Análisis Geostadístico de Datos Funcionales

María José Ginzo Villamayor (mariajose.ginzo@usc.es) - Manuel Febrero Bande (manuel.febrero@usc.es)
Departamento de Estadística e Investigación Operativa (USC)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

Resumen

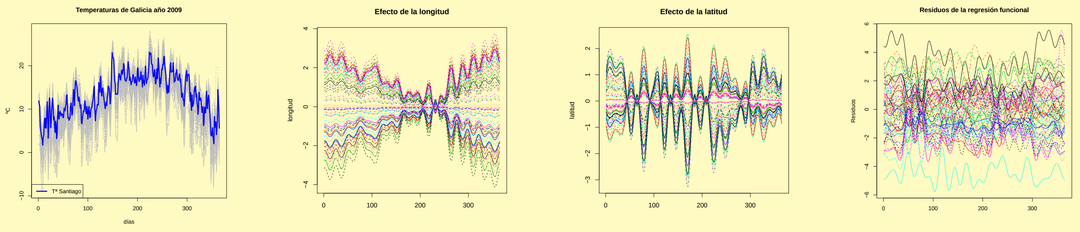
Este trabajo trata la predicción de curvas, cuando se dispone de una muestra de las curvas de una región con continuidad espacial. La motivación de este trabajo es ofrecer una solución al problema de predecir curvas en aquellas zonas no muestreadas de una región, basándonos en estas las ramas estadísticas, el análisis de datos funcionales y la estadística espacial.
Se revisan tres métodos para la predicción espacial de los datos funcionales.
Inicialmente, se propone un predictor que tiene la misma forma que el predictor kriging clásico, pero teniendo en cuenta las curvas en lugar de datos de una sola dimensión. Los otros predictores surgen de adaptaciones de modelos lineales funcionales con respuesta funcional en el caso de datos funcionales espacialmente correlacionados. Por un lado, se define un predictor que es una combinación de kriging y del modelo funcional lineal point-wise (concurrente). Por otra parte, se utiliza el modelo funcional lineal total para extender dos métodos clásicos geoestadísticos multivariantes para el contexto funcional. El primer predictor se define en términos de parámetros escalares. En el resto de los casos, los predictores implican parámetros funcionales. Se adapta un criterio de optimización, criterio utilizado en predicción espacial multivariante para estimar los parámetros escalares y funcionales que intervienen en los predictores propuestos. En todos los casos se da un enfoque no paramétrico basado en la expansión en términos de bases de funciones que se usa para obtener las curvas a partir de datos discretos.
De la misma manera que los métodos estadísticos estándares han sido generalizados para ser utilizados en FDA, es posible pensar que los métodos geostadísticos pueden ser adaptados a este tipo de datos.
Las metodologías propuestas se ilustran mediante el análisis de un conjunto de datos real correspondiente a la curva de temperatura que es función del tiempo.
En resumen, lo que contiene este proyecto fin de máster es una revisión crítica de los métodos que se han considerado previamente, en estadística espacial con datos funcionales y aplicados a un conjunto de datos real como es el de la temperatura en Galicia. Este conjunto de datos tiene tanto componente espacial y funcional.

Datos

Se dispone de datos de temperatura media del ambiente para Galicia en 66 estaciones distribuidas por Galicia del siguiente modo: 13 en la provincia de A Coruña (94 municipios), 20 en la provincia de Lugo (67 municipios), 15 en la provincia de Orense (92 municipios) y 18 en la provincia de Pontevedra (62 municipios), durante el año 2009. Los datos se obtuvieron de la página web de Meteogalicia - Xunta de Galicia (http://www.meteogalicia.es/web/index.action) y también las coordenadas geográficas de las estaciones meteorológicas. El punto marcado en rojo corresponde a la Facultad de Matemáticas, punto no muestreado. La máxima distancia de una estación a otra son 225.32Km y la máxima de cualquier estación al no muestreado es 155.82Km.

