Módulo 4:
Mejorando la Exactitud

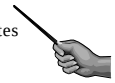
Carlos López Vázquez

carlos.lopez@ieee.org

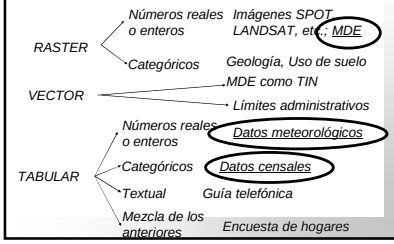
Parte II

Plan

- ✔ Introducción
- ✔ Revisión de herramientas estadísticas
- ■ Detectando problemas
- ■ Imputando valores ausentes
- ■ Ejemplos



Una posible categorización de datos



El cómo de la detección de outliers

- ■ Métodos tradicionales para el caso multivariado
  - ➤ Distancia de Mahalanobis

    $(x-\underline{x}) / \sigma <==> (x-\underline{x})^T C^{-1} (x-\underline{x})$

    ¿Cómo hallar C y $\underline{x}$? => Clásico, MCD, MVE, Hadi (1994), Rocke (1996), etc.
  - ➤ Análisis de Componentes Principales (*PCA*)

    Hawkins, 1974; López, 1994a,b, 1996, 1997
  - ➤ Otros métodos...

Mahalanobis de vuelta…

- ■ Si $d^2(x)=(\underline{x}-\underline{T})^T C^{-1} (\underline{x}-\underline{T}) > d_{crit}$ ➔ outlier
- ■ Depende de cómo se construyen C y $\underline{T}$ puede ser inapropiado si hay outliers (¡!)
- ■ Ej: Philips data

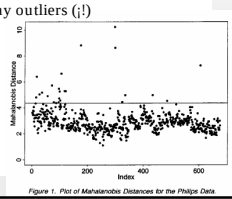
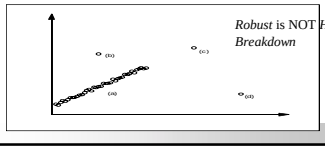


Figure 1. Plot of Mahalanobis Distances for the Philips Data

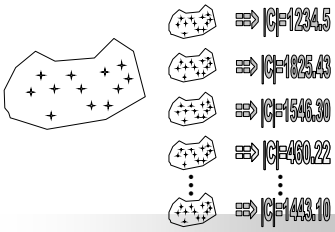
¿Cómo hallar C y $\underline{T}$?

- •Estimadores clásicos de media y varianza
- •Estimadores robustos
- •*High Breakdown estimators*



Robust is NOT *High Breakdown*

MCD (Rousseeuw *et al*., 1987)



FAST-MCD (Rousseeuw *et al*., 1999)

- ■ Mismo criterio, otro algoritmo
- ■ Más rápido, etc.
- ■ Maneja "exact fit"

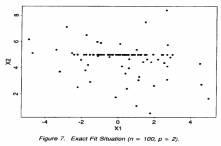


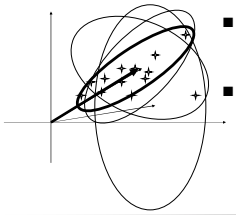
Figure 7. Exact Fit Situation (n = 100, p = 2).

Hadi (1992,1994)

- ■ Similar al MCD, pero no combinatorio
- ■ Más rápido que otros
- ■ Implementado en SAS y otros paquetes estadísticos

¿¿¿¿Y la esfinge???

