LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Incorporación de Información Satelital Grillada para la Producción de una Base de Datos de Temperaturas Mínimas de Alta Resolución

LatinR 2018.
Conferencia Latinoamericana sobre Uso de R en Investigación + Desarrollo.
4 y 5 de Setiembre del 2018.
Buenos Aires, Argentina.

Ing. Pablo Alfaro. Facultad de Ingeniería
M.Sc. Leonardo Moreno. Instituto de Estadística
D. Sc. Marco Scavino. Instituto de Estadística

Agenda

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Objetivos del Trabajo
- Base de Datos Puntual de las Estaciones de Observación
- Base de Datos Grillada Satelital
- Información Faltante y Controles de Calidad
- Modelado Geoestadístico
  - Kriging Ordinario vs Regression Kriging
  - Cross Validation
  - Estadísticos de Validación
- Aspectos Computacionales, Big Data, R
- Evaluación de Modelos Geoestadísticos
- Resultados Obtenidos
- Trabajo Futuro

Introducción

- Uruguay cuenta con redes de observación meteorológica de diversas instituciones
  - INUMET
  - INIA
- La recolección de información meteorológica de largo aliento y espacialmente representativa en todo el país es dificultosa
  - Costos elevados
  - Series incompletas, cierre de estaciones
  - Errores en las observaciones
- Existe información satelital complementaria, de relativamente largo aliento y excelente cobertura espacial disponible en forma gratuita

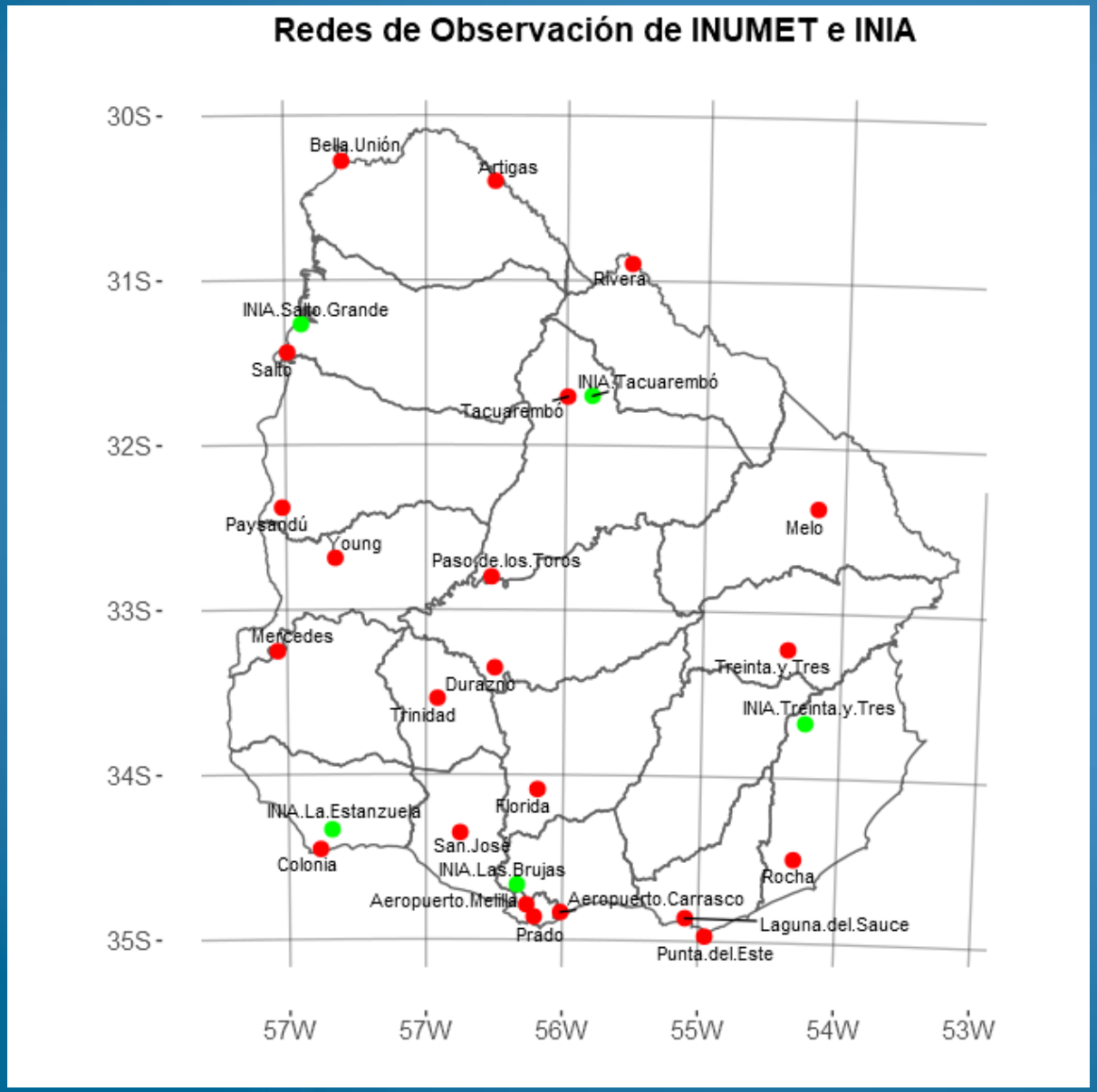
Objetivos

- Incorporar a los datos puntuales tradicionales, fuentes de información grillada satelital
  - Mejor cobertura espacial que las estaciones
  - Calibración del satélite de acuerdo a las estaciones
- Incorporar efectos locales al modelado
- Mejorar la predicción (espacial) de variables meteorológicas, en particular TMin
- Estimar Información Faltante
- Detectar Valores Sospechosos
- A futuro: incorporar las mejoras a productos elaborados: Balance Hídrico, Propagación de Plagas o Enfermedades, Detección de Heladas

