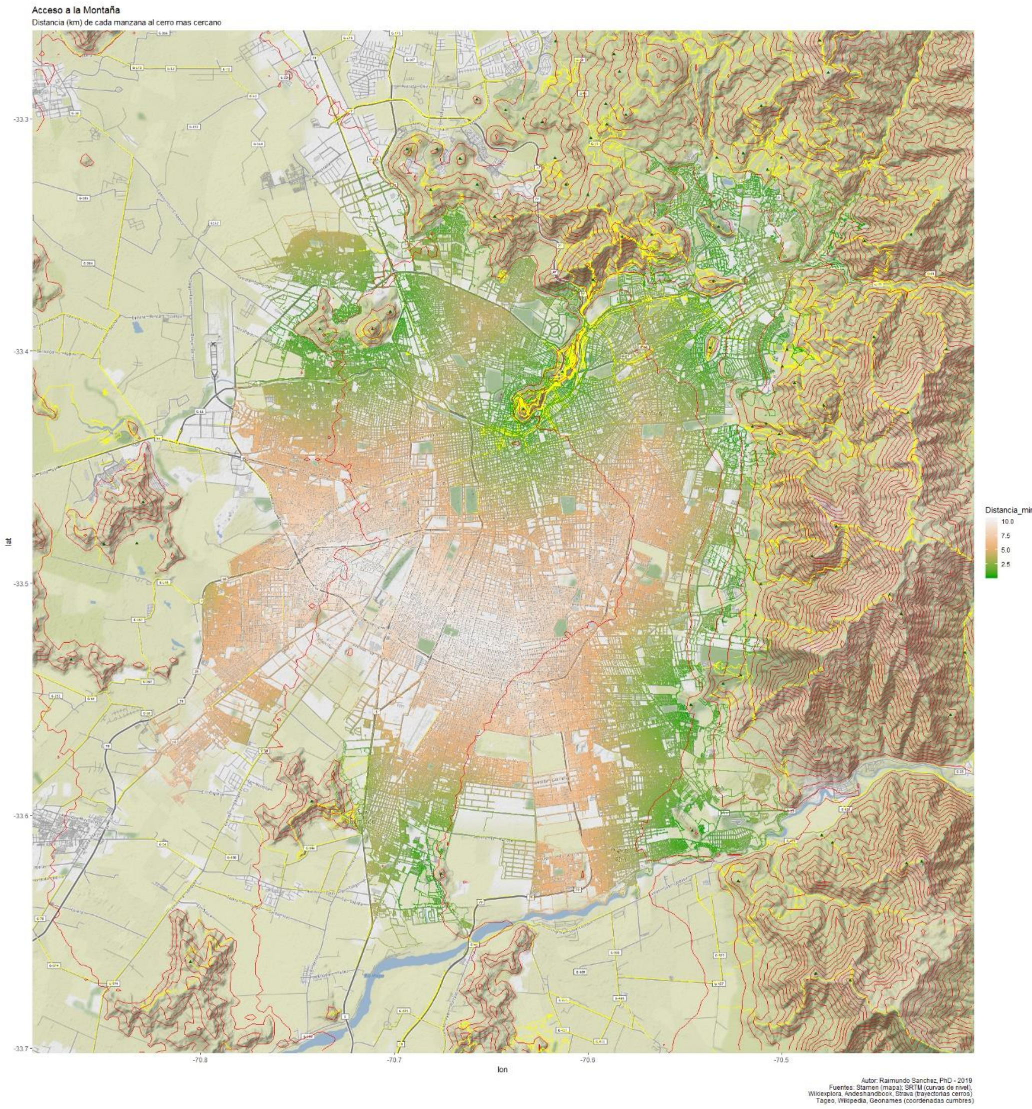
- Un grafo es una estructura matemática que describe un **conjunto de nodos** y las **relaciones entre ellos**
- El problema de encontrar el camino mas corto entre 2 nodos en un grafo con pesos no es trivial, pero el mundo de las matemáticas ya encontró una solución suficientemente buena (**Dijkstra, 1959**)
- Estos problemas suelen buscar minimizar la distancia (horizontal), pero a veces hay otras variables relevantes como terreno o elevación



