Artículo de Investigación. Revista Ciencia e Ingeniería. Vol. 39, No. 2, pp. 199-212, abril-julio, 2018.
ISSN 1316-7081. ISSN Elect. 2244-8780. Universidad de Los Andes (ULA)

Exploración espacio temporal de la distribución de datos faltantes de precipitación mensual en el centro occidente de Venezuela, con fines de selección de estaciones

Space-Time exploration of the distribution of missing data for the selection of monthly precipitation stations in central western of Venezuela

Andrades-Grassi, Jesús Enrique* [1]**; Torres-Mantilla, Hugo Alexander** [2]**; López-Hernández, Juan Ygnacio** [1]**; Goitía-Acosta, Arnaldo**[3]**, Mejías-Delgado, Jesús Enrique** [1]

[1] Universidad de Los Andes, Facultad de Ciencias Forestales y Ambientales, Departamento de Ordenación de Cuencas Mérida, 5101, Venezuela;

[2] Universidad de Santander, Facultad de Ciencias Exactas, Naturales y Física, Departamento de Matemáticas y Física Bucaramanga, Colombia;

[3]Universidad de Los Andes, Facultad de Ciencias Económicas y Sociales, Instituto de Estadística Aplicada y Computación Mérida, 5101, Venezuela

koflasjesus@gmail.com*

Resumen

Los datos faltantes se definen como un mecanismo de no respuesta de una variable o una serie de variables, en este caso se exploró la distribución espacial y temporal de los datos faltantes de precipitación mensual en el centro occidente de Venezuela, con el fin de seleccionar las estaciones y la serie temporal más adecuada. Se plantearon diversas preguntas espacio-temporales generales e individuales que derivaron en la aplicación de técnicas espaciales y temporales de Análisis Exploratorio de Datos. Los resultados indican de un Mecanismo No Aleatorio de Datos Faltantes (MNAR) Espacio-Temporal y el diagnostico de los denominados Block Missing Data. Sin embargo, se logró realizar una selección espacial de un conjunto de 88 estaciones que aproximadamente tienen un comportamiento de un Mecanismo Completamente Aleatorio de Datos Faltantes en el Espacio-Tiempo. Se recomienda trabajar con un conjunto de 42 estaciones de precipitación mensual ubicadas en 5 estados de Venezuela, ya que el comportamiento de la mismas sigue una Distribución Espacial Aletoria.

Palabras clave: Datos Faltantes, Espacio-Tiempo, Análisis Exploratorio de Datos.

Abstract

The missing data values are defined as a non-response mechanism of a variable(s), in this case we explored the spatial and temporal distribution of the missing data values of monthly precipitation in central western Venezuela, in order to select the stations and the most appropriate time series. General and individual space-time questions were raised that resulted in the application of spatial and temporal techniques of Exploratory Data Analysis. The results indicate a Non-random Missing Data Mechanism (MNAR) and the diagnosis of the so-called Block Missing Data. However, it was possible to perform a spatial selection of a set of 88 stations that approximately have a Missing Completely at Random (MCAR) in Space-Time, it is recommended to work with a set of 42 monthly precipitation stations located in 5 states of Venezuela because follows a random spatial distribution.

Key words: Missing Data, Space-Time, Exploratory Data Analysis.

1 Introducción

El manejo de datos espacio temporales ha sido definido por diversos autores como la nueva frontera de la ciencia. Cressie y col. (2015) llaman a este tipo de análisis como el *Santo Grial de la Ciencia*, sin embargo, este tipo de análisis resulta de alta complejidad, debido a que incorpora las múltiples dimensiones y datos altamente volumétricos.