Base de Datos Puntual

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

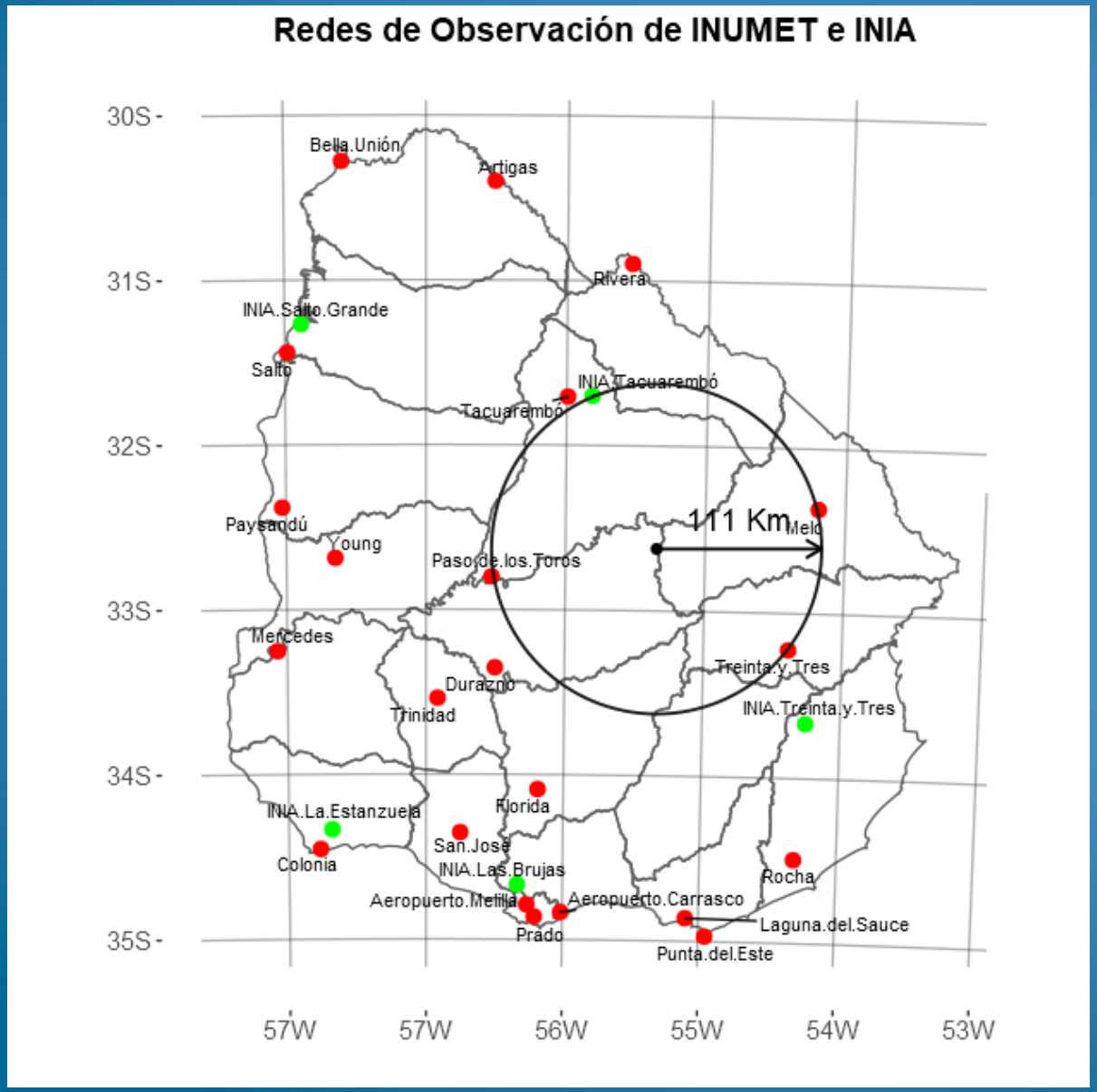


- 22 Estaciones de INUMET
- 5 Estaciones de INIA
- Período 2002-2014
- En general buena distribución espacial
- Algunas zonas con poca cobertura

Base de Datos Puntual

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY



- 22 Estaciones de INUMET
- 5 Estaciones de INIA
- Período 2002-2014
- En general buena distribución espacial
- Algunas zonas con poca cobertura
  - Dist. Máx: 111.6 Km
  - Dist. Media: 46.3 Km

Base de Datos Grillada

- Temperatura de Superficie Terrestre (LST) de plataforma MODIS. [NASA, 9]
- Datos Grillados
  - Disponibles sobre todo el país
  - 2002-07-08 al 2014-12-31
- Dos satélites
  - Terra: Pasaje por el ecuador a las 10:30 am / pm
  - Aqua: Pasaje por el ecuador a la 1:30 am / pm
- Cuatro Productos
  - MOD11A1, Terra, Diario
  - MOD11A2, Terra, Promedio 8 días de ventana fija
  - MYD11A1 , Aqua, Diario
  - MYD11A2 , Aqua, Promedio 8 días de ventana fija

Información Faltante y Controles de Calidad

- Controles de Calidad para Base Puntual de TMin
  - Menor o igual que TMax y TempAire Horaria
  - Correlación VS Distancia
  - Spatial Regression Test [Hubbard 4,5]
- Base Grillada de LST
  - Información Faltante
    - Reconstrucción Temporal
      - Promedio 8 días ventana fija (MOD11A2, MYD11A2)
      - Promedios de 3, 5, 7 y 9 días, ventana centrada
    - Reconstrucción Espacial
      - Cálculo de Climatologías
      - Regression Kriging utilizando la climatología como regresor
  - Control de calidad
    - Filtrado de Anomalías Espaciales
    - Nueva Reconstrucción Espacial