MVE (Rousseeuw *et al.,* 1990)



- C y T tal que hay ε datos fuera de cada elipse
- C y T no son ahora función directa de un subconjunto de datos

Comentario…

- Los anteriores son casos particulares de estimadores más generales
  - ➤Estimador-S
  - ➤Estimador-M
- Veamos apenas una definición de cada uno de ellos

Estimador-S

- C=C(X) y T=T(X) tales que:

$$\begin{cases} \det(\mathbf{C}) \text{ sea mínimo} \\ \frac{1}{n}\sum_i \rho(d_i) = b_0 \end{cases}$$

$$d_i = \sqrt{(x_i - \mathbf{T})^T \mathbf{C}^{-1}(x_i - \mathbf{T})}$$

- ρ(d) función no decreciente
- MVE: ρ pertenece al conjunto {0,1}

Estimador-M

- C=C(X) y T=T(X) tales que:

$$\begin{cases} \det(\mathbf{C}) \text{ sea mínimo} \\ \sum_i (x_i - \mathbf{T}) * u_1(d_i) = 0 \\ \frac{1}{n}\sum_i (x_i - \mathbf{T})(x_i - \mathbf{T})^T u_2(d_i^2) = \mathbf{C} \end{cases}$$

- $u_1(d)$ y $u_2(d)$ ni negativas ni decrecientes si d>0

Aplicación: Philips data

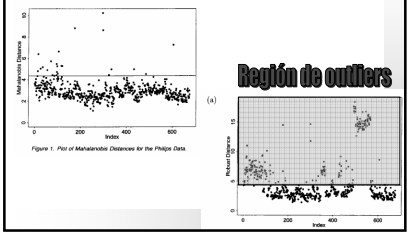


Figure 1: Plot of Mahalanobis Distances for the Philips Data

Recapitulando…

- Estimadores basados en Mahalanobis
- Difieren entre sí en la forma de estimar unos C y T *apropiados*
- Dados C y T son *simples* de aplicar
- Matemáticamente tratables
- Para datos tabulares
- Requieren datos sin ausencias
  - ➔quizá hay que imputar primero

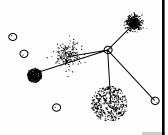
Minor ACP (Hawkins 1974)

- Los CP *mayores* (i.e. con gran valor propio) tienen interpretación física
- Normalmente se retienen, y los *menores* se descartan
- Hawkins propone utilizar los *scores* asociados como detectores de errores
- Son típicamente pequeños, e indican algo inusual cuando son grandes

Otra ventaja: Conglomerados

Ventaja: es un subproducto de la clasificación

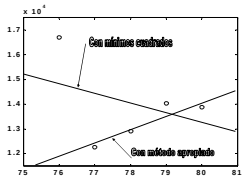
- Idea básica:
  - ➤Clasificar los datos en grupos
  - ➤Los puntos en los grupos pequeños son candidatos a outliers
  - ➤Calcular la distancia entre los puntos candidatos y los grupos no-candidatos
  - ➤Si los puntos candidato están *lejos* de *todos* los otros no candidatos, entonces se les declara outliers



Detección en el contexto de regresión

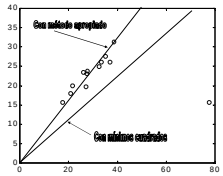
- Idea: ajustar con una función, analizar las discrepancias y separar las más groseras
- Problema: los errores pueden enmascararse unos a otros
- Problema: los errores pueden afectar significativamente la función de ajuste (Ejemplo: OLS)
- Solución ==> High breakdown methods (LTS, LMS, etc.)

Efecto de errores en regresión(1)



Reclamos por incendio en Bélgica

Efecto de errores en regresión(2)



Caudal en un dos puntos de un mismo río

OLS (Gauss, 18XX?)

- Minimiza la *suma* de cuadrados de residuos
- Sensible a outliers en varias formas
- Muy afectado por *enmascaramiento*
- ¡Implementado everywhere!
- En problemas tabulares tolera ausencias
  - ➤Requiere un OLS por cada combinación de ausencia / presencia➔puede ser pesado…

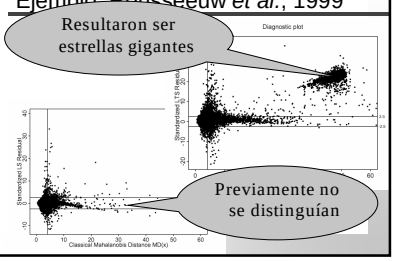
LMS (Rousseeuw 1984)

- Minimiza la *mediana* de los residuos
- Es como OLS si se descartara *cierto* 50% de la población
- Es por lo tanto del tipo *High Breakdown*
- Problema combinatorio ➔¡explota!
- Existen alternativas no combinatorias (Hawkins 1993)
- Igual que OLS: tolera ausencias

LTS (Rousseeuw 1984)

- Minimiza la *suma ponderada* de los residuos
- Pesos se eligen del conjunto {0.0,1.0}
- El total de casos 0.0 se especifica a priori
- Es por lo tanto del tipo *High Breakdown*
- Problema combinatorio ➔¡explota!
- Existen alternativas no combinatorias (Hawkins 1993)
- También tolera ausencias

Ejemplo: Rousseeuw *et al.*, 1999



Datos raster: caso del MDE

- Ampliamente estudiado en Agrimensura
- Antes: pocos puntos, muy precisos
  - ➤Típicamente formato TIN
  - ➤Raster se *calculaba* a partir del TIN
- Antes: fotogrametría aérea
  - ➤Típicamente curvas de nivel
  - ➤Raster se *calculaba* a partir de las curvas
- El productor dispone de controles internos
- También se detectan errores al comparar con la hidrografía, etc.

Situación presente

- Surgen otros métodos de creación (satélite, GPS, LIDAR, etc.)
- Muchos más puntos, algo menos precisos
  - ➤Imagen de satélite, etc.
  - ➤Se genera directamente el raster
- ¡El usuario puede ahora ir al campo y controlar!
- Detalle esencial: un pixel puede estar errado sin implicar al vecino
- Amerita otros métodos de control…

Método de Hannah (1981)

- MDE en formato raster
- Establece límites en la pendiente y cambio de pendiente
  - ➤¡Requiere especificarlos a priori!
- Usa interpolante local
- Fácilmente implementable en GIS
- Poco impacto en la literatura; mencionado aquí por *completness*

Método de Felicísimo (1994)

- Imputa interpolando con los vecinos
- Cualquier interpolante sirve; propone polinomio de 2do. grado en (i,j)
- Analiza la distribución de la diferencia del interpolado vs. el verdadero valor
- Asume normalidad, y saca límites
- Relativamente simple, implementable en SIG
- Veremos un ejemplo más adelante

Métodos mixtos

- Usan indirectamente métodos de regresión para detectar los errores
  - ➤Uso de la verosimilitud (*likelihood*)
  - ➤Interpretación de los roles de las neuronas en redes neuronales artificiales

Función de Verosimilitud (López, 1997)

- En un contexto de Kriging aparece el Variograma
- Depende de: Tipo, Alcance **a** y Meseta **S**
- Método de VCMV (Samper *et al., 1987)*
  - ➤Elegir **a** y **S** que maximicen la VCMV asumiendo que no dependen del tiempo…
- ¿Cómo es el método de VCMV?

VCMV (Samper, 1987)

- VCMV: Elijo **a** y **S,** y para cada fecha repito para los n puntos disponibles:
  - ➤Retiro el i-ésimo
  - ➤Interpolo mediante krigeado
  - ➤Conservo la discrepancia observada
- Luego se calcula la Verosimilitud
- Nuestro n era relativamente pequeño
- En general el proceso requiere minimizar una función no lineal costosa…

Nuestro problema particular de VCMV

- En realidad, nosotros no necesitábamos **a** y **S** …
- Sólo interpolaríamos en los puntos dato
- Para nuestros fines sólo necesitábamos una C y T obtenida de los datos experimentales
  - ➤No hubo necesidad de minimizar la función
- Se asumió homogeneidad e isotropía
- Se asumió también **a** y **S** constantes en el tiempo
- Idea: dados **a** y **S**, la Verosimilitud pasa a ser función del tiempo; un número por día
- Días con valores extremos➔¡outliers!