Una gran cantidad de observaciones se pueden obtener a partir de numerosas ubicaciones espaciales a través del tiempo; según Gujarati y col. (2010), estos corresponden estadísticamente con una estructura de datos agrupados (series de tiempo con datos transversales en el espacio). Esto significa que los datos son una agrupación de matrices de observaciones que combina los datos transversales en N unidades espaciales y períodos de tiempo T, para producir un conjunto de datos de observaciones NxT Podestà (2002), Stadelmann-Steffen y col. (2008) (figura 1). La gestión de los valores faltantes (también conocidos como valores perdidos) es importante cuando se realiza un análisis de datos porque, a pesar de que los valores no son muy abundantes, la mayoría de los métodos estadísticos asumen una matriz de datos completa. La imputación tiene el objetivo de completar los datos mediante la sustitución de los valores perdidos por valores válidos estimados, preferentemente utilizando un camino sin sesgo y computacionalmente eficiente. Con los datos agrupados espacio temporales (comunes en los estudios hidrológicos, climáticos y muchos otros campos de investigación ambiental y territorial) surge un problema para la imputación de los datos, ya que la falta de exhaustividad de los datos plantea retos especiales para el análisis estadístico y de modelado espacio-temporal Schneider (2001).

A partir de los datos espacio temporales se pueden plantear múltiples preguntas. Peuquet (1994) menciona que este tipo de datos se caracteriza por poseer tres componentes: Espacio (Dónde), Objeto (Qué) y Tiempo (Cuándo). Basado en estos componentes Andrienko y col. (2003), clasifican los datos espacio-temporales de acuerdo al tipo de cambio en el tiempo, estos autores reconocen los cambios existenciales, como lo es la aparición y desaparición de un objeto, que puede ser asociado con el componente tiempo. Estos tres componentes pueden trabajarse de forma general, teniendo en cuenta el grupo total de datos, o de forma elemental, individualizando el abordaje de los objetos de trabajo.

Los datos faltantes o conocidos también como *Missing Values*, formalmente se definen como valores *No-data* o mecanismos de no respuesta que no son almacenados en la variable de observación, Enders (2010), Daniels y col. (2008). Las series históricas de clima completas y de períodos largos son un requisito para la elaboración de estudios confiables climáticos, meteorológicos, hidrológicos, entre otros. En la medida que los datos sean el resultado de un proceso veraz y bien planteado, con registros periódicos que den continuidad a las observaciones, será factible obtener resultados ajustados a la realidad del fenómeno que se estudia.

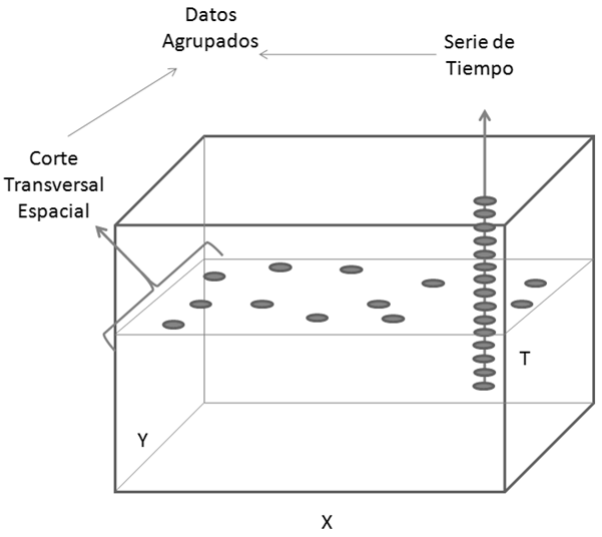


Fig. 1: Estructura de Datos Agrupados.

Sin embargo, existen situaciones que pueden afectar la calidad de los datos o registros de las series históricas de precipitación, tales como el cambio de lugar o movimiento de instrumentos (estaciones climáticas), la transformación del espacio físico del entorno donde se encuentra la estación, o no registrar el dato en el momento adecuado, lo que conlleva a la inconsistencia o ausencia de datos, Medina y col. (2008). Los datos faltantes son comunes en los datos climáticos, ellos afectan los modelados e inferencias de procesos usuales y relevantes como lo son el análisis de oferta hídrica. La exploración de los datos faltantes es un proceso clave y fundamental en el pre-procesado de los datos climáticos y su posterior modelamiento, ya que la selección de las estaciones y la serie temporal de trabajo están condicionadas por los mismos y los cambios existenciales, como lo es la aparición y desaparición de una estación o el daño de un instrumento; de no considerarse esta situación, las estimaciones pueden ser erróneas o poseer varianzas de grandes proporciones que aumenten la incertidumbre de los datos estimados. Un condicionante importante a evaluar es que siempre una estación climática tiene una localización cartográfica fija, y los datos climáticos de la estación corresponden a medidas repetidas en una nueva dimensión temporal de la variable de análisis. Es por ello, que se introduce el concepto de que el análisis de datos perdidos debe manejarse de forma espacio temporal, ya que incluirían nuevas herramientas en el proceso de selección de las estaciones y la serie operativa de trabajo.