Modelado Geoestadístico

- Modelo universal de variación [Hengl, 1]
  - $Z(s) = Z^*(s) + \varepsilon'(s) + \varepsilon''$
    - $Z(s)$: variable objetivo en la ubicación s
    - $Z^*(s)$: componente determinista
    - $\varepsilon'(s)$: componente aleatoria espacialmente correlacionada
    - ε'': ruido blanco
- Kriging Ordinario
  - $\hat{Z}(s_0) = m + \varepsilon'(s_0)$
    - m: media constante (localidad) y desconocida
- Regression Kriging
  - $\hat{Z}(s_0) = f(s_0) + \varepsilon'(s_0)$
    - f: función de tendencia conocida para todo el dominio

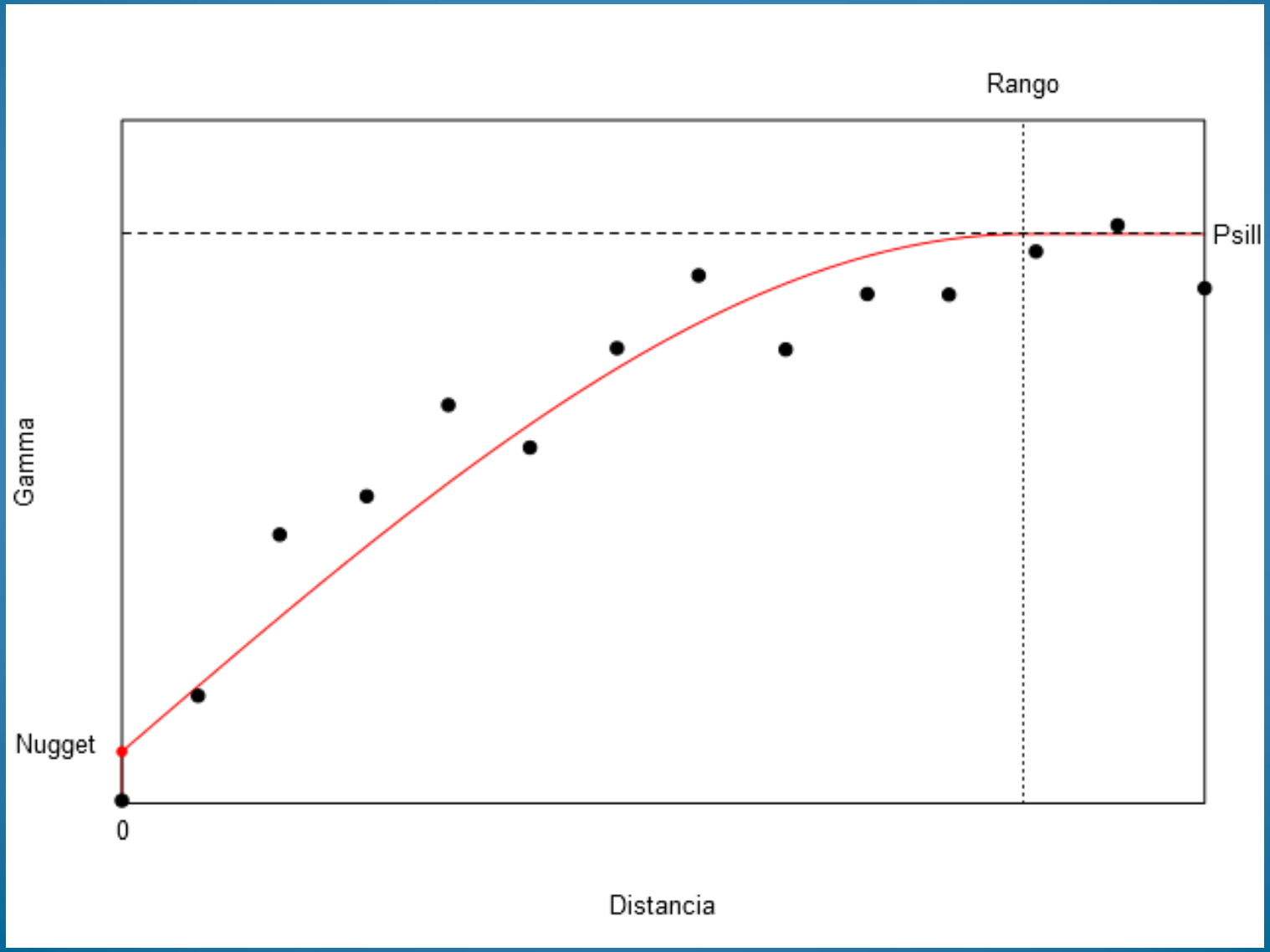
- Código base en los paquetes `gstat` e `intamap`
- Gráficos mediante `ggplot2`
- Ajuste de parámetros e hiperparámetros realizado en forma automática por la librería con código R de autoría propia

Kriging Ordinario

- Kriging Ordinario
  - $\hat{Z}(s_0) = \sum_{i=1}^{n} w_i(s_0) * z(s_i)$
- Variograma y Autocovarianza:
  - $\gamma(h) = \frac{1}{2} E[(Z(s_i) - Z(s_i + h))^2]$
  - $C(h) = sill - \gamma(h)$
- Pesos de Kriging Ordinario

$$\begin{pmatrix} C(s_1,s_1) & \cdots & C(s_1,s_n) & 1 \\ \vdots & \ddots & \vdots & 1 \\ C(s_n,s_1) & \cdots & C(s_n,s_n) & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} C(s_0,s_1) \\ \vdots \\ C(s_0,s_n) \\ 1 \end{pmatrix} = \begin{pmatrix} w_1(s_0) \\ \vdots \\ w_n(s_0) \\ \varphi \end{pmatrix}$$

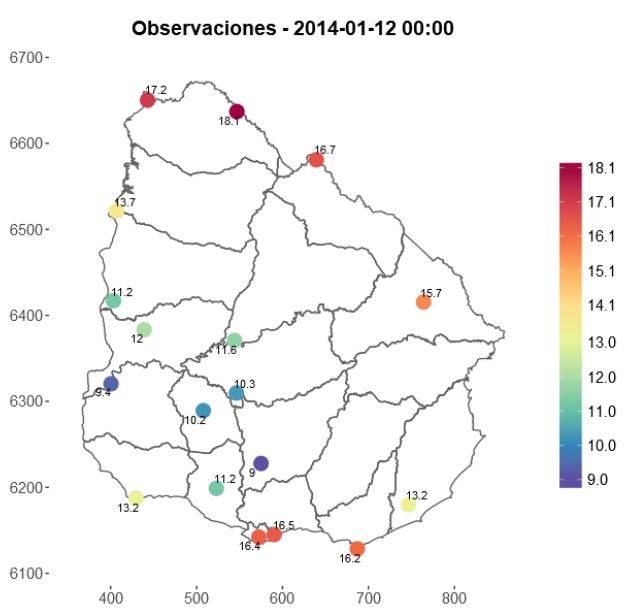
Variograma

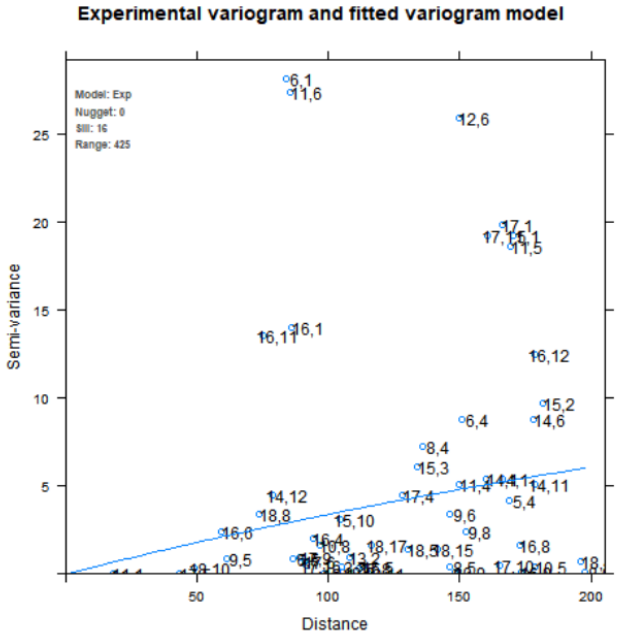


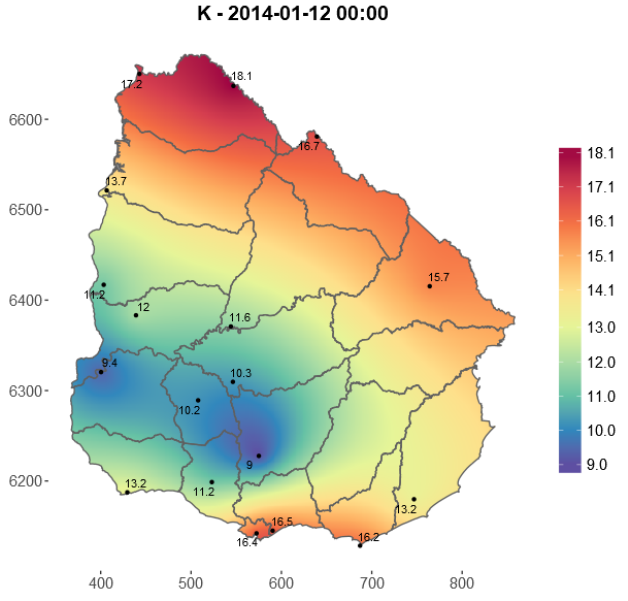
Kriging Ordinario

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY







Regression Kriging

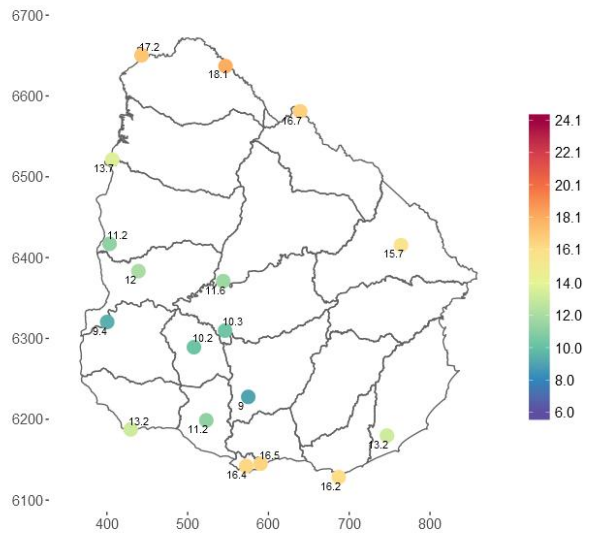
- Regression Kriging
  - $\hat{Z}(s_0) = f(s_0) + \varepsilon'(s_0)$
- En nuestro caso
  - $\hat{Z}(s_0) = \sum_{k=1}^{N_U} \hat{\beta}_k * U_k(s_0) + \sum_{i=1}^{n} w_i(s_0) * r(s_i)$
  - Con:
    - $r(s_i) = \mathrm{Z}(s_i) - \sum_{k=1}^{N_U} \hat{\beta}_k * U_k(s_i)$
    - $\hat{\beta}_k$ coeficientes de regresión: LM, RLM, GLS, …

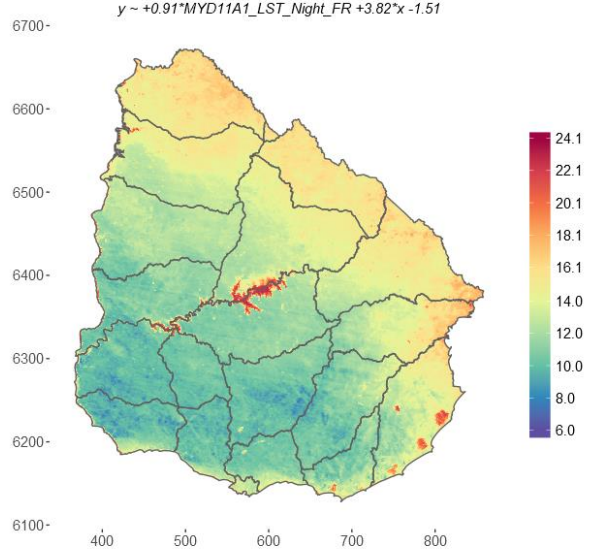
Regression Kriging

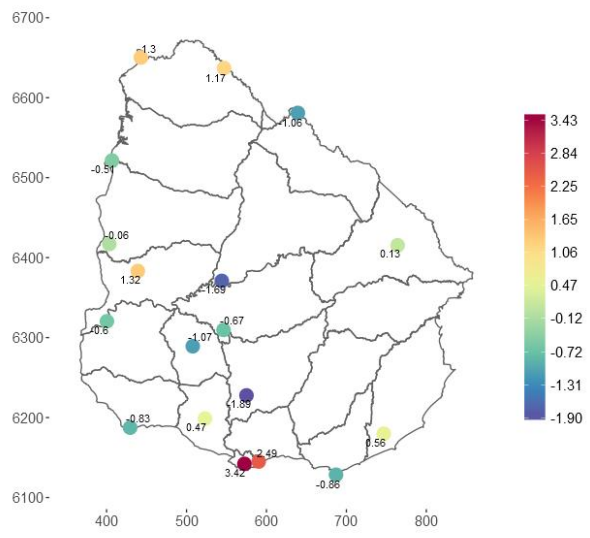
LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

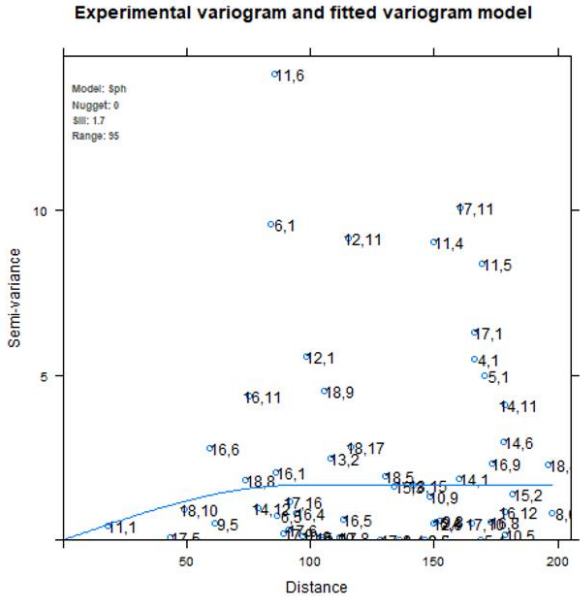
FACULTAD DE INGENIERIA

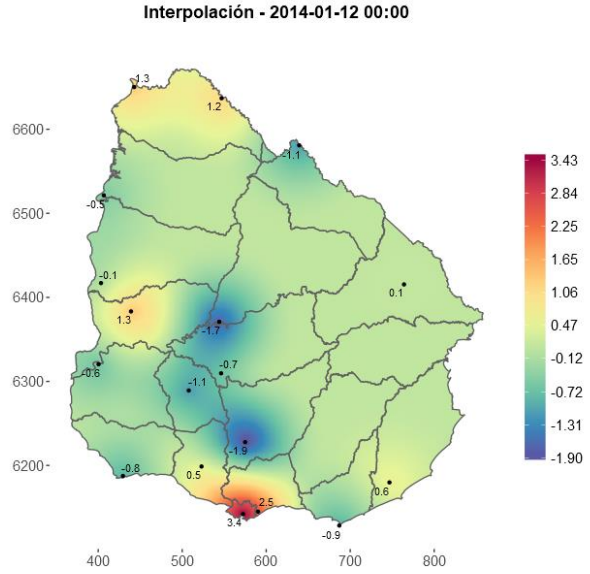
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

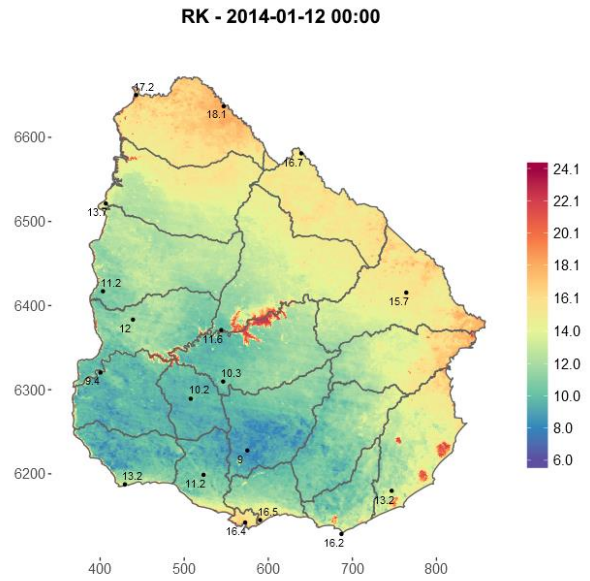








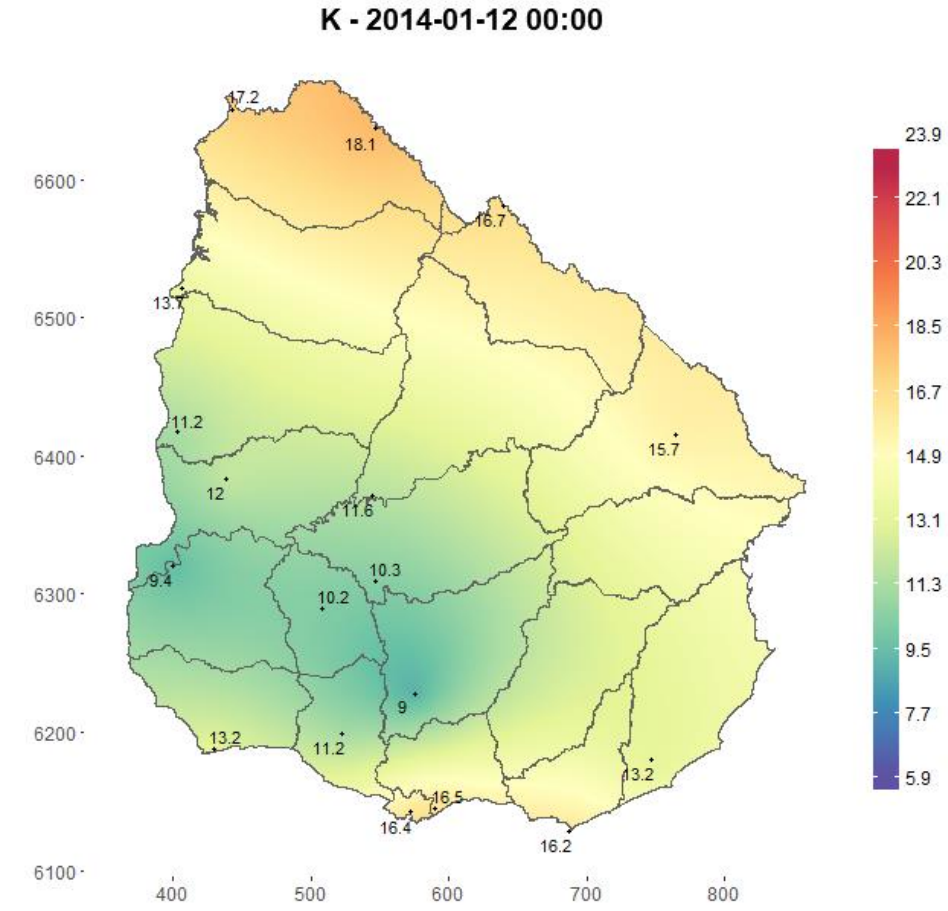


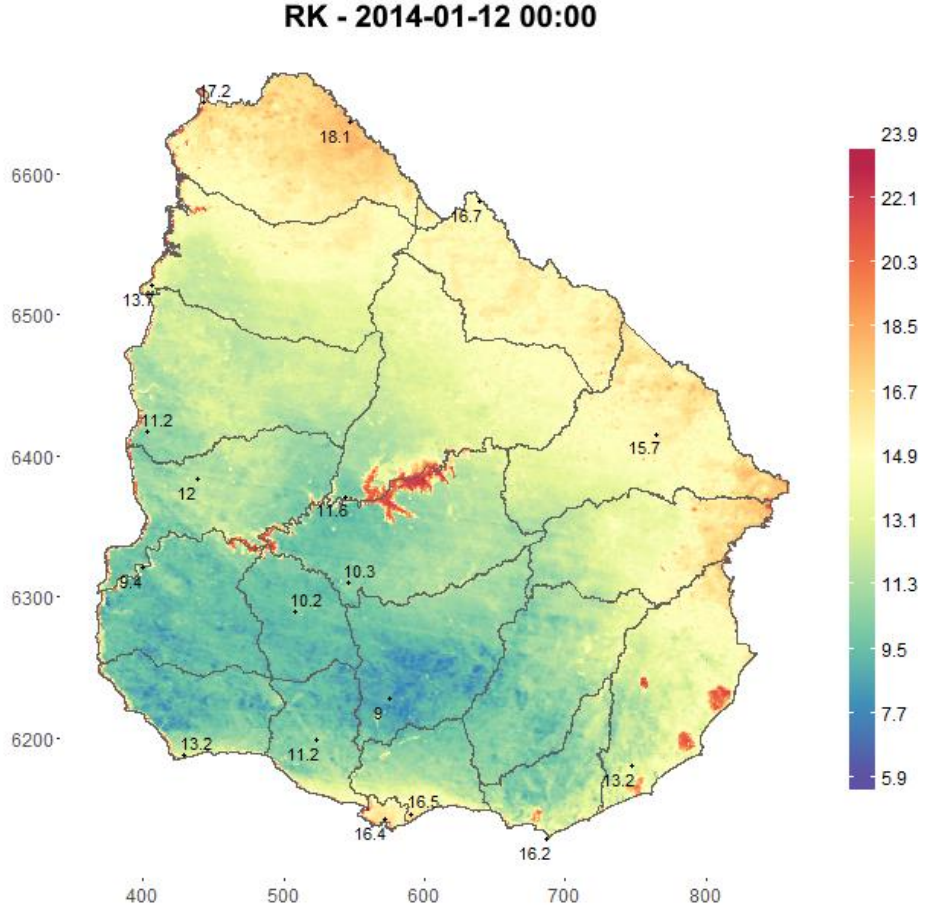


Comparación K vs RK

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY





Cross Validation

- Todos los modelos:
  - $\hat{Z}(s_0) = \mathrm{g}(s_0, Z(s_i)) \forall i \in \{1,2,\ldots,n\}$
- Leave One Out Cross Validation
  - $\hat{Z}_{CV}(s_i) = \mathrm{g}\big(s_0, Z(s_j)\big) \forall j \in \{1,\ldots,i-1,i+1,\ldots,n\}$
- Estadísticos de Validación
  - $e_i = \hat{Z}_{CV}(s_i) - Z(s_i)$
  - $ME = \frac{\sum_{i=1}^{n} e_i}{n}$
  - $MAE = \frac{\sum_{i=1}^{n} |e_i|}{n}$
  - $RMSE = \sqrt{\frac{\sum_{i=1}^{n} {e_i}^2}{n}}$

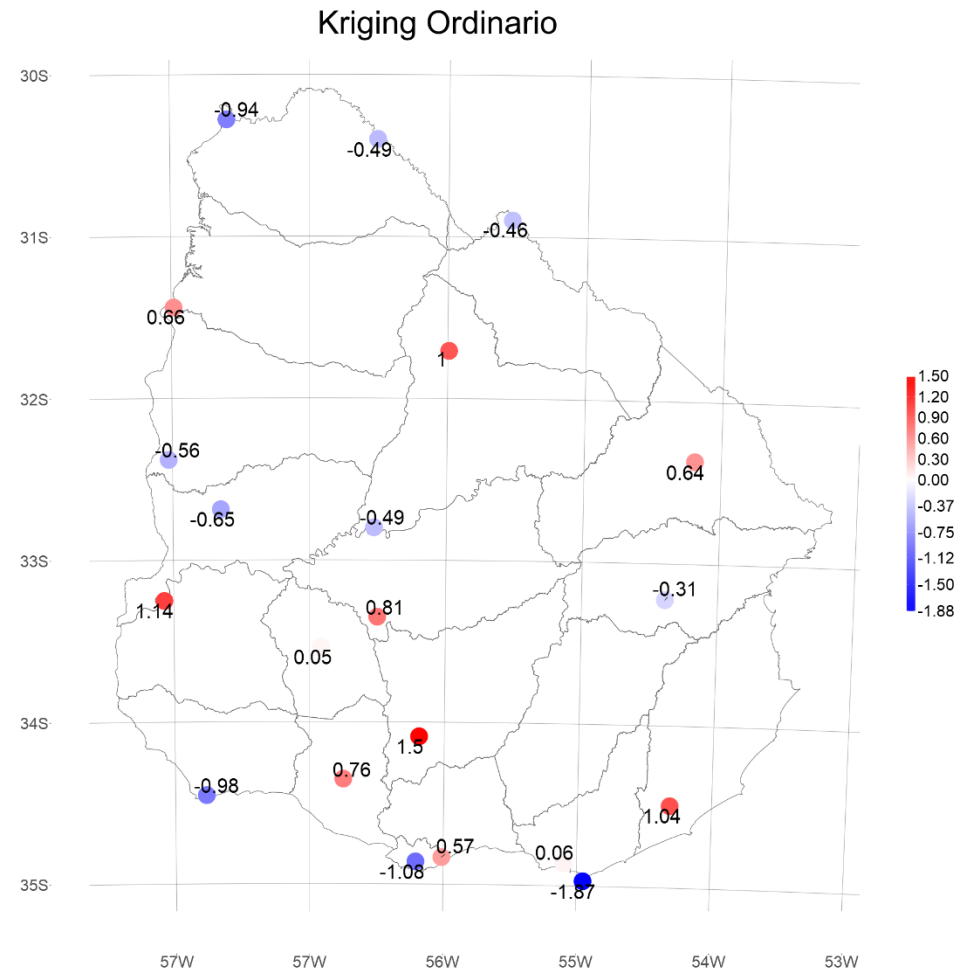
Resultados

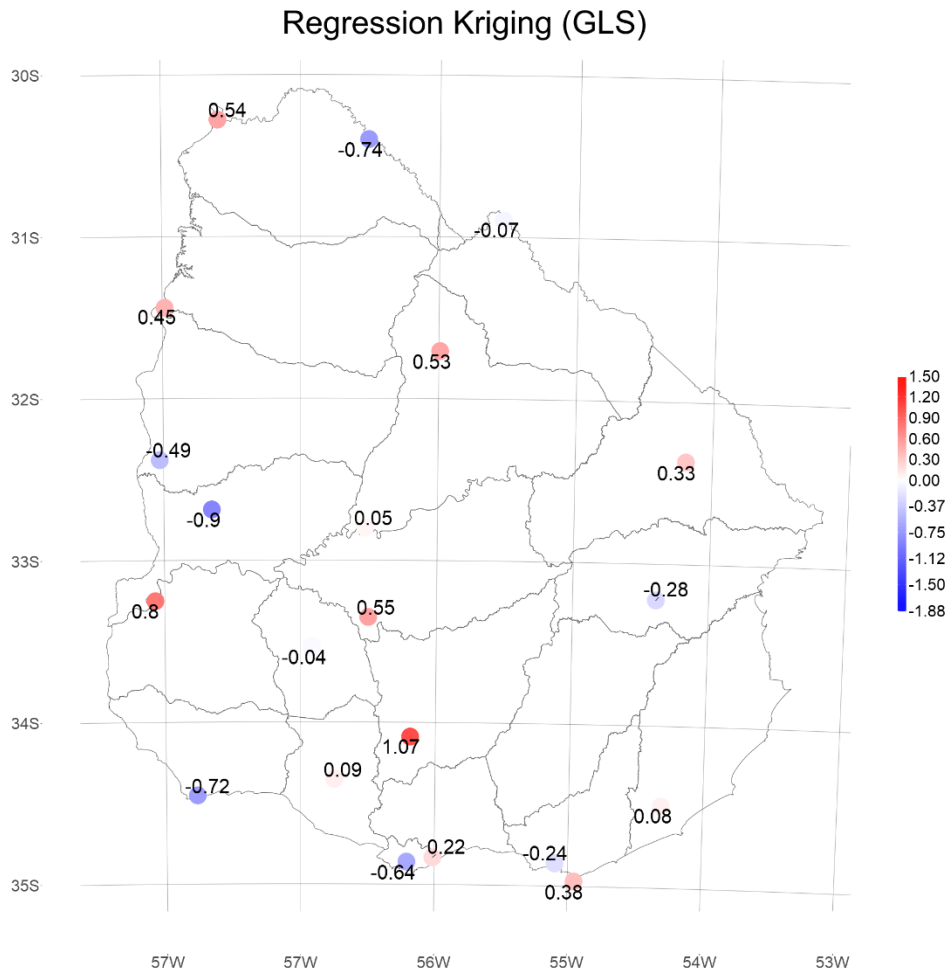
Modelo	ABS(ME)	MAE	RMSE	Corr
K	0.02	1.34	1.75	0.95
RK (RLM)	0.03	1.27	1.66	0.96
RK (GLS)	0.02	1.26	1.65	0.96

Resultados ME

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY





Aspectos Computacionales, Big Data, R

- Fue necesario descargar 165.8 GB de datos en 40874 archivos para disponer de la información satelital necesaria
- Paquete “rts”
  - Descarga a través de HTTP autenticado
  - Contribuciones al paquete:
    - Descargas paralelas y HTTP keepalive
    - Speedup combinado de 7X. Reduce tiempos de descarga de una semana a un día.
    - Además, mayor robustez a las descargas: detección de errores y reintentos ante descargas fallidas
  - Incorporadas a partir de la versión 1.0-43
- Otras contribuciones:
  - Paquete gstat: estimación de parámetros de variogramas teóricos mediante GLS para variogramas sin nugget [Müller 2, Genton 3]
  - Paquete intamap: habilitar pasaje de parámetros para cálculo de Kriging Simple
- Compresión
  - De los 165.8 GB se extrajeron 21.77 GB solo con la información de LST Nocturna
  - Usando compresión diferencial de geoTiffs la base se redujo a 2.72 GB
  - Razón de Compresión de 0.12

Aspectos Computacionales, Big Data, R

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- RAM necesaria para una grilla de datos con el dominio definido ~ 2 MB
- 4560 días en el período de trabajo
- (1 regresor + grilla de resultados) * 4560 días ~ 18.9 GB
- El tamaño es aún mayor con más regresores, un período más largo o un dominio más extenso
- Requerimiento de memoria no razonable para los asumidos de la biblioteca
- Big Data: algoritmos para trabajar con información que no cabe enteramente en memoria RAM
- La biblioteca implementada carga los datos a medida que los necesita y libera los recursos cuando no los necesita más

Aspectos Computacionales, Big Data, R

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Técnicas de HPC para mejorar performance
- Paralelismo
  - Descomposición de dominio
    - Modelos no autorregresivos: descomposición temporal
    - Modelos autorregresivos: descomposición espacial
  - Speedup lineal con la cantidad de procesadores
  - Uso de RAM por proceso ~ 600 MB
  - Solución escalable
- Cache de funciones complejas
  - Función de hash en base a los parámetros para generar un identificador de la llamada
  - Si el resultado se encuentra en disco se carga directamente, sino se evalúa la función y se guardan los resultados para futuras ejecuciones

Trabajo Futuro

- Mejoras al filtro de información satelital
  - Filtro Global
    - RLM contra el otro satélite
    - Data Interpolating Empirical Orthogonal Functions (DINEOF)
    - Detección de discontinuidades mediante nuggets
- Comparación de resultados con medidas de otros satélites
- Mejoras a la estimación de covarianzas para GLS
  - Actualmente solo exponencial y esférica
  - Paquete geoR permite trabajar con varios modelos más
- Aplicación de la técnica a TMax
  - 3 propuestas en [Ceccato, 7] implementables con la librería
- Incorporación de los resultados en productos elaborados

Muchas gracias por su tiempo!

- Alguna pregunta?

Referencias Bibliográficas

1. Hengl, T. (2009). A Practical guide to Geostatistical Mapping.
2. Müller. W. G. (1999), "Least-squares fitting from the variogram cloud," Stat. Probab. Lett., vol. 43, no. 1, pp. 93–98.
3. Genton. M. G. (1998), "Variogram fitting by generalized least squares using an explicit formula for the covariance structure," Math. Geol., vol. 30, no. 4, pp. 323–345.
4. Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N., & Osugi, T. T. (2005). Performance of Quality Assurance Procedures for an Applied Climate Information System. J. Atmos. and Oceanic Tech., 22, 105-112.
5. Hubbard, K. G., You, J. & Shulski, M. (2012). Toward a Better Quality Control of Weather Data. Practical Concepts of Quality Control. Nezhad, Mohammad Saber Fallah. Rijeka. InTech.
6. Ceccato, P., Vancutsem, C., and Temimi, M. (2010). Monitoring air and Land Surface Temperatures from remotely sensed data for climate-human health applications. Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International, pages 178–180.
7. Vancutsem C., Ceccato P., Dinku T., and Connor S. J., (2010). "Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa," Remote Sens. Environ., vol. 114, no. 2, pp. 449–465.
8. Kilibarda, M., Hengl, T., Heuvelink, G. B. M., Gräler, B., Pebesma, E., Percec Tadic, M., and Bajat, B. (2014). Spatio-temporal Interpolation of daily temperatures for global land areas at 1 km resolution. Journal of Geophysical Research Atmospheres, 119(5):2294–2313.
9. NASA LP DAAC (2001a). NASA Land Processes Distributed Active Archive Center (LP DAAC). MOD11A1, MOD11A2, MYD11A1, MYD11A2.
10. Wan, Z. (2009). MODIS Land Surface Temperature Products Users' Guide.
11. R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.