Modelación Estadística

Charla de Especialidad

Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Estructura de la presentación

0. Presentación del grupo de Estadística.
1. El problema de modelamiento estadístico.
2. Motivación mediante dos problemas.
3. Posibles tópicos de investigación.
4. Perfil de alumnos en la línea.
5. Algunos comentarios finales.

Modelación Estadística: Académicos

(a) A. Alegría

(b) F. Osorio

(c) R. Vallejos

Modelación Estadística: Alfredo Alegría

- Ph.D. Mathematics, PUCV-UTFSM-UV, Chile (supervisor: Emilio Porcu).
- Áreas de investigación: Multivariate spatial statistics, Geostatistics for large datasets, Non-gaussian random fields.
- Red de Colaboración: Emilio Porcu, Peter Diggle (Reino Unido), Jorge Mateu (España), Reinhard Furrer (Suiza), Stefano Castruccio (USA), Moreno Bevilacqua (Chile).
- Trabajos relevantes: Electronic Journal of Statistics, International Statistical Review, Journal of Multivariate Analysis, Journal of Statistical Computation and Simulation, Spatial Statistics, Stat, Statistics and Computing.
- Proyecto FONDECYT Iniciación.
- Postdoctorado en la Universidad de Newcastle, UK.

Modelación Estadística: Ronny Vallejos

- Ph.D. Statistics, University of Maryland Baltimore County, USA. (supervisor: Andrew L. Rukhin).
- Áreas de investigación: Spatial statistics, Robust modelling, Statistical image modelling, Time series.

- Red de Colaboración: Daniel Griffith, Aaron Ellison (USA), Silvia Ojeda, Oscar Bustos (Argentina), Hannah Buckley, Bradley Case (New Zealand).
- Trabajos relevantes: Chance, Electronic Journal of Statistics, Journal of Mathematical Imaging and Vision, Journal of Statistical Planning and Inference, Spatial Statistics, Stochastic Environmental Research and Risk Assessment.
- Proyectos FONDECYT, de cooperación internacional (CECYT, Math-AmSud), PIA.
- Ex Editor en Jefe de la revista *Chilean Journal of Statistics*, miembro de los centros AM2V y AC3E.

Modelación Estadística: Felipe Osorio

- D. Sc. Statistics, Universidade de São Paulo, Brasil.
  (supervisor: Gilberto A. Paula).
- Áreas de investigación:
  Modelos para datos longitudinales,
  Diagnóstico de influencia,
  Funciones de inferencia.

- Red de Colaboración: Gilberto A. Paula, Cibele Russo (Brasil), Manuel Galea (Chile), Federico Crudu (Italia).
- Trabajos relevantes: Annals of the Institute of Statistical Mathematics, Biometrical Journal, Computational Statistics & Data Analysis, Economics Letters, Spatial Statistics, Statistical Papers, Statistics and Computing.
- Proyectos FONDECYT, de cooperación internacional (PROSUL, CNPq).
- Creador de paquetes contribuídos a R (heavy, L1pack, MVT, SpatialPack).
- Editor de la revista *Chilean Journal of Statistics*.

Ideas subyacentes

"Todos los modelos son errados, pero algunos son útiles."

– George Box.

"Aunque puede parecer una paradoja, toda la ciencia exacta está dominada por la idea de aproximación."

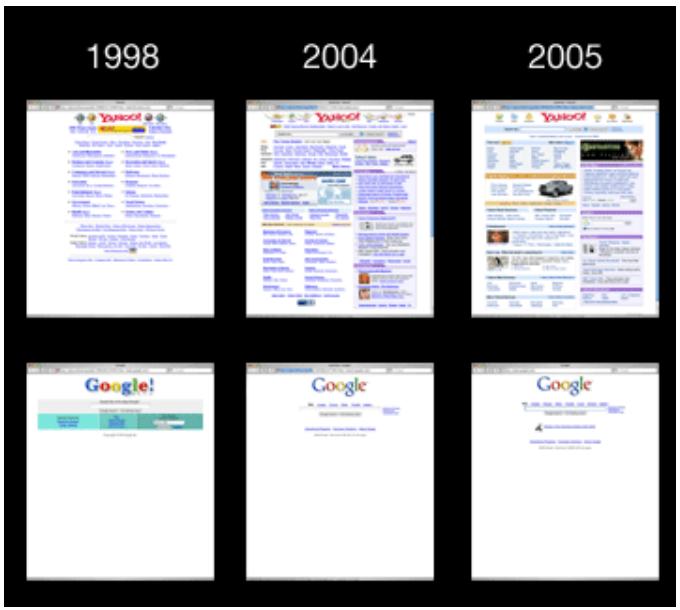
– Bertrand Russell.

Principio KISS: "Keep It Short and Simple."

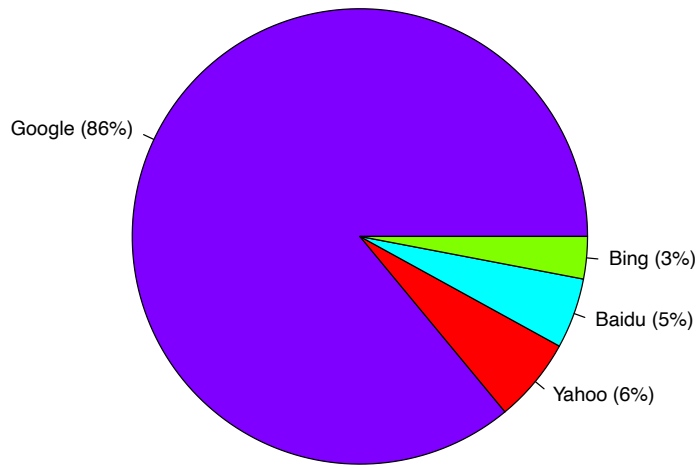
– Clarence "Kelly" Johnson.