En este caso se desea explorar la distribución espacial y temporal de los datos faltantes de precipitación mensual en el centro occidente de Venezuela, con el fin de seleccionar las estaciones y la serie temporal más adecuada. Ello implica que deben considerarse preguntas y tareas espacio temporales

en el proceso de selección, así como diversas metodologías estadísticas para el análisis de los mismos. Se pretende con ello llegar a una selección de estaciones y período de trabajo que afecte el mínimo la escala (que utilice el máximo número de estaciones), que considere una distribución espacial de los datos, y que la cantidad de datos temporales sea lo máximo posible. Todo ello utilizando técnicas de exploración espacio temporal.

2 Procedimiento Experimental

En Venezuela el Instituto Nacional de Meteorología e Hidrología (INAMEH), dispone a nivel nacional de 2471 estaciones y el portal oficial del Instituto Nacional de Investigaciones Agrícolas (INIA) tiene 20 estaciones adicionales, lo que da un total de 2491 estaciones con datos oficiales de tipo Geocodificación. Utilizando los datos oficiales de precipitación mensual disponibles en estos portales, se trabajó con un total de 961 estaciones de precipitación mensual, que corresponden al área de interés de la zona Centro Occidental de la República Bolivariana de Venezuela; los datos de trabajo corresponde a 11 estados: Táchira, Mérida, Trujillo, Barinas, Portuguesa, Cojedes, Yaracuy, Carabobo, Lara, Falcón y Zulia, (figura 2). Los datos, según Gujarati y col. (2010), corresponden estadísticamente con una estructura de datos agrupados, (series de tiempo con datos transversales en el espacio). Se está trabajando con una variable dicotómica ó binaria, de ausencia y presencia de datos. Dado que se está explorando espacial y temporalmente los datos faltantes, en primer lugar partiendo de las tipologías y niveles de preguntas planteadas por Andrienko y col. (2003), Li y col. (2008) se plantearon preguntas espacio temporales.

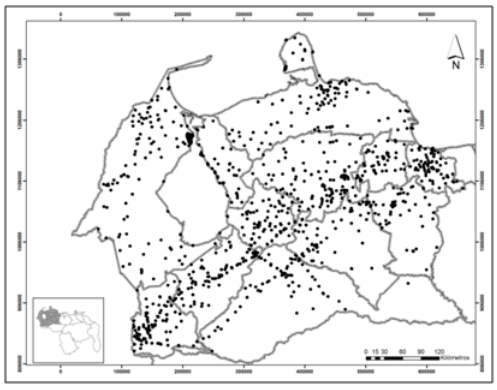


Fig. 2: Estaciones de precipitación mensual seleccionadas.

2.1 Pregunta referida al objeto: «Qué» General

2.1.1 Número mínimo de datos para ajustar un modelo paramétrico

El ajuste de un modelo general estocástico paramétrico tiene como restricción un número datos mínimos de acuerdo al número de parámetros a estimar. Aunque no existe un acuerdo sobre límite mínimo de datos para un modelo ARIMA o algún modelo espacio temporal, Gujarati y col. (2010) mencionan que el número de observaciones n debe ser mayor que el número de parámetros por estimar. Esto es refrendado por Hyndman y col. (2007), cuando mencionan que desde un punto de vista meramente estadístico, siempre es necesario contar con más observaciones que parámetros.

Los modelos ARIMA estacionales se describen utilizando la notación $ARIMA(p,d,q)(P,D,Q)_n$, donde cada una de las letras dentro del paréntesis indica algún aspecto del modelo. Un modelo ARIMA estacional posee $p+q+P+Q$ parámetros, sin embargo, en ellos se requiere de diferenciaciones adicionales $d+mD$ para lograr que la varibale sea estacionaria en segundo orden (medias y varianzas constantes), por lo tanto en este último proceso se pierden observaciones. Así, un total de $p+q+P+Q+d+mD$ parámetros efectivos se usan en el modelo. En consecuencia, para el ajustar correctamente un modelo ARIMA estacional se requieren al menos $p+q+P+Q+d+mD+1$ observaciones. Por ejemplo, el modelo que de series de tiempo ARIMA de Box y col. (2015) que utilizó datos de aviación, es un modelo mensual ARIMA $(0,1,1)(0,1,1)_{12}$ y por lo tanto contiene $0+1+0+1+1+12=15$ parámetros, por lo que se necesitan de al menos de 16 observaciones.

Partiendo de este criterio, dado que se está manejando la posibilidad de un modelo espacio temporal de series de tiempo como lo es el Spatio Temporal Autorregresive Model (STARIMA), desarrollado por Pfeifer y col. (1980) y descrito apliamente por Cressie (1993), Cressie y col. (2015), Kamarianakis y col. (2003), Kamarianakis y col. (2005). Si se utiliza este caso el modelo más sencillo de STARIMA, es decir,un modelo $STARIMA(1_1,0,1_1)$, que está especificado de la siguiente manera: $y_t=\phi_1 y_{(t-1)}+\phi_{11}W_1 y_{(t-1)}+\theta_{10}\varepsilon_{(t-1)}+\theta_{11}W_1\varepsilon_{(t-1)})+\varepsilon_t$. Donde: ϕ_{10} y θ_{10} son los coeficientes AR y MA con un retardo temporal, ϕ_{11} y θ_{11} son los coeficientes espaciales escalares para el primer retardo espacial y W_1 corresponde con la matriz de ponderación espacial. Por lo cual, ya que existen 10 parámetros, se requieren $10+1$ observaciones como mínimo, es decir 11.

Ahora, suponiendo para los datos de precipitación mensual (datos fuertemente estacionales), con una covaribale (altitud), y se desconoce del número de parámetros y la complejidad del modelo STARIMA, se debe seleccionar el modelo más complejo posible a ser ajustado en el que se sobreestimen los parámetros, en este caso un $STARIMA(2_2,2,2_2)(2_2,2,2_2)_{12}$, por lo tanto contiene $4+4+2+4+4+2x12+1=42$ parámetros, lo que quiere decir que se requieren como mínimo 43 observaciones, sin interrupciones para ajustar un STARIMA $(2_2,2,2_2)(2_2,2,2_2)_{12}$ sin covariable, sumando la covariable altitud entonces se requieren de 44 observaciones ininterrumpidas por estación.

Adicionlmente también se pueden utilizar modelos tipo Models For Multiple Spatial Time Series Relations, en

este caso se presenta un modelo autoregresivo de primer orden $Y_t = \tau Y_{(t-1)} + \delta W Y_t + \eta W Y_{(t-1)} + \beta_1 X_t + \beta_2 X_{(t-1)} + \beta_3 W X_t + \beta_4 W X_{(t-1)} 6 + \varepsilon_t$ Kamarianakis y col. (2003), Kamarianakis y col. (2005), López y col. (2005), López y col. (2007). Suponiendo ahora el STARIMA $(2_2, 2, 2_2)(2_2, 2, 2_2)_{12}$, con la mezcla de la covariable altitud, con un retardo espacial, que introduzca una nueva matriz de ponderación espacial de tal manera que $z_t = \rho_1 W_2 z_t + \rho_2 X_t + \rho_3 W_2 X_t + \varepsilon_s$ Anselin (1988), Anselin (2005), Anselin (1999), Anselin y col. (2013), Anselin y col. (2010), Cressie (1993), por lo tanto se necesitan las 43 observaciones del STARIMA$(2_2, 2, 2_2)(2_2, 2, 2_2)_{12}$ sin la covariable, mas 5 observaciones del modelo con retardo espacial con covariable, sumando así 48, por lo que se necesitan 48 observaciones mínimo. Ahora, suponiendo un modelo Panel de tipo Seemingly Unrelated Regressions (SUR) (Regresiones aparentemente no relacionadas) $y_i = x_i B_i + \varepsilon_i$ donde:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_m \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} \quad (1)$$

El cual es aplicable cuando el número de observaciones temporales es mayor que las espaciales, es decir, $T > N$ Anselin (1988). Siendo el número de parámetros dependiente del número de niveles (NIV) que se deseen trabajar, por lo tanto el número de datos mínimos será $(NIV) + 1$. Sin embargo, lo anterior se refiere a que sería imposible estimar los parámetros si no se cumple con ese número de datos mínimo. Aún cuando se cumpla con ese mínimo número de datos, para que la estimación sea adecuada se debería contar con suficiente cantidad de datos para cada parámetro, que podría variar subjetivamente según el modelo entre 10 a 30 datos por cada parámetro Hair J y col. (2010). Finalmente, los datos mínimos necesarios en cuestión, en particular para modelos que involucran relaciones temporales o espaciales en estructura Lattice, deben encontrarse en una secuencia ininterrumpida o continua. Para el caso de interrupciones o datos faltantes aislados se podrían intentar solventar mediante técnicas de imputación cualesquiera, teniendo en cuenta que cada dato imputado involucra un dañino aumento en el sesgo en las estimaciones finales.

Partiendo de esto se planteó la pregunta referida al Qué y se decidió colocar tres restricciones condicionadas por un máximo 30 %, 25 % y 20 % de datos faltantes para cada una de las estaciones (recuérdese que se desea maximizar periodo y el uso de estaciones y minimizar el número de datos a imputar), basado en esto, se definieron 4 periodos de tiempo tentativos y para cada uno de ellos la cantidad máxima de datos faltantes tolerada (tabla 1).

Posteriormente se decidió evaluar la estructura de los datos faltantes realizando el test de Little para evaluar la ocurrencia de un Mecanismo Completamente Aleatorio de

Tabla 1: Máximo número de datos faltantes tolerados por estación.

Periodo (años)	**30 %**	**25 %**	**20 %**
1949-2008	212	177	142
1949-2005	202	168	134
1949-2000	184	153	122
1949-1998	176	147	118

Valores Faltantes (MCAR), (se dedinen tres tipos de mecanismos estocásticos de datos faltantes; para ello ante de describirlos deben plantearse unos elementos relevantes, $Ycom = (Yobs, Ymis)$, en donde $Ycom$, $Yobs$ y $Ymis$ corresponden al vector de datos completos, observados y faltantes respectivamente de la variable Y, Los tres mecanismos se describen a continuación: a) Missing at Random (MAR), en este la distribución de los valores observados no depende del patrón de comportamiento de los registros sin información $Ymis : P(Z/Ycom) = P(Z/Yobs)$, por ello los datos perdidos pueden ser obviados; b) Missing completely at random (MCAR), el cual ocurre cuando la omisión no depende de los datos observados: $P(R/Ycom) = P(Z)$, este corresponde con un caso espacial MAR, en él los datos perdidos siguen una distribución aleatoria, pero los mismos no pueden ser obviados; y c) Cuando existe dependencia entre los datos completos y los faltantes cuando este ocurre se le denomina proceso no aleatorio (Missing not at Random, MNAR), la perdida de datos no es aleatoria ni ignorable). La hipótesis nula en el test de Little es que los datos faltantes siguen un MCAR, se utilizó un nivel de significancia del 5 %, para cada uno de los periodos seleccionados, Enders (2010); Daniels y col. (2008). Sin embargo, esta pregunta abarca únicamente un contexto general asumiendo que los datos de trabajo son multivariantes, caso que no es correcto en los datos de precipitación mensual, ya que esta es univariante con múltiples localizaciones espaciales.

2.2 Pregunta referida al tiempo: «Cuándo»

Se definió una pregunta a nivel general referida a Cuándo se generaron los cambios existenciales en la dimensión temporal correspondiente a la fecha de instalación y desinstalación del conjunto de estaciones climáticas?, esto con el fin de identificar el momento del tiempo en donde se encuentra la mayoría de las observaciones de precipitación. Para ello, en primer lugar se generaron los mapas de datos faltantes para cada una de las estaciones separadas por estado, se utilizó R como lenguaje y entorno de programación para análisis estadístico y gráfico R Project (2016) y la librería Amelia II diseñada para el análisis de valores perdidos Honaker y col. (2011). No se generó el mapas de datos faltantes en conjunto para toda la zona, debido a la gran cantidad de observaciones por estaciones, lo cual impedía el análisis visual de los datos (el periodo de análisis se llevó desde enero de 1900 hasta

diciembre 2014). Partiendo de esta pregunta general temporal se identificó el periodo de tiempo en donde se encuentran el gran volumen de los datos observados, se realizó una primera selección de estaciones, siendo el período seleccionado entre enero de 1949 a diciembre de 2008.

Medina y col. (2010) mencionan que un aspecto crucial en el análisis de datos se vincula al porcentaje máximo de omisiones que deben aceptarse. No existen criterios objetivos para dilucidar este tema, por lo que cada investigador debe hacerse cargo de sus propias decisiones. Rubin (1988) menciona que algunos métodos de imputación de datos faltantes generan buenos resultados, aún con porcentajes de omisión del 30, 40 o 50 %. No obstante, es preciso señalar que cuando se trabaja con este tipo de datos, el tamaño de muestra garantiza cierta precisión para una tasa máxima de no respuesta, y en la medida de que la omisión supera el umbral establecido se pone en riesgo la confiabilidad estadística de las variables principales. Por tanto, no se recomienda imputar datos en situaciones en que la omisión en una o más variables alcance porcentajes superiores al 20 %. Si se trabaja, por ejemplo, con una base de datos en donde la tasa de omisión en las variablesde interés se ubica en 25 %, se debe tener presente que modelar la respuesta en una de cada cuatro observaciones puede resultar adecuado en el ámbito académico, pero se considera poco útil desde el punto de vista práctico, sobre todo cuando los resultados se utilizarán para apoyar el diseño o evaluación de políticas públicas.

2.3 Pregunta referida al espacio: «Dónde»

Se plantearon preguntas referidas al Dónde, es decir, se analizó desde un contexto general las propiedades de primer orden, la escala y distribución espacial de las estaciones. Para ello se utilizó la función K de Ripley que incorpora la escala dentro del proceso de análisis. Para caracterizar las propiedades de primer orden se definieron las siguientes hipótesis:

- H_0: El proceso puntual espacial subyacente en las estaciones de precipitación mensual es de tipo Completamente Aleatoria (CSR).
- H_1: El proceso puntual espacial subyacente en las estaciones de precipitación mensual es distinto del tipo Completamente Aleatoria (CSR).

De rechazar la hipótesis nula se debe evaluar qué tipo de derivación del CSR se está manifestando (agregados o regulares), bajo el supuesto de que las estaciones no poseen datos discontinuos en el espacio tiempo. Para comprobar estas hipótesis se procedió a ejecutar la función K de Ripley a través de su parámetro, $L(d) - d$, estimándose de la siguiente manera:

$$L(d) - d = \sqrt{\frac{K(d)}{\pi}} - d \tag{2}$$

Donde: $K(d)$ equivale a la función K de Ripley, siendo su estimador, $\widehat{K}(d) = \frac{|A|}{n^2} \sum_{i=1}^{n} \#(C(x_i, d))$, resultando esta igual a la distancia, n es el número de muestras, A representa el área total de las muestras, d corresponde con el comportamiento en caso de una distribución espacial Completamente Aleatoria (CSR), valores por encima de 0 indican un alto grado de agrupación (Clustering), valores menores que 0 indican dispersión y los valores cercanos a 0 indican aleatoriedad espacial (CSR) Lloyd (2010), Baddeley (2008), Baddeley y col. (2005).

Para evaluar el factor de distribución espacial (Dónde) de los datos faltantes y la incertidumbre de una posible estimación de los datos faltantes, se procedió a realizar un Kriging Indicador, en donde el umbral de probabilidad es para aquellos valores que superen el 30 % (el modelaje se realizó con todos los porcentajes de los valores faltantes para cada estación), este tipo de interpolador se obtiene por el indicador de la transformación de una variable continua, o reducción de una variable categórica, a una variable binaria del cual se obtiene un estimador de la probabilidad de que un área supere o no un umbral dado, en este caso el 30 % de datos faltantes. Este tipo de Kriging es del tipo no lineal y no necesita evaluación de la distribución de frecuencia de los datos Bivand y col. (2008), Viera (2002), Li y col. (2008), se basa en la teoría y el desarrollo de estimadores no paramétricos de distribuciones espaciales de la manera:

$$i(x, z_c) = \begin{cases} 1, \; si \; Z(x) \leq z_c \\ 0, \; si \; Z(x) > z_c \end{cases} \tag{3}$$

Donde: $Z(x)$ es el valor observado en el punto x y z_c es el valor umbral definido por el usuario (30 %) Viera (2002). Para que sea efectiva la metodología Kriging debe realizarse lo que Matheron (1970) denomina el Análisis Estructural de la Dependencia Espacial, Viera (2002) define este como uno de los tópicos más importantes de la Geoestadística, puesto que se encarga de la caracterización de la estructura espacial de una propiedad o fenómeno regionalizado. Es el proceso en el marco del cual se obtiene un modelo geoestadístico para la función aleatoria que se estudia. En pocas palabras se puede decir que el análisis estructural consiste en estimar y modelar una función que refleje la correlación espacial de la variable regionalizada, a partir de la adopción razonada de la hipótesis más adecuada acerca de su variabilidad. Esto quiere decir, que en dependencia de las características de estacionaridad del fenómeno se modelará la función de covarianzas o la de semivarianzas. El Análisis Estructural de la Dependencia Espacial, el cual lleva dos etapas Gallardo (2006):

1) Estimación del semivariograma empírico: a través de la función de la semivarianza, esta es una medida de la autocorrelación espacial de una variable x entre dos puntos i, j. Puesto que puede calcularse la distancia entre dichos puntos, pueden representarse los valores de frente a las distancias h. El cálculo de la varianza

entre pares de individuos separados por intervalos de distancia se conoce como semivarianza(γ), estimada como: $\gamma_{(h)} = 1/2N_{(h)} \sum[Z_{(x)} - Z_{(x+h)}]^2$, Donde: $\gamma_{(h)}$ es la semivarianza para todas las muestras localizadas en el espacio, separado por el intervalo de distancia h; $N_{(h)}$, es el número total de pares de muestras separados por un intervalo de distancia h ; $Z_{(x)}$ es el valor de la muestra en una localización x; y $Z_{(x+h)}$ es el valor de la muestra a la distancia de intervalo h desde x. Generando un gráfico con las siguientes partes fundamentales: la curva que los puntos del variograma experimental da lugar a la definición de unos elementos básicos que lo caracterizan: a) *Rango*: este representa la máxima distancia hasta la cual existe dependencia espacial. Es el valor en el que se alcanza la máxima varianza, o a partir del cual ya presenta una tendencia asintótica; b) *Sill*: representa la máxima variabilidad en ausencia de dependencia espacial, la máxima semivarianza encontrada entre pares de puntos y debe coincidir con la varianza de la población. Da el grado de variación espacial, y por tanto el grado de incertidumbre a la hora de interpolar puntos en el espacio. Un alto cociente indica una variable espacialmente muy predecible. c) *Nugget*: conforme la distancia tiende a cero, el valor de la semivarianza tiende a este valor. Representa una variabilidad que no puede explicarse mediante la estructura espacial. El valor de la función ha de ser, lógicamente, cero en el origen. Representa la varianza no explicada por el modelo, y se calcula como la intercepción con el eje Y. Se conoce también como varianza error, puesