Impacto de la multicolinealidad en modelos jerárquicos Bayesianos ajustados con INLA a datos espaciales. Una aplicación a la degradación del suelo.

Tesista: Guillermina Senn

Directora: PhD MSc. Mónica Balzarini

Co-director: PhD MSc. Raúl Macchiavelli

Proyecto de tesis para optar al grado de
Magíster en Estadística Aplicada.

Universidad Nacional de Córdoba
Argentina
Junio 2021

Índice

1. Introducción

En en marco de una operación conjunta entre la Universidad Nacional de Córdoba, el Consejo Nacional de Actividades Científicas y Tecnológicas (CONICET) y la Comisión de Enlace de Entidades Agropecuarias (CEEA) Regional Córdoba se construyó un Sistema de Información Geográfica (SIG) entre los años 2020 y 2021 con gran cantidad de datos geo-referenciados del territorio conocido como Cuenca del Carcarañá, obtenidos mayoritariamente de imágenes por sensado remoto.

El cambio climático, la extensión de las coberturas antrópicas y la consecuente disminución de las coberturas naturales en una región que provee de servicios ecosistémicos a millones de habitantes justifica nuestro interés en analizar los datos de este GIS. Específicamente, nos interesa generar modelos estadísticos a escala de paisaje con potencial para predecir riesgos emergentes en la Cuenca del Carcarañá en relación con la degradación de recursos naturales y la producción agropecuaria.

Dentro de las técnicas estadísticas más difundidas para generar modelos predictivos se encuentra la regresión lineal. La regresión lineal predice la respuesta a partir de los datos utilizando una relación lineal en los parámetros y asumiendo que la respuesta observada es la realización de un vector de variables aleatorias normales e independientes (Casella y Berger, 2002). En el enfoque frecuentista, este modelo produce estimadores puntuales y un intervalo de confianza para los parámetros a través de estimación por mínimos cuadrados ordinarios (OLS, del inglés, Ordinary Least Squares) o máxima verosimilitud.

Los modelos de regresión también pueden enfocarse de manera Bayesiana. En contraste con el enfoque frecuentista, en el enfoque Bayesiano los parámetros son variables aleatorias cuya distribución se puede especificar antes de conocer los datos (Blangiardo y Cameletti, 2015). La inferencia Bayesiana usa el teorema de Bayes para combinar la verosimilitud de los datos con las distribuciones a priori de los parámetros, produciendo distribuciones a posteriori para los parámetros que son compromisos entre las distribuciones a priori y la evidencia presente en los datos. Las conclusiones estadísticas sobre un parámetro se hacen en términos de afirmaciones probabilísticas (Moraga, 2019).

La elección del tipo de distribución a priori de los parámetros y de los hiperparámetros - los

parámetros de estas distribuciones de segundo nivel - es vital. Si no se cuenta con información a priori sobre los parámetros, se debe usar una distribución a priori no informativa que tenga mínima influencia sobre la inferencia (Wang et al., 2018).

Calcular la distribución a posteriori de los parámetros implica generalmente realizar una integración altamente dimensional que no está disponible en forma cerrada. Este cálculo se hizo tradicionalmente con Cadenas de Markov - Monte Carlo (MCMC), que son métodos de muestreo computacionalmente demandantes (Moraga, 2019). En lugar de usar MCMC, en los modelos Gaussianos latentes se puede realizar una inferencia aproximada con aproximaciones de Laplace integradas anidadas (INLA). El método INLA, propuesto por Rue, Martino y Chopin (2009), tiene una performance superior a cualquier alternativa MCMC, tanto en términos de precisión como de velocidad computacional (Rue et al., 2009). INLA se puede utilizar en R fácilmente gracias al paquete R-INLA (Lindgren y Rue, 2015).

Cuando las variables respuesta están geo-referenciadas, las regresoras tienen una parte de variabilidad propia y otra causada por la dependencia espacial entre observaciones. En el modelo de regresión lineal normal, no reconocer la presencia de correlación positiva en los errores genera intervalos de confianza para los parámetros demasiado angostos y, por lo tanto, un aumento en el error Tipo I (Cressie, 1993; Beale et al, 2010). Cressie también indica que, desde el punto de vista predictivo, cuando las correlaciones espaciales decaen geométricamente con la distancia, los intervalos de predicción clásicos son frecuentemente válidos, pero pueden ser altamente ineficientes.