El éxito de Google: Aplicar el principio KISS[1]

Evolución de Yahoo vs. Google:



Cuota de mercado de los motores de búsqueda:



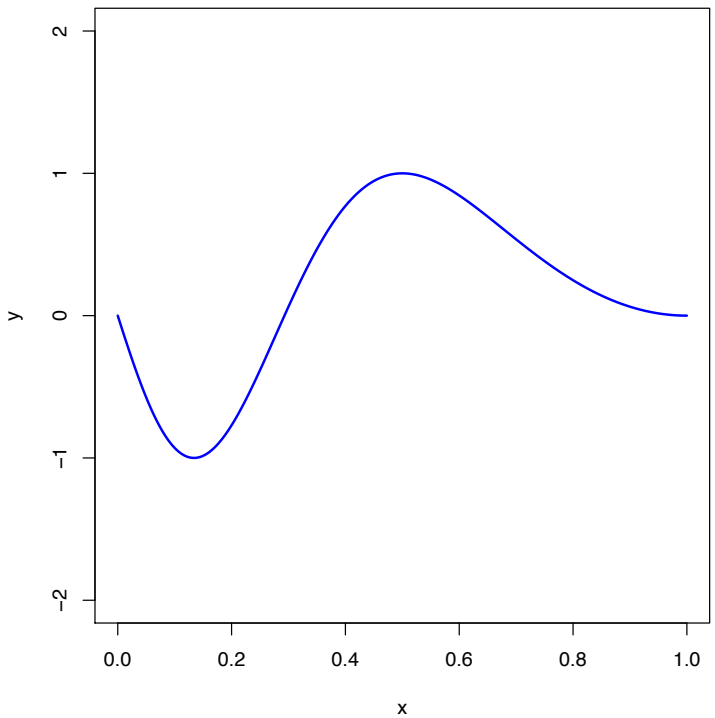
[1] En estadística este se conoce como Principio de Parsimonia.

El problema del modelado

Considere la función

$$Y = \text{sen}\{2\pi(1-x)^2\},$$

cuyo gráfico es dado por:

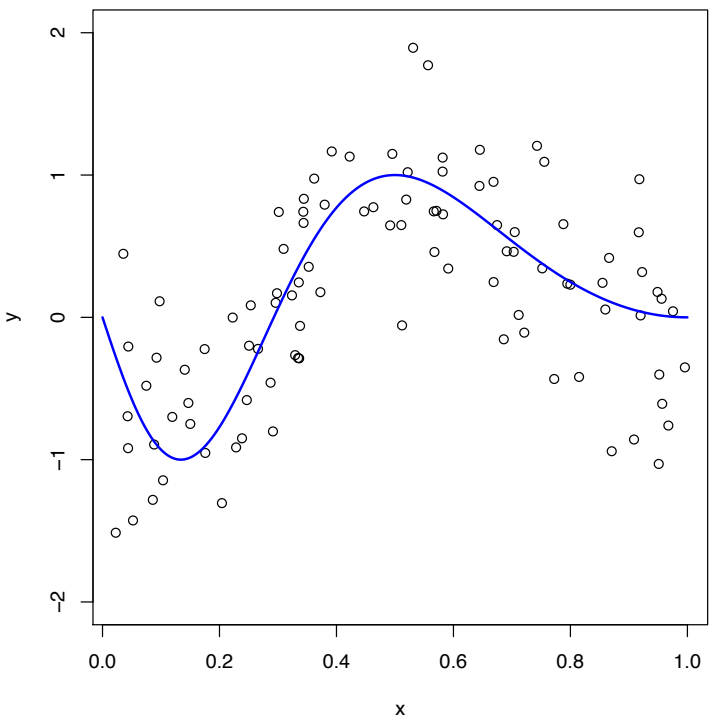


El problema del modelado

Suponga que "generamos" datos, usando

$$Y_i = \text{sen}\{2\pi(1 - x_i)^2\} + \sigma\epsilon_i, \qquad i = 1, \ldots, 100,$$

donde $x_i \sim \text{Unif}(0, 1)$, $\epsilon_i \sim \mathcal{N}(0, 1)$ y $\sigma = 1/2$,

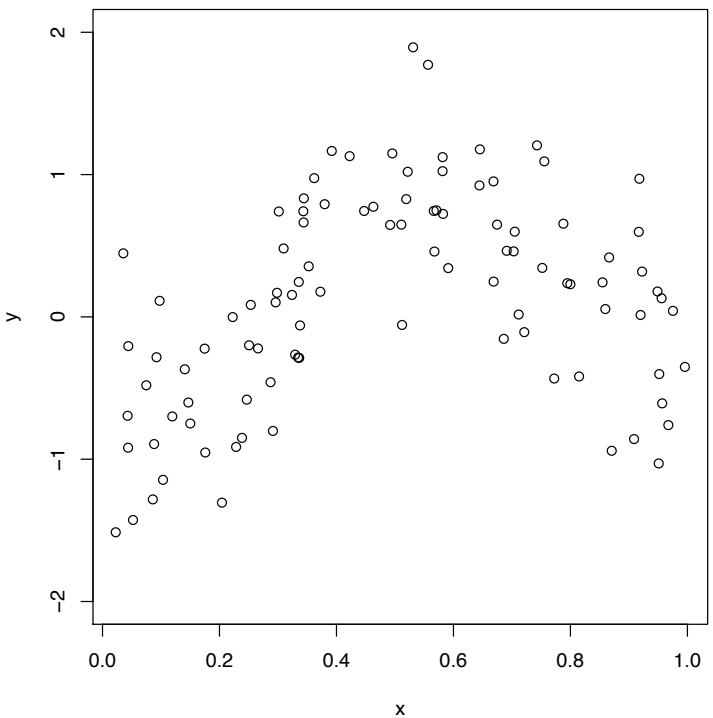


El problema del modelado

Lamentablemente, en la práctica **sólo** disponemos de los datos observados:

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_{100}, Y_{100}),$$

el primer paso es hacer un análisis exploratorio:

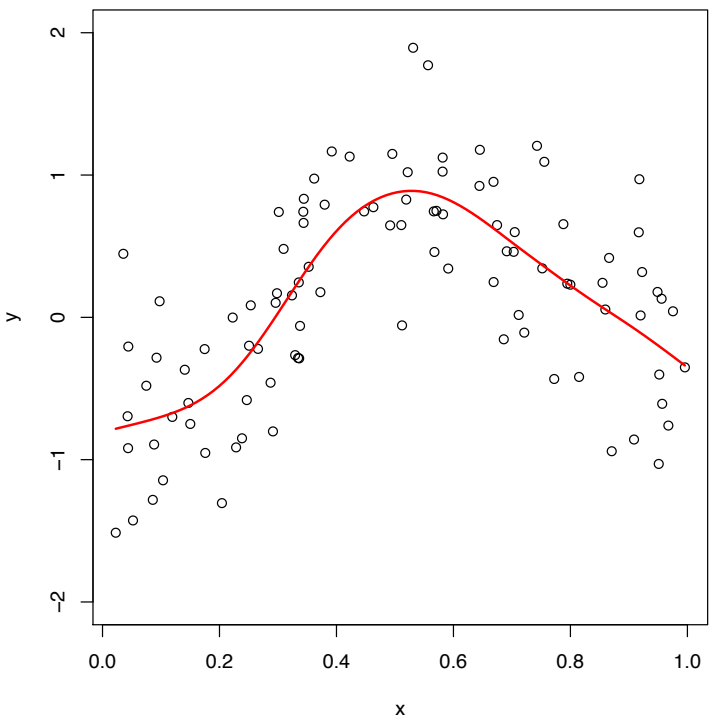


El problema del modelado

El analista propone el modelo:

$$Y_i = g(x_i) + \epsilon_i, \qquad i = 1, \ldots, 100,$$

y su objetivo es "estimar" la función $g(\cdot)$ desde los datos, obteniendo

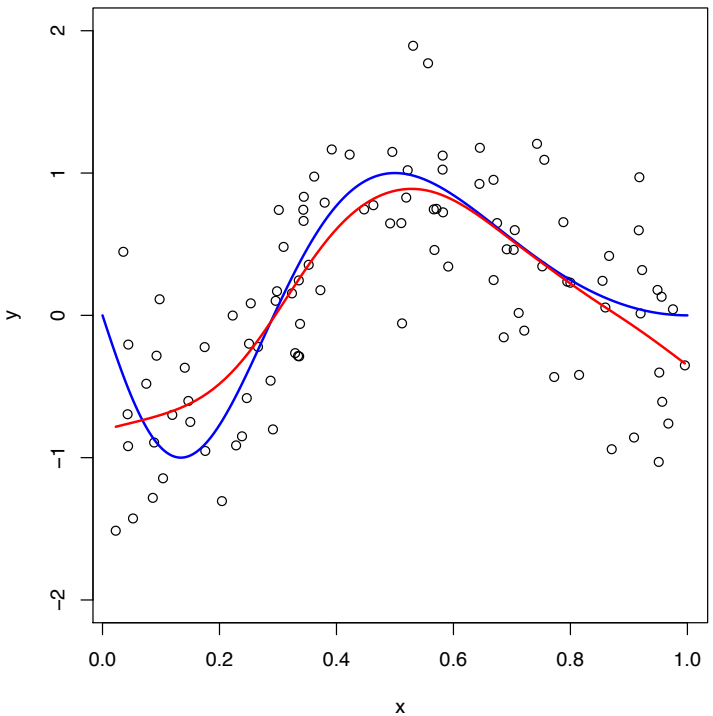


El problema del modelado

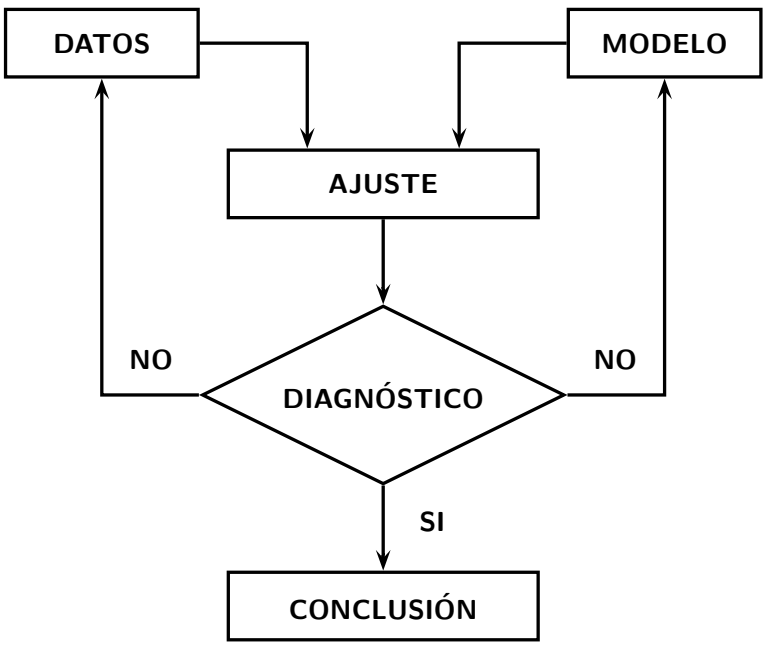
En Estadística se estudia teóricamente, la "bondad del modelo" comparando

$$\widehat{Y} = \widehat{g}(x), \qquad \text{v.s.} \qquad Y = \text{sen}\{2\pi(1-x)^2\},$$

esto es, el modelo ajustado v.s. el modelo subyacente (verdadero).



Esquema de Modelación Estadística



Recolección de datos: Muestreo.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comunique sus resultados!

Estimación del SSIM

Trabajo:

Osorio, F., **Vallejos, R.**, **Barraza, W.**, Ojeda, S., Landi, M.A. (2020). Estimation of the structural similarity index for remote-sensing data. *Statistics and Computing* (enviado).

Considere dos imágenes $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_{+}^{N}$, el coeficiente de similaridad estructural (SSIM) es dado por:

$$\text{SSIM}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = l(\boldsymbol{x}, \boldsymbol{y})^{\alpha} \cdot c(\boldsymbol{x}, \boldsymbol{y})^{\beta} \cdot s(\boldsymbol{x}, \boldsymbol{y})^{\gamma},$$

donde

$$l(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\overline{x}\,\overline{y} + c_1}{\overline{x}^2 + \overline{y}^2 + c_1}, \qquad c(\boldsymbol{x}, \boldsymbol{y}) = \frac{2s_x\, s_y + c_2}{s_x^2 + s_y^2 + c_2},$$

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{s_{xy} + c_3}{s_x\, s_y + c_3}.$$

Objetivo:

Basado en imágenes observadas $\boldsymbol{x}$ e $\boldsymbol{y}$, estimar $\boldsymbol{\theta} = (\alpha, \beta, \gamma)^{\top}$ y probar la hipótesis:

$$H_0 : \alpha = \beta = \gamma = 1.$$

Estimación del SSIM

Se consideró un modelo no-lineal heteroscedástico[2] bajo el supuesto de normalidad:

$$Z_i \sim \mathsf{N}(\phi f_i(\boldsymbol{\theta}), f_i^2(\boldsymbol{\theta}) g^2(\phi)), \qquad i = 1, \ldots, n,$$

donde

$$f_i(\boldsymbol{\theta}) = \mathrm{SSIM}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\theta}), \qquad g^2(\phi) = \phi^2(\phi^2 - 1),$$

corresponden a funciones de media y de varianza y $Z_i = 1/\,\mathrm{RMSE}(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

Resultados:

- Algoritmo de estimación: Híbrido entre método secante multivariado (BFGS) con pseudo-verosimilitud (método de Brent).[3]
- Test de hipótesis usando el estadístico gradiente (Terrell, 2000).
- Matriz de información de Fisher y método eficiente para evaluar función score.
- Experimento numérico con datos sintéticos (desde base de datos USC-SIPI).[4]

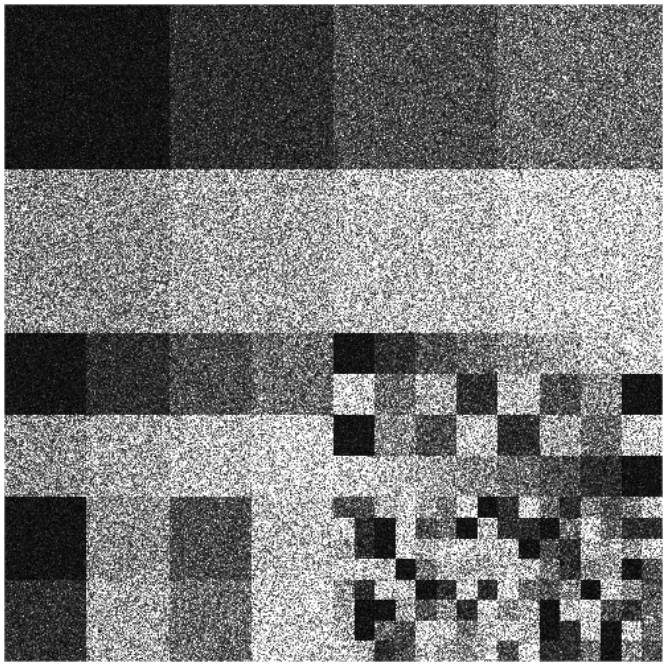
[2] Inspirado en el contexto de funciones de producción.

[3] Código C y R disponible en `github.com/faosorios/SSIM`.

[4] Se ajustaron 144 000 modelos (tiempo: 34 hrs, 40 min, 16 seg).

Imagen de referencia (texmos2) vs imagen distorsionada

(a)



(b)

Para este par de imágenes, obtenemos

$$\widehat{\alpha} = 1.5255, \quad \widehat{\beta} = 1.6509, \quad \widehat{\gamma} = 1.5188, \quad \mathrm{SSIM}(\widehat{\boldsymbol{\theta}}) = 0.8777,$$

y rechazamos $H_0 : \alpha = \beta = \gamma = 1$.

Influencia local para el análisis de acuerdo/concordancia

Suponga que deseamos evaluar el grado de acuerdo entre dos instrumentos de medición y sea $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ observaciones bivariadas con vector de medias $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$.

Considere

$$\rho_{\mathsf{c}} = \frac{2\sigma_{12}}{\sigma_{11} + \sigma_{22} + (\mu_1 - \mu_2)^2}, \qquad \psi_{\mathsf{c}} = \mathsf{P}(|X_{i1} - X_{i2}| \leq c), \qquad c > 0.$$

el coeficiente de concordancia (Lin, 1989) y la probabilidad de acuerdo (Stevens et al., 2017), respectivamente. Bajo normalidad,

$$\psi_{\mathsf{c}} = \Phi\Big(\frac{c - \mu_D}{\sigma_D}\Big) - \Phi\Big(-\frac{c - \mu_D}{\sigma_D}\Big),$$

con $\mu_D = \mu_1 - \mu_2$, $\sigma_D^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$.

Influencia local para el análisis de acuerdo/concordancia

Trabajo:

Leal, C., Galea, M., **Osorio, F.** (2019).
Assessment of local influence for the analysis of agreement.
Biometrical Journal **61**, 955-972.

Objetivo:

Considerar funciones de influencia $f(\boldsymbol{\omega})$ y estudiar el comportamiento del modelo perturbado

$$\mathcal{P}_\omega = \{p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\omega}) : \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}.$$

con $\boldsymbol{\omega}_0$ vector de perturbación nula, tal que $\mathcal{P}_{\omega_0} = \mathcal{P}$, con

$$\mathcal{P} = \{\mathsf{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})^{\otimes n} : \boldsymbol{\mu} \in \mathbb{R}^2, \boldsymbol{\Sigma} \geq 0\}.$$

Llevar a cabo influencia local de primer y segundo orden sobre $\rho_{\mathsf{c}}(\boldsymbol{\omega})$ y $\psi_{\mathsf{c}}(\boldsymbol{\omega})$.

Influencia local para el análisis de acuerdo/concordancia

Se obtuvo las curvaturas normal (Cook, 1986) y conformal (Poon y Poon, 1999)

$$C_{f,h} = \frac{\boldsymbol{h}^\top \boldsymbol{H}_f \boldsymbol{h}}{(1 + \nabla_f^\top \nabla_f)\boldsymbol{h}^\top (\boldsymbol{I} + \nabla_f \nabla_f^\top)\boldsymbol{h}},$$

$$B_{f,h} = \frac{\boldsymbol{h}^\top \boldsymbol{H}_f \boldsymbol{h}}{\|\boldsymbol{H}_f\|_M \boldsymbol{h}^\top (\boldsymbol{I} + \nabla_f \nabla_f^\top)\boldsymbol{h}},$$

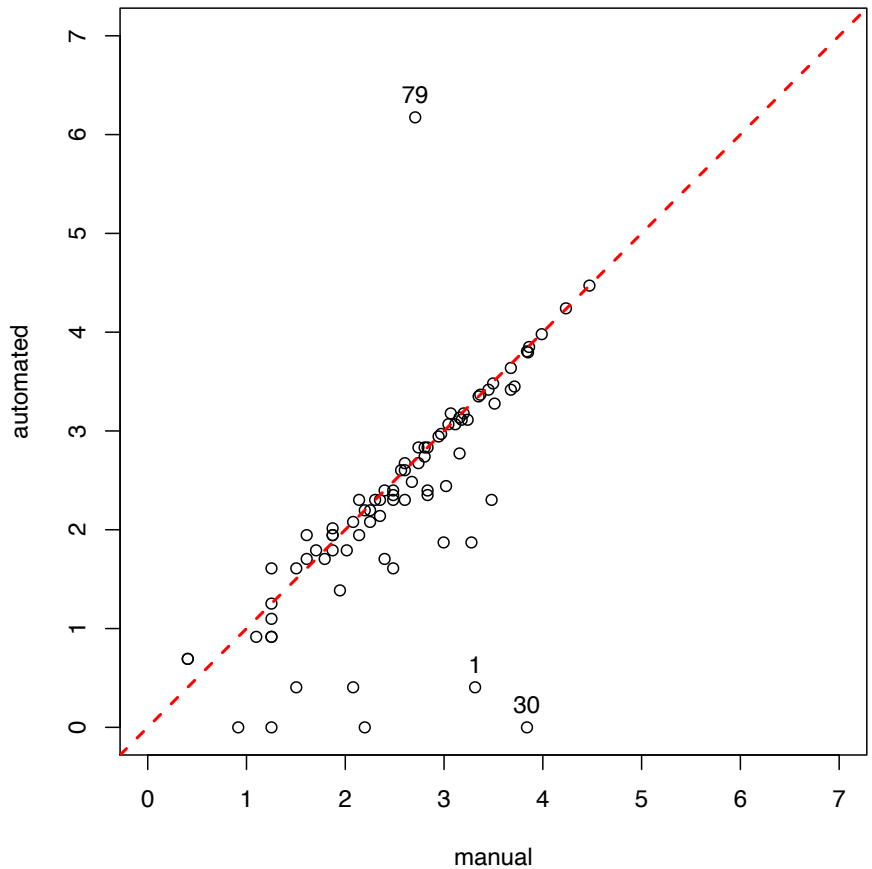
y medidas de influencia de primer y segundo orden (Zhu et al., 2007)

$$\mathrm{FI}_{f,h} = \frac{\boldsymbol{h}^\top \nabla_f \nabla_f^\top \boldsymbol{h}}{\boldsymbol{h}^\top \boldsymbol{G}(\boldsymbol{\omega}_0)\boldsymbol{h}}, \qquad \mathrm{SI}_{f,h} = \frac{\boldsymbol{h}^\top \tilde{\boldsymbol{H}}_f \boldsymbol{h}}{\boldsymbol{h}^\top \boldsymbol{G}(\boldsymbol{\omega}_0)\boldsymbol{h}},$$

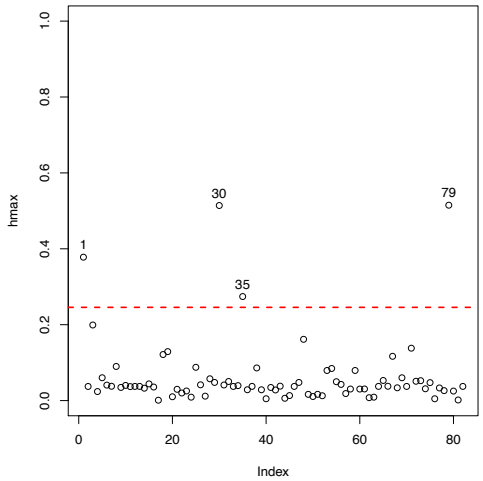
donde $\nabla_f = \partial f(\boldsymbol{\omega})/\partial \boldsymbol{\omega}\big|_{\omega=\omega_0}$ y $\boldsymbol{H}_f = \partial^2 f(\boldsymbol{\omega})/\partial \boldsymbol{\omega}\partial \boldsymbol{\omega}^\top\big|_{\omega=\omega_0}$, con $\boldsymbol{G}(\boldsymbol{\omega})$ matriz de información de Fisher con relación a $\boldsymbol{\omega}$,

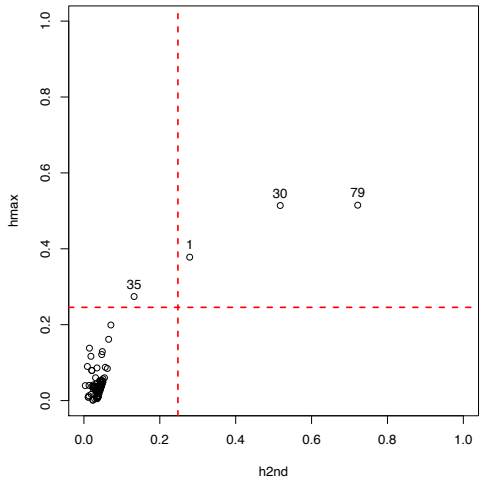
$$g_{ij}(\boldsymbol{\omega}) = \mathsf{E}_\omega \left\{ \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \omega_i} \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \omega_j} \right\}, \qquad i, j = 1, \ldots, n.$$

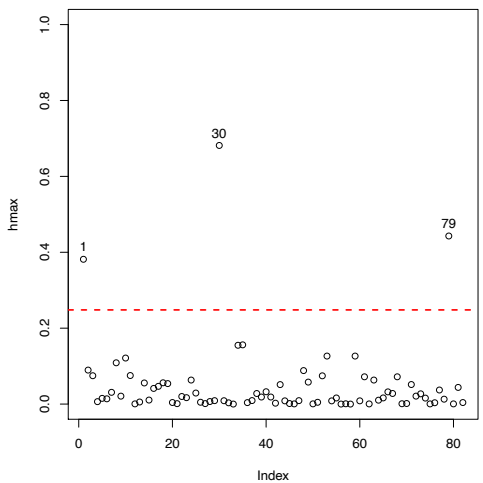
Datos de desorden transitorio de sueño (Svetnik et al., 2007)

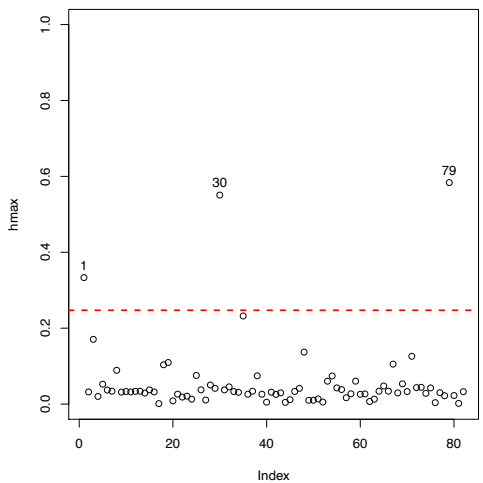


Datos de desorden transitorio de sueño, objetivo: $\rho_c(\omega)$

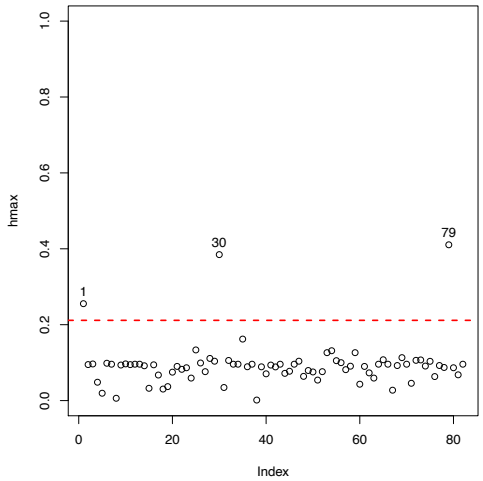


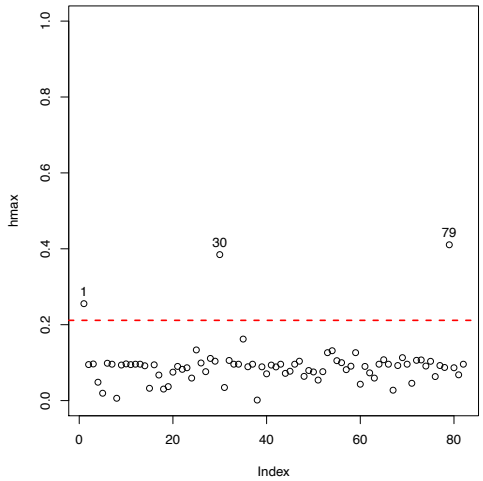


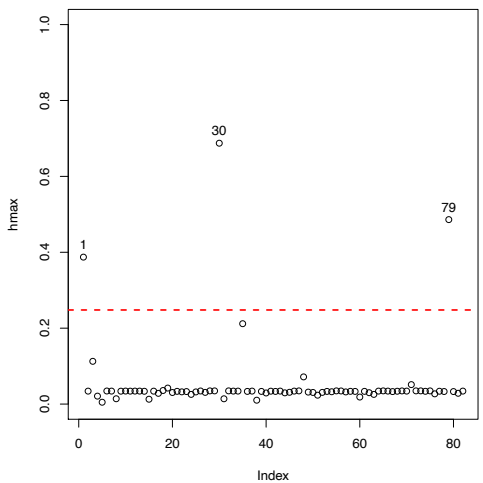


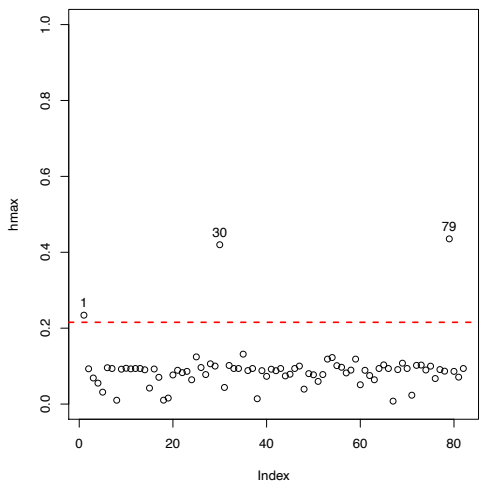


Datos de desorden transitorio de sueño, objetivo: $\psi_c(\omega)$









Resultados:

- Influencia local de primer y segundo orden aplicado a funciones objetivo ρ_c y ψ_c.
- Experimentos numéricos: estudio de simulación Monte Carlo y aplicación a datos de la realidad sugieren que influencia de primer orden es más potente para la detección de observaciones atípicas.
- Se determinó que el esquema de perturbación de ponderación de casos corresponde a una perturbación apropiada.
- Código en R para la estimación y el diagnóstico de influencia se encuentra disponible en `github`.[5]
- Resultado lateral: distribución asintótica para la probabilidad de concordancia, ψ_c.

[5] `github.com/faosorios/CCC`

Temas tentativos para tesis: Alfredo Alegría

- Campos aleatorios no estacionarios sobre grandes porciones del planeta.
- Funciones de covarianza cruzada flexibles y su aplicación en el análisis de datos espaciales multivariados.
- Métodos de estimation y predicción para campos aleatorios no-gaussianos.

Temas tentativos para tesis: Ronny Vallejos

- Desarrollo de medidas de concordancia para datos espacio-temporales.
- Extensión de la probabilidad de concordancia para datos espaciales.
- Estimación del coeficiente de similaridad entre imágenes.
- Tamaño muestral efectivo para datos espacio-temporales.

Temas tentativos para tesis: Felipe Osorio

- Influencia local para regresión LAD.
- Probabilidad de concordancia para varios instrumentos de medición.
- Diagnóstico en análisis multivariado bajo estimación máximo Lq-verosímil.

Modelación Estadística: Estudiantes

Pregrado:

- Wilson Barraza. (U-Planner)
- Claudio Henríquez. (PUC)
- Jorge Littin.
- Jean Paul Maidana. (UV)
- Marcela Miranda.[6]
- Consuelo Moreno. (DELPHOS)
- Eric Muñoz.[6] (AFP Provida)
- Ignacio Vásquez.
- Gabriel Vidal.[6]

Postgrado:

- Jonathan Acosta. (PUCV)
- Francisco Alfaro. (U-Planner)
- Paola Carvajal.
- Francisco Cuevas. (UQÀM)
- Ángelo Garate. (PUC)
- Diego Mancilla. (DELPHOS)
- Alonso Ogueda.[6] (U-Planner)
- Javier Perez.[6]
- Danilo Pezo. (TU Kaiserslautern)
- Carlos Schwarzenberg.[6]

[6] Actualmente en nuestros programas de pre y postgrado.

Algunos trabajos recientes

Alegría, A. (2020+).
Cross-dimple in the cross-covariance functions of bivariate isotropic random fields on spheres.
Stat (to appear).

Alegría, A., Cuevas, F. (2020+).
Karhunen-Loève expansions for axially symmetric gaussian processes: Modeling strategies and L2 approximations.
Stochastic Environmental Research and Risk Assessment (to appear).

Alegría, A., Emery, X., Lantuéjoul, C. (2020+).
The turning arcs: A computationally efficient algorithm to simulate isotropic vector-valued gaussian random fields on the d-sphere.
Statistics and Computing (to appear).

Crudu, F., **Osorio, F.** (2020).
Bilinear form test statistics for extremum estimation.
Economics Letters **187**, 108885.

Vallejos, R., Pérez, J., Ellison, A., Richardson, A. (2020).
A spatial concordance correlation coefficient with an application to image analysis.
Spatial Statistics (in press).

Algunos trabajos recientes

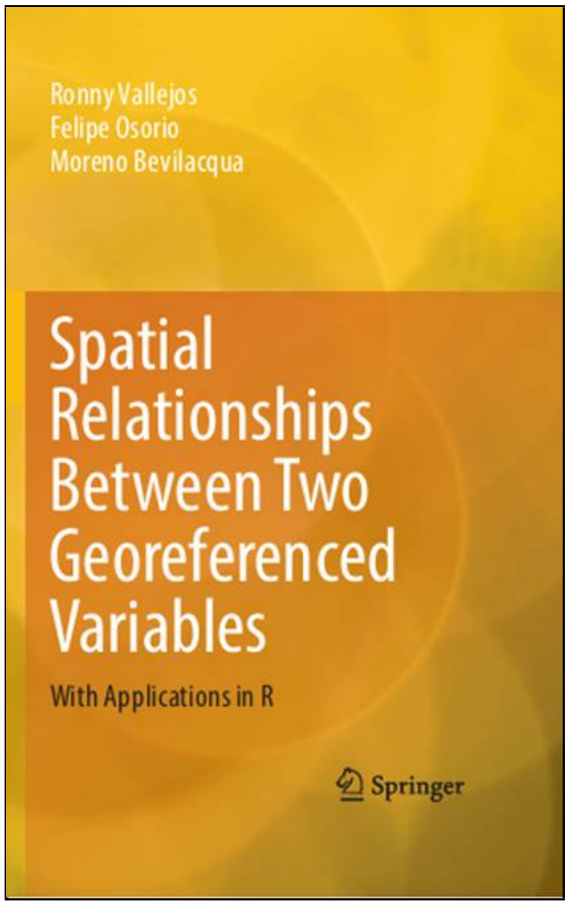
Alegría, A., Porcu, E., Furrer, R., Mateu, J. (2019).
Covariance functions for multivariate Gaussian fields evolving temporally over planet earth.
Stochastic Environmental Research and Risk Assessment **33**, 1593–1608.

Leal, C., Galea, M., **Osorio, F.** (2019).
Assessment of local influence for the analysis of agreement.
Biometrical Journal **61**, 955-972.

Porcu, E., Castruccio, S., **Alegría, A.**, Crippa, P. (2019).
Axially symmetric models for global data: A journey between Geostatistics and Stochastic generators.
Environmentrics **30**, e2555.

Vallejos, R., Gárate, A., Gomez, M. (2019).
Comovement among returns of private chilean pension system.
International Review of Applied Economics **33**, 228-240.

Asociación espacial entre dos procesos estocásticos (libro Springer)



- Ronny Vallejos, Felipe Osorio (USM) y Moreno Bevilacqua (UAI).
- Asociación entre dos procesos espaciales:
  - procedimientos de test de hipótesis.
  - coeficientes de asociación/codispersión.
  - asociación entre imágenes.
- Paquetes en R: `SpatialPack` y `GeoModels`.
- Proximamente publicado por Springer (Diciembre 2020).

Cursos de especialidad: Modelación Estadística

Especialidad:

- Análisis de regresión.
- Análisis multivariado.
- Series de tiempo I.
- Series de tiempo II.

Magíster:

- Estadística espacial.
- Modelos lineales generalizados.
- Series de tiempo avanzadas.
- Simulación estocástica.

Otros cursos:

- Base de datos (INF)
- Física computacional (FIS)
- Inteligencia artificial (INF)
- Minería de datos (ELO)
- Proc. imágenes digitales (ELO)
- Teoría de información (ELO)

Comentarios/dudas?