Reflexión…

- Casi nadie usa o referencia la Verosimilitud misma; sólo la maximiza
- Algo parecido a los Mínimos Cuadrados
  - ➤¿alguien se fija si los mínimos cuadrados son *pequeños*?
- Resultó ser uno de los mejores métodos en nuestros experimentos

No por trillado el camino es conocido
López (2005)

Métodos mixtos

- Usan indirectamente métodos de regresión para detectar los errores
  - ➤Uso de la verosimilitud (*likelihood*)
  - ➤Interpretación de los roles de las neuronas en redes neuronales artificiales

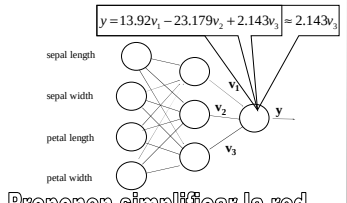
Uso de redes neuronales

Se reconocen dos líneas posibles

- Línea 1: Clasificación
  - ➤Clasificar en forma no supervisada
  - ➤Clusters con pocos elementos ➔ outliers
- Línea 2: Regresión
  - ➤Ajustar por MC y analizar discrepancias
- Línea 2.5: Regresión+…
  - ➤Ídem 2, pero luego interpretar roles
  - ➤Unpublished work, by López

Tomado de Benítez *et al.*, 1997

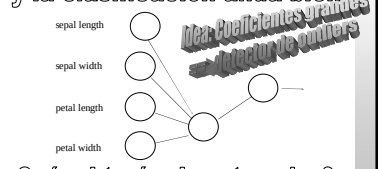
ANN para regresión



Proponen simplificar la red…

ANN para regresión

y la clasificación anda bien

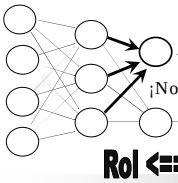


¿Qué rol tenían las otras dos?

Versión modificada

$$y = 13.92v_1 - 23.179v_2 + 2.143v_3$$

$$\begin{cases} if\ (v2 > 0.45\ or\ v1 > 0.73) \\ then\ z = 1 \\ else\ z = 0 \end{cases}$$



Rol <==> coeficiente

Ventajas…

- La Red se entrena como siempre para regresión / clasificación
- Se inspeccionan los pesos; no hay que reentrenar
- Los outliers no se decretan; ¡surgen!
- Desventaja: los pesos pueden ser muy sensibles a los outliers➔*masking*
- Fue testeado en el ejemplo (caso pequeño, *de paper*) y con lluvia, etc.
- ¡Fue el óptimo!
- Es aún una teoría. Queda mucho por hacer…

Ejemplos de detección de outliers

- Comentaremos algunos casos
- Tabular Cuantitativo: datos meteorológicos
  - Observados en una red de puntos fijos
  - Muchas medidas en el tiempo
  - ➢ Viento horario
    - Fuerte correlación espacio-tiempo
  - ➢ Lluvia diaria
    - En Uruguay, sólo correlación espacial
- Tabular Categórico: Datos de un Censo
- Raster: MDE

Datos tabulares: lluvia y viento

Cuantitativos

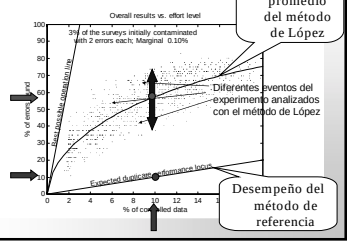
- Usamos lluvia diaria y viento horario
  - ➢ Lluvia tiene sólo correlación espacial
  - ➢ Viento tiene espacio-temporal
- Para el viento, 35% de los errores simulados aparecieron en el primer paso de depuración
- Para lluvia, 81% de los errores simulados aparecieron en el primer paso de depuración

Datos tabulares: censo nacional

Cualitativos

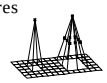
- Sólo para datos categóricos puros
- Pudimos remover 50% de los errores revisando un 10% del conjunto
  - ➢Cinco veces mejor que digitar de nuevo
- Método general, automatizable, basado en ACP

Gráficamente…



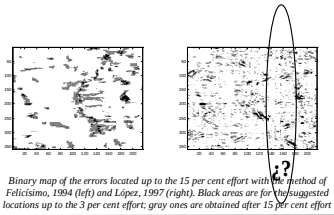
Datos raster: MDE (López 1997)

- Buscamos algunos tipos de errores
  - ➢ *Salt and pepper*
  - ➢ *Spike*
  - ➢ *Pyramid*



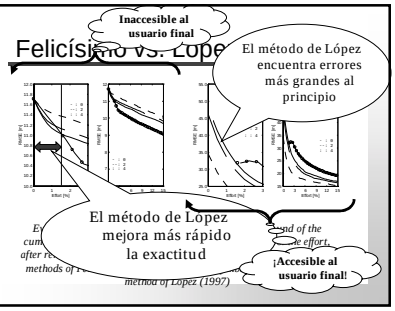
- El método es aplicable para cualquier raster *cuantitativo* (imágenes, fotos, etc.)
- En el artículo, 40% de los errores fueron encontrados con probabilidad > 88%
- Podría ser una herramienta útil para productores y usuarios

Felicísimo vs. López ☺



Binary map of the errors located up to the 15 per cent effort with the method of Felicísimo, 1994 (left) and López, 1997 (right). Black areas are for the suggested locations up to the 3 per cent effort; gray ones are obtained after 15 per cent effort

Felicísimo vs. Lopez



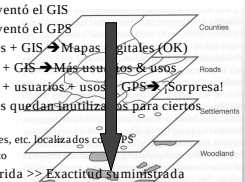
¿En qué estábamos? ¿Para dónde íbamos?

Otros casos

- Si la Exactitud es muy afectada por outliers
  - ➔¡detectar y remover outliers!
- ¿Y después?
  - ➤Errores sistemáticos
  - ➤Errores no groseros
- Requieren otro tipo de enfoque
- Ej.: PAI

Precision Accuracy Improvement

- Hecho #1: se inventó el GIS
- Hecho #2: se inventó el GPS
- Mapas existentes + GIS ➔ Mapas digitales (OK)
- Mapas digitales + GIS ➔ Más usuarios & usos
- Mapas digitales + usuarios + usos + GPS ➔ ¡Sorpresa!
- Mapas existentes quedan inutilizados para ciertos propósitos
  - ➤ Ductos, desagües, etc. localizados con GPS
  - ➤ Tráfico y tránsito
- Exactitud requerida >> Exactitud suministrada



¿Dentro?¿Fuera?¿Borde?

Caso típico (OS GB)

CA Access Map
Follows old OS well

MAGIC
Follows new OS badly !

Fuente: Robin Waters

¿Qué puede hacer el productor?

- Alternativa 1: Dejar todo como está
- Alternativa 2: Hacer todo de nuevo
- Alternativa 3: Intentar arreglar ➔ PAI
- Problema: [illegible] error planimétrico
- Mapas existentes tienen un gran valor "residual"
  - ➤ Actualizados (±…)
  - ➤ Populares (muchas veces únicos…)
  - ➤ Muchísimos atributos (¡cierto!)
  - ➤ Son base para otros mapas derivados (¡muy cierto!)

muchos $

PAI

- Idea: corregir *masivamente* la planimetría
- Ej.: OS GB; TIGER files USA
- Problema internacional
- Datos digitalizados… o no
- Impactos
  - ➤del lado del productor
  - ➤del lado del usuario

del lado del productor…

- Reingeniería de procesos
- Probable actualización tecnológica
- Alternativas:
  - ➤ Tercerización / Trabajo propio
  - ➤ Modificación / Nuevo relevamiento
  - ➤ Incorporación de otras fuentes
- ¿Financiación?
- Plan de actualización
  - ➤ Por dónde empezar
  - ➤ Cronograma de entre

¿investigación?

del lado del usuario…

- ¿Impacta / no impacta?
  - ➤ ¿Sólo datos PAI-compatibles?
  - ➤ ¿Datos generados internamente?
- Análisis de riesgo: *do nothing* vs. arreglo
- Si impacta ➔ alternativas:
  - ➤ Esperar a que PAI termine
  - ➤ Acompasar entregas con modificaciones internas
- Ambas tienen pros y contras

Resumen: No matar al mensajero…

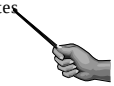
- No puede des-inventarse el GPS
- No puede des-inventarse el GIS
- No pueden des-digitalizarse los mapas

- Hay que *entender* los problemas
- Hay que *tomar* decisiones apropiadas
- Hay que *aprender* de otros ejemplos

¿Estamos perdidos?

Plan

- ✔ Introducción
- ✔ Revisión de herramientas estadísticas
- ✔ Detectando problemas
- Imputando valores ausentes
- Ejemplos



Recordemos: ¿Porqué imputar?

- Detectado un error…
- Modelos que no toleran ausencias
- Bajar costo al medir menos
- Típicamente métodos de *Interpolación*
  - ➤ Medidas escasas, de alta exactitud
- Actualmente más y más *Aproximación*
  - ➤ Más automatismo, menos control humano
  - ➤ Medidas abundantes, de menor exactitud

Enfoque es función del dato…

- Datos "*puramente*" espaciales
  - ➤ Caso más familiar para la audiencia
  - ➤ Métodos de Interpolación:
    - – TIN, Splines, Kriging, Cressman, etc.
  - ➤ Ej.: MDE, tipo de suelo, etc.
- Datos espacio-temporales
  - ➤ Correlación espacial + temporal
  - ➤ Ciencias de la Tierra, pero no Agrimensura
  - ➤ Ej.: Meteorológicos, uso el suelo, etc.
- Formulación sensiblemente específica

Datos *puramente* espaciales

- En la gran mayoría son Métodos lineales
- Coeficientes son función de punto
- Toleran ausencias
- A veces son lineales pero complicados
  - ➤ Cokriging
- Hay también métodos no lineales
  - ➤ Redes neuronales
  - ➤ Ecuaciones constitutivas (EDP)

Datos espacio-temporales

- Típicamente equi-muestreados en el tiempo
- Problema no resuelto: covarianza cruzada tiempo-espacio
- Muy usual en las Ciencias de la Tierra
  - ➤ Ej.: Meteorología, Hidrología, etc.
- Habitual en las aplicaciones GIS
  - ➤ Ej.: Tráfico / Tránsito, uso del suelo (!)
- Poco o mal manejado en GIS comerciales

Muchos métodos…

Procedimiento sugerido…

- Repita un número grande de veces
  - ➤ Generar ausencias al azar
  - ➤ Imputar con método1, método2, etc.
  - ➤ Calcular estadísticos de ajuste (distancias)
- Comparar estadísticos, y luego elija…
- Ventajas:
  - ➤ Tiene base estadística
  - ➤ Lo puede hacer el productor o el usuario
  - ➤ ¡No requiere ir al campo a medir!
- ¿Y las desventajas?

Desventajas o problemas…

- No todos los métodos están en los GIS
- ¿Cómo generar ausencias?
  - Al azar (MCAR)
  - En rachas (usual en datos meteorológicos)
- Hay que caracterizar primero SUS ausencias
- Otro tema: los estadísticos de éxito
  - Datos cuantitativos
  - Datos categóricos
  - Considerar o no el *impacto en el modelo*
- Un detalle más: el tiempo de cálculo

¿Cómo generar ausencias?

- Es más fácil que generar errores
- Hipótesis inicial: MCAR
  - Test descrito en Little (1988)
- En la práctica también había *rachas*
  - Rotura de instrumento
  - Pérdida de documento original en papel
- Quizá parezca excesivo detalle, pero…

Estadísticos de éxito

- Métricas usuales:
  - RMSE: Da mucho peso a errores groseros
  - MAD (Promedio): ídem RMSE
  - Percentiles: quizá más apropiado
- Asumiendo que existe un dato *verdadero* existe un Método Óptimo que lo asigna
- No existe en cambio un Peor Método
  - Podría usarse un *Naive* como referencia

Más sobre Estadísticos

- Podría considerarse el modelo
  - Errores sistemáticos pueden ser peores que errores groseros
  - Groseros son detectables; sistemáticos no
  - Ej.: errores en una factura:
    - Sesgados: ¡el cliente se queja dependiendo del signo!
- Otro problema: RMSE vs. Exactitud original
  - Ej.: RMSE lluvia ~7 mm / día; Exactitud 5 mm / día, pero ¡¡precisión 0.1 mm / día!!

Plan

- ✔ Introducción
- ✔ Revisión de herramientas estadísticas
- ✔ Detectando problemas
- ✔ Imputando valores ausentes
- Ejemplos

Caso del Viento horario

Problema:

- Completar un banco de datos de viento de superficie horario
- Comparar diferentes métodos, en dos diferentes casos:
  - ⇒ Ausencias al azar
  - ⇒ Ausencias planificadas

Fuente: Proyecto CONICYT / BID 51 / 94 (1999)

Diseño de la metodología

- Seleccionar un banco apropiado, lo más completo posible
- Ocultar temporalmente los valores a ser imputados (elegidos al azar o no)
- Para cada método
  - ⇒ imputar todos los valores ausentes
  - ⇒ calcular RMSE y MAD de las discrepancias entre el valor real y el imputado

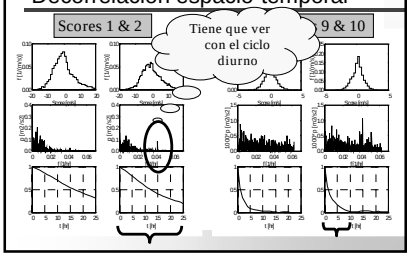
El banco de datos

- Cinco estaciones meteorológicas, separadas no más de 400 km, en terreno suave
- Aproximadamente 25 meses de registros horarios

Descripción de los métodos

- Interpolación Óptima (equivalente a Krigeado Ordinario)
- Métodos basados en el Análisis de Componentes Principales:
  - **T**emporal **I**nterpolation of **P**rincipal **S**cores (TIPS)
  - **P**enalty **O**f the **P**rincipal **S**cores (POPS)

Decorrelación espacio-temporal



Resultados preliminares obtenidos

a) Ausencias sistemáticas

- Se asumieron tres lecturas diarias (8, 14 y 21 hs.), en cuatro de cinco estaciones
- Con TIPS se logra un RMSE de 2.05 m/s
- Con POPS se logra RMSE de 2.84 m/s
- La Interpolación Óptima produce 2.84 m/s
- Asignando simplemente la media histórica el RMSE es de 3.24 m/s

Resultados preliminares obtenidos

b) Ausencias al azar

- Se ocultó aleatoriamente un 20% de los datos, criterio MCAR
- Con TIPS se logra un RMSE de 1.67 m/s
- Con POPS se logra RMSE de 2.33 m/s
- La Interpolación Óptima produce 2.37 m/s
- Asignando la media histórica el RMSE es de 2.76 m/s

Conclusiones

- El uso de la información temporal da resultados más precisos, sugiriendo un muestreo excesivo para esta zona
- Los resultados deben ser corroborados en ensayos más extensos, para darle validez estadística
- Otros métodos deben ser incluídos en la comparación

Ver informe final de 1999

Caso de la lluvia diaria

- Nuevamente, un problema *tabular*
- 10 estaciones, registros diarios (mm/día)
- Correlación espacial pero no temporal
  - ➤TIPS falla miserablemente
- Problema difícil
  - ➤RMSE del Mejor vs. Peor método evaluado difieren en 30%
- Mejor RMSE: 7 mm/día; según los expertos, la Exactitud~5 mm/día (!)

Sugerencias para lectura…

- Informe CONICYT/BID 51/94 (1999)
  - ➤Análisis comparativo de ~30 métodos
    - – Imputación
    - – Detección de outliers
  - ➤Descripción de métodos, referencias, etc.
  - ➤No orientado a meteorología
  - ➤Único estudio sistemático conocido

Módulo 4:
Mejorando la Exactitud

Carlos López Vázquez

carlos.lopez@ieee.org