Para modelar la dependencia espacial se pueden usar modelos de regresión mixtos, que incorporan la variabilidad espacial utilizando efectos aleatorios. Los efectos aleatorios espaciales generalmente modelan la variabilidad a pequeña escala y por ello estos modelos se pueden ver como modelos con errores correlacionados (Krainski et al., 2019). Los mismos autores señalan que en los modelos de regresión mixtos para datos regionales los efectos mixtos tienen una matriz de precisión cuya estructura está dada por el hiperparámetro de precisión y una matriz de vecindades. A estos modelos se les puede dar un enfoque Bayesiano si se les asignan distribuciones a priori a los hiperparámetros y se los estima con MCMC o INLA. En este escenario, los modelos resultantes se llaman modelos de regresión jerárquicos Bayesianos.

Los datos de nuestro GIS no solo están geo-referenciados, si no que también muestran multicolinealidad. En el caso de la regresión clásica, la colinealidad genera una matriz mal condicionada en

el proceso de estimación que causa inestabilidad y mayor varianza en los coeficientes de regresión estimados (Belsley et al., 1980). En otras palabras, cuando hay un problema de colinealidad, la inferencia clásica, que no considera información a priori, falla como método para interpretar la evidencia de los datos (Leamer, 1973).

En la inferencia Bayesiana, cuando hay multicolinealidad, la distribución a posteriori puede ser altamente sensible a cambios en la distribución a priori; es decir, pequeños cambios en la distribución a priori podrían causar diferencias significativas en la distribución a posteriori (Leamer, 1973).

Resumiendo, la multicolinealidad en los datos es un problema en la inferencia de los modelos de regresión frecuentistas y Bayesianos, pero en el segundo caso se puede tratar si se tienen y se desean usar distribuciones a priori informativas para los parámetros del modelo (Belsley et al., 1980).

El efecto de la multicolinealidad en modelos de regresión para datos espaciales también ha sido en cierta medida estudiado con anterioridad. Por ejemplo, Beale et al. (2010), en un estudio con datos simulados, compararon dos modelos estimados con GLS que incluían la estructura espacial en el término de error, y encontraron evidencia de que las variables espaciales altamente cross-correlacionadas causan estimados imprecisos para los parámetros.

Sin embargo, la literatura sobre el impacto de la multicolinealidad en modelos jerárquicos Bayesianos, tanto sobre la inferencia como sobre la capacidad predictiva, es escasa, y es nuestro deseo ampliarla con este trabajo.

Objetivo general

El objetivo general de este proyecto de tesis es estudiar los efectos de la multicolinealidad en los modelos de regresión jerárquicos Bayesianos para datos espaciales, utilizando como aplicación los datos de tipo areal disponibles para la Cuenca del Carcarañá.

Objetivos específicos

1. Identificar técnicas para la selección de variables en presencia de colinealidad.

2. Encontrar el mejor modelo predictivo para los índices de servicios ecosistémicos ESPI, sCOS, COV y LPD.

3. Evaluar el impacto de la multicolinealidad en modelos de regresión jerárquicos Bayesianos para datos espaciales areales.

2. Metodología

2.1. Datos

El área de estudio comprende el territorio del sistema hidrológico del Carcarañá dentro de los limites jurisdiccionales de la Provincia de Córdoba. Se recolectaron diversas capas de datos georreferenciados comunicados en diversas fuentes de dominio público para construir un Sistema de Información Geográfico (SIG) que alberga variables edáficas, climáticas, topográficas y de vegetación para la Cuenca del río Carcarañá. Los datos fueron depurados y re-escalados previo a la conformación del SIG. Se describieron con medidas resumen las unidades de gestión hídrica existentes en la cuenca.

Se realizó una segmentación multidimensional usando cinco capas de información para obtener unidades homogéneas en cuanto a clima, suelo, topografía y vegetación. La optimización del proceso de segmentación arrojó un total de 4676 unidades homogéneas.

Para cada unidad se calcularon tres índices que describen servicios ecosistémicos: índice de fertilidad de los suelos (IP), almacenamiento de carbono orgánico en suelo (sCOS) y almacenamiento de carbono en biomasa vegetal (COV). Además, se calculó un índicador global de SE basado en la dinámica de series de 30 años de NDVI, llamado ESPI, del inglés Ecosystem Services Provisioning Index. La dinámica expuesta en series largas de NDVI también se resumió con otro índice, denominado LPD, del inglés Land Productivity Dynamics.

2.2. Procedimientos

Evaluación de la capacidad predictiva

Para las cuatro variables respuesta (LPD, ESPI, sCOS y COV) se buscará mejor modelo predictivo. La capacidad predictiva del modelo se evaluará comparando el error de predicción entre modelos con métodos de cross-validación. En el caso Bayesiano, el predicho puntual se calculará como la esperanza de la distribución a posteriori del predicho. Además, en los modelos espaciales, se calculará y comparará la varianza del error de predicción.

Técnicas de selección de variables

El paso de selección de variables puede ser muy importante en un modelo predictivo. En este trabajo proponemos dos mecanismos para seleccionar variables.

El primer mecanismo será para disminuir el número de variables y obtener un modelo parsimoniosos. Este mecanismo se definirá luego de una revisión de la literatura más profunda.

El segundo mecanismo será para tratar las variables multicolineales. Utilizaremos análisis de componentes principales (PCA) de dos maneras:

- Aplicar PCA a los subconjuntos de variables altamente multicolineales y reemplazar estos subconjuntos con las primeras componentes principales obtenidas para cada subconjunto. De esta manera eliminamos gran parte de la multicolinealidad en los datos.
- Aplicar PCA a todas las p covariables disponibles y reemplazar las p covariables con las p componentes principales ortogonales. De esta manera eliminamos toda la multicolinealidad de los datos.

Se calculará la capacidad predictiva del modelo obtenida luego de aplicar cada técnica de selección. Reportaremos las mejores técnicas como aquellas que al ser aplicadas mejoren la capacidad predictiva de los modelos en mayor medida.

Evaluación del impacto de la colinealidad en la regresión Bayesiana jerárquica

Para evaluar el impacto de la colinealidad en modelos espaciales proponemos mirarlo desde dos puntos de vista:

1. Quitando la colinealidad de los datos. En este caso se aplicará PCA a los datos siguiendo la metodología propuesta para la selección de variables.
2. Utilizando distribuciones a priori informativas y no informativas para los parámetros. Los datos se mantendrán inalterados.

En ambos casos se comparará la capacidad predictiva de los modelos siguiendo la metodología propuesta previamente para comparar la capacidad predictiva. En caso de que la capacidad pre-

dictiva se vea inalterada en todos los casos, adicionalmente evaluaremos el sesgo y la varianza en las distribuciones a posteriori para los parámetros. Además de la regresión Bayesiana jerárquica ajustaremos modelos de regresión normal por OLS con fines de referencia.

La combinación de todos los posibles escenarios sugiere una grilla de modelos como la que se muestra en el Cuadro 1.

Para cada una de las cuatro variables respuesta (LPD, ESPI, sCOS y COV) se ajustará cada una de las posibles combinaciones en la grilla de modelos. Cabe aclarar que, como se utilizarán parcialmente los mismos datos para modelar las distintas respuestas, y la multicolinealidad es una propiedad de los datos, no se pueden considerar como cuatro escenarios distintos para evaluar la multicolinealidad.

ID	Modelo	Efectos espaciales	Selección de variables y multicolinealidad
1	OLS	No	(SS) Sin selección
2	OLS	No	(PCA1) PCA sobre los grupos multicolineales
3	OLS	No	(PCA2) PCA sobre todos los datos
4	OLS	No	(PAR) Selección para modelo parsimonioso
5	HB	No	(SS) Sin selección
6	HB	No	(PCA1) PCA sobre los grupos multicolineales
7	HB	No	(PCA2) PCA sobre todos los datos
8	HB	No	(PAR) Selección para modelo parsimonioso
9	HB	No	(UP) Distribución a priori no informativa
10	HB	No	(IP) Distribución a priori informativa
11	HB	Si	(SS) Sin selección
12	HB	Si	(PCA1) PCA sobre los grupos multicolineales
13	HB	Si	(PCA2) PCA sobre todos los datos
14	HB	Si	(PAR) Selección para modelo parsimonioso
15	HB	Si	(UP) Distribución a priori no informativa
16	HB	Si	(IP) Distribución a priori informativa

Cuadro 1: Grilla de modelos

3. Referencias

Beale, C., Lennon, J., Yearsley, J., Brewer, M., Elston, D. (2010). Regression analysis of spatial data. Ecology Letters, Volume 13, Issue 2 p. 246-264

Belsley, D., Kuh, E., Welsch, R. (1980). Regression diagnostics: identifying influential data and sources of collinearity. John Wiley and Sons.

Blangiardo, M., Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley and Sons.

Casella, G., Berger, R. L. (2002). Statistical inference. Belmont, CA: Duxbury.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H. (2018). Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall/CRC.

Leamer, E. (1973). Multicollinearity: A Bayesian Interpretation. The Review of Economics and Statistics, MIT Press, vol. 55(3), pages 371-380, August.

Lindgren, F., Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software, 63(19), 1 - 25. doi:http://dx.doi.org/10.18637/jss.v063.i19

Moraga, P. (2019). Geospatial health data: Modeling and visualization with R-INLA and shiny. CRC Press.

Rue, Håvard, Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series b (Statistical Methodology), 71(2), 319–392.

Wang, X., Ryan, Y. Y., Faraway, J. J. (2018). Bayesian Regression Modeling with INLA. Chapman and Hall/CRC.