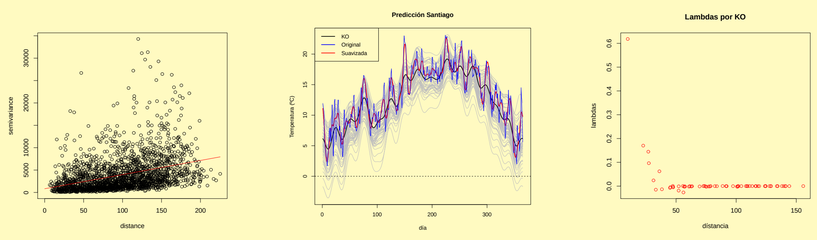
Técnica Descriptiva

Modelos con latitud y longitud:
$Temp_0(t) = \alpha_0(t) + \beta_1(t) \times Lat_0 + \beta_2(t) \times Long_0 + \hat{\epsilon}_0(t)$.
KU a los datos eliminándole a los residuos el efecto de la latitud y longitud.



Técnica 1 ($\lambda \in \mathbb{R}$)

Kriging ordinario para funciones-valores de datos espaciales.Procedimiento de kriging funcional donde la curva a predecir es c.l. de las curvas observadas y donde los coeficientes son **números reales**. Se aplica un ajuste no paramétrico al pre-proceso de las funciones observadas, mediante, bases de Fourier en la suavización. En el ajuste se usa un criterio de GCV lo que concluye que el número de bases consideradas sea 35.



De CV se concluye que la estación más influyente para la predicción en la Facultad de Matemáticas (s_0) es Sergude, la más cercana. Validación cruzada funcional (FCV) usado para para elegir el número de parámetros de suavización, también puede ser considerado una herramienta útil para comparar curvas observadas y predichas, define una medida de distancia entre estas dos curvas.

Referencias principales

[1] C. Comas, P. Delicado, R. Giraldo and J. Mateu. Statistics for spatial functional data, *Environmetrics*, 2009.

[2] H. Wackernagel. Multivariate Geostatistics, *Springer-Verlag*, 2nd edition, 1998.

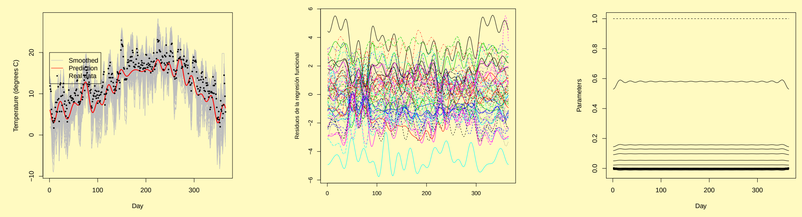
[3] J.O. Ramsay and B.W. Silverman. Functional Data Analysis, *Springer-Verlag*, 2nd edition, 2005.

[4] N. Cressie. Statistics for Spatial Data, *Probability and Statistics. Wiley*, 1993.

[5] R. Giraldo. A geostatistical analysis of functional data, *THESIS*, 2009.

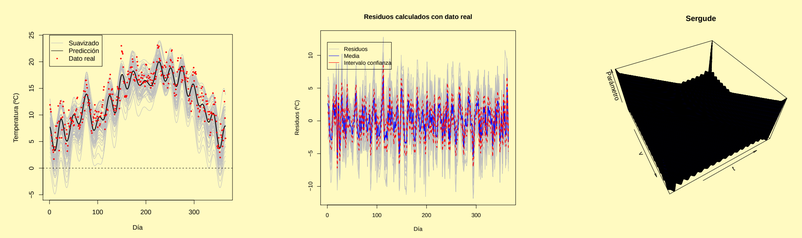
Técnica 2 ($\lambda(t) \in \mathbb{L}^2(t)$)

Kriging variación de tiempo continuo para la predicción espacial de datos funcionales. Se considera el problema de predicción espacial de datos funcionales con ponderación de cada curva observada por un parámetro funcional. Es una combinación del KO y el modelo funcional lineal concurrente (punto-wise). Se propone una solución basada en bases de funciones. Tanto las curvas como los parámetros funcionales se expanden en términos de un conjunto de bases de funciones.



Técnica 3 ($\lambda(s,t) \in \mathbb{L}^2(s) \times \mathbb{L}^2(t)$)

De multivariante a geostadística funcional. Se propone un predictor Cokriging haciendo una predicción univariante, considerando como información auxiliar funciones en lugar de observaciones de v. a. Extensión Kriging multivariante de v.a. al contexto funcional mediante la definición de un predictor Kriging funcional que permite hacer una predicción de la curva completa de la zona no muestreada mediante información de las curvas muestreadas a sitios cercanos al sitio de predicción. Flexibilidad ↑, mayor complejidad, el número de parámetros se elevan al cuadrado. Cada curva se pondera por un parámetro funcional para llevar a cabo la realización de la predicción en cada momento. Filosofía del modelo lineal funcional de respuesta funcional de $Y_i(t) = \beta_0(t) + \int_T X_i(v)\beta\,(v,t)\,dv + \epsilon_i(t)$, donde se debe estimar el coeficiente de regresión bivariante. Para cada caso, se estimó una LMC, para ajustarlo todos los variogramas modelados como una c. l. de modelos con efecto pepita y modelo exponencial.

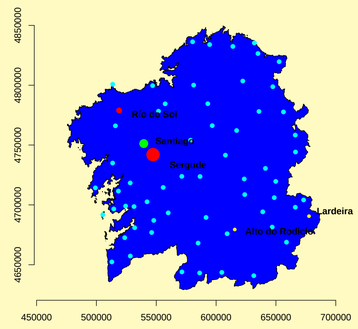


Se aprecia que la desviación estándar residual es menor en el otoño y en el invierno donde las curvas suavizadas y pronosticadas tienen menor variación.

La media residual varía alrededor de cero, lo que indica que las predicciones son insesgadas. Finalmente se ha aplicado esta técnica a los datos eliminándoles a los residuos el efecto de la latitud y longitud.

Conclusiones

Para comparación se el estadístico $SSE_{FCV}(i)$: $SSE_{FCV} = \sum_{i=1}^{n}\sum_{j=1}^{M}\left(\chi_{si}(t_j) - \hat{\chi}_{si}^{(i)}(t_j)\right)^2$.
Se evalua en $j = 1, \cdots, 365$ las predicciones obtenidas por VC para OKFD, CTKFD y FKTM. Efecto Pantalla. Se observa que donde existen las mayores diferencias entre los métodos son en términos de los valores mínimos o máximos y también influenciadas por el efecto de la longitud y la latitud.



- En el mapa el cículo verde, corresponde con Santiago, punto no muestreado. Los que están en rojo son los que tienen los autovalores más grandes, en cyan valores intermedios y amarillo residuos más grandes.

	OKFD	OKFD*	CTKFD*	FKTM	FKTM*
Mínimo	1326.0	812.6	865.3	1041.0	805.3
1st Qu.	2931.0	2017.0	2373.8	1971.0	1951.0
Mediana	3861.0	2626.0	2768.0	2554.0	2499.0
Media	4306.0	2990.0	3212.3	3030.0	3009.0
3rd Qu.	4847.0	3403.0	3793.2	3261.0	3468.0
Máximo	12880.0	11090.0	10061.0	10470.0	10180.0
Sd	2279.5	1551.3	1474.8	1852.1	1780.6
Suma	284222.1	197345.3	212009.0	199959.1	198569.1

(Nota: * método de la columna anterior sin los efectos de la longitud y la latitud).

- Los predictores tienen un comportamiento similar cuando las curvas son relativamente homogéneas. Las diferencias entre los mismos con los datos de temperatura para Galicia son fundamentalmente debidas a su desempeño en las estaciones con temperaturas extremas.
- Los resultados indican que la inclusión de un doble índice funcional en los parámetros (efecto temporal en el conjuntos de datos) no reflejan cambios sustanciales en las predicciones del análisis. El coste de estimar tantos parámetros no revierte en mejores resultados.
- Si a los datos se les elimina el efecto de la latitud y la longitud (estimando la tendencia) se obtienen mejores resultados.
- La mediana en el FKTM es menor y en el resto de estadísticos es mejor el OKFD*.

Motivación

Este póster es fruto del proyecto fin de máster de María José Ginzo Villamayor dirigido por el profesor Manuel Febrero Bande, como iniciación para futuros trabajos de investigación.