Accesibilidad

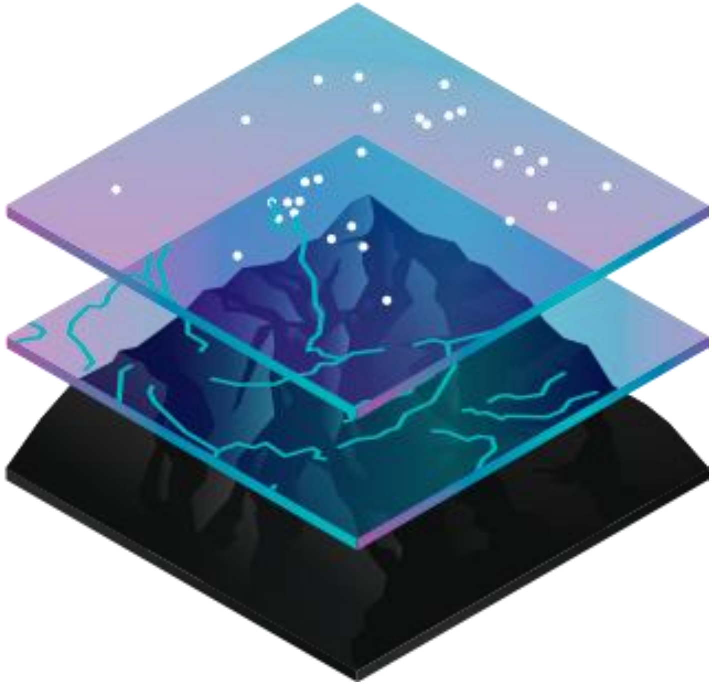
- La accesibilidad debe comprenderse como la facilidad con la cual una oportunidad (servicios, actividades, destinos, etc.) puede ser alcanzada.
- El acceso depende de 4 tipos de variables de decisión descritas por Geurs y van Wee (2004):
    - Land-use (Uso del suelo)
    - Transportation (Transporte)
    - Temporal (Relacionado al tiempo)
    - Individual (Individual)

- Cada una tienen sus propias restricciones de espacio-tiempo, llamadas impedancias.



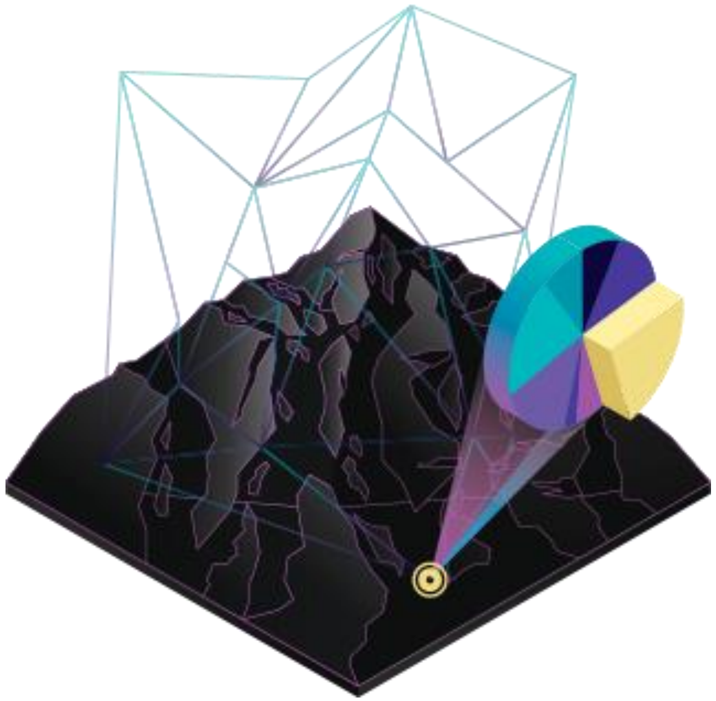
Uso de suelo

- Está compuesto por la oferta de las oportunidades, la cual contiene; sus tamaños, capacidades y ubicaciones (distribución).
- Considera tanto la oferta tanto como la demanda, y la competencia:
  - Ofertas de trabajo
  - Matrículas escolares



Transporte

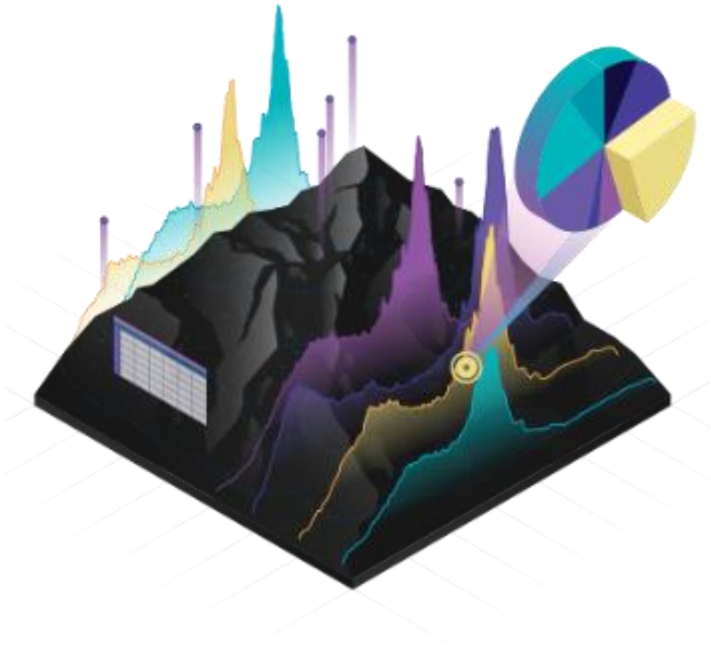
- Es el medio de transporte utilizado para alcanzar una oportunidad desde su origen:
  - Caminar
  - Transporte público
  - Automóvil, etc
- El acceso se relaciona al costo de alcanzar una oportunidad desde un origen al destino:
  - Tiempo de traslado
  - Distancia
  - Costo monetario
  - Conectividad



Temporalidad

- Es la disponibilidad en el tiempo de las oportunidades, es decir, el tiempo necesario para las actividades.

- Las impedancias temporales son:
    - Horarios de apertura o cierre
    - Horarios de trabajo o libre



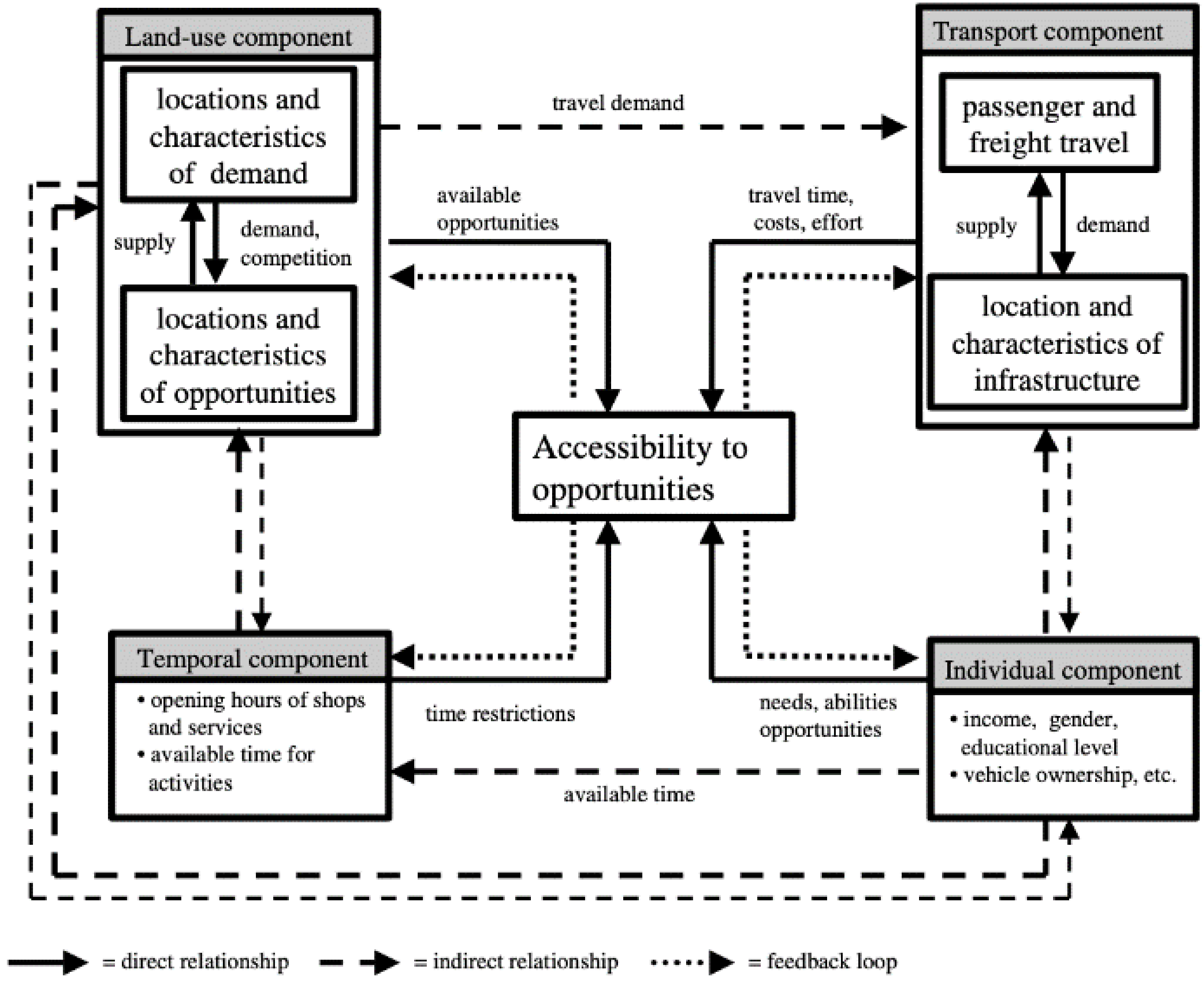
Individual

- Son las necesidades, habilidades y capacidades individuales de los demandantes.
    - Propósito (hacer, libre)
    - Percepciones.
    - Condición física.
    - Nivel educacional.

Integración de dimensiones

- En la práctica es difícil integrar todas las variables y restricciones a la vez.
- Todas las variables tienen su importancia en el cálculo de la accesibilidad
- Resultados dependerán de su planteamiento y pesos asignados.
- Mantener los modelos simples

K.T. Geurs, B. van Wee / Journal of Transport Geography 12 (2004) 127–140

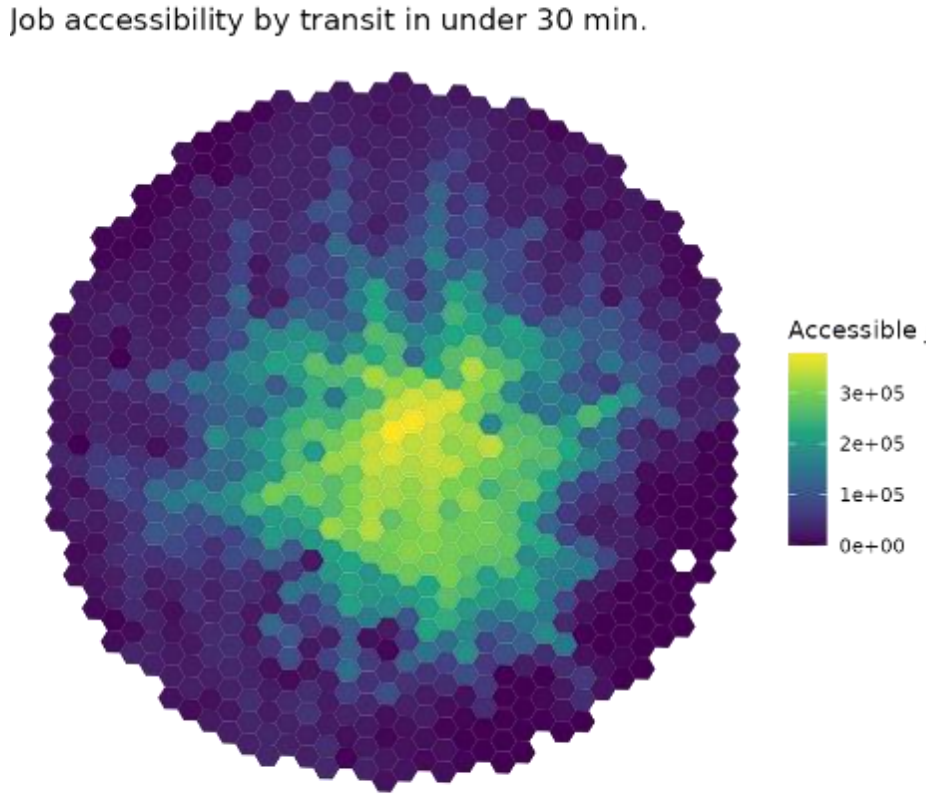


Modelos gravitacionales

Miden la accesibilidad una oportunidad (servicios, actividades, destinos, etc.)

Se puede abordar de diferentes maneras:

- Cost to k closest
- Cumulative opportunities



Impedancia

- Las funciones de impedancia se utilizan en los modelos de Gravity-based para el cálculo de la accesibilidad.
- Integran el acceso a un servicio con el costo de desplazamiento (distancia, tiempo, dinero, etc.).
- A mayor costo de desplazamiento genera influencias decrecientes sobre la accesibilidad.

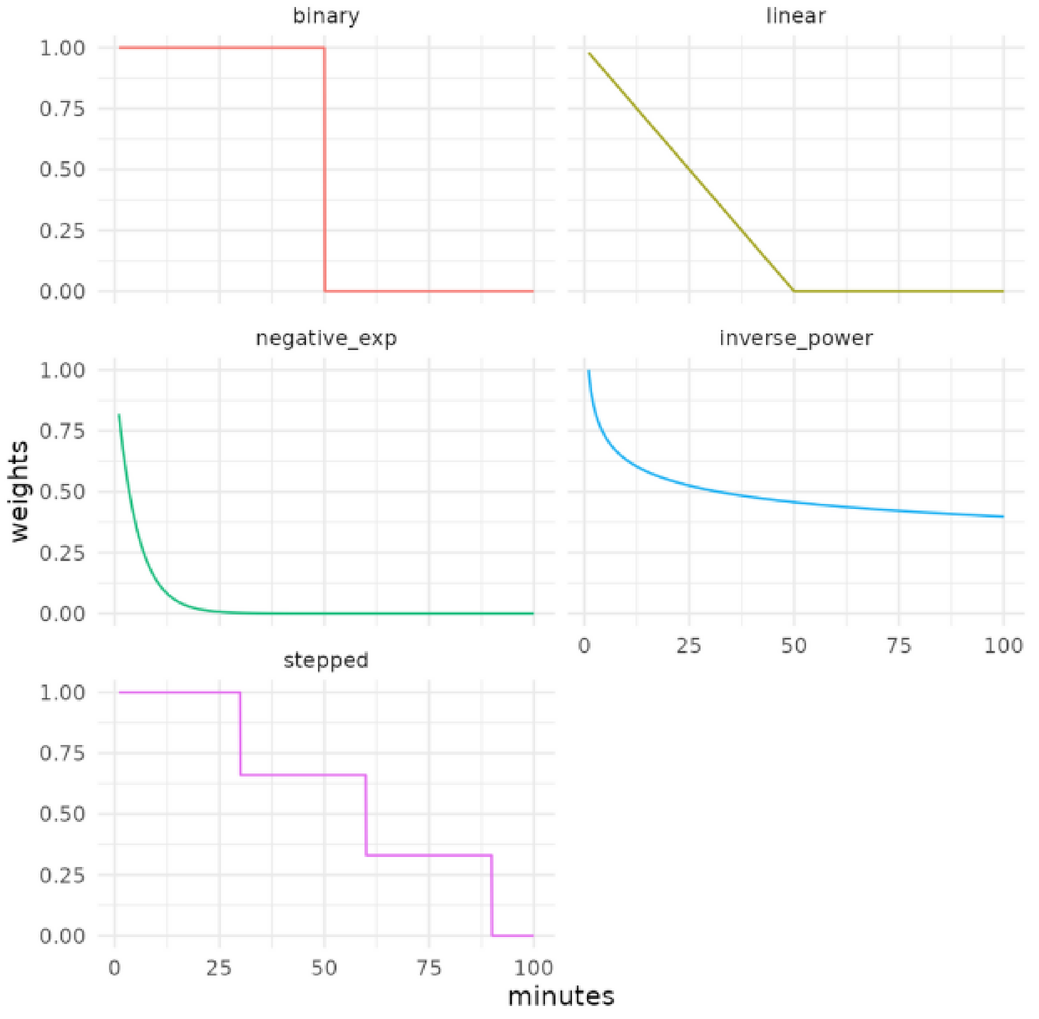


Imagen 10: Ejemplo gráfico de distintas funciones de decaimiento. (Pereira y Herszenhut, 2022)

Distancias (km)									
	Clientes								
Desde	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**
Fábrica	15	10	20	18	22	20	10	9	24
C1	0	12	10	12	25	24	20	21	30
C2		0	13	11	18	15	12	15	20
C3			0	7	22	23	22	25	28
C4				0	18	18	18	22	16
C5					0	7	17	23	12
C6						0	13	19	10
C7							0	9	15
C8								0	22

Isócronas

- Una isócrona se define como una línea en que algo ocurre o llega a la misma hora
- Se utilizan en la planificación del transporte y planificación urbana
- El mapa de isócronas muestra las áreas relacionadas con isócronos entre diferentes puntos
- Esto permite construir áreas de servicio de acuerdo a el tiempo que un usuario este dispuesto a viajar por este servicio.

Analisis de redes

Dr. Raimundo Sánchez
raimundo.sanchez@uai.cl
@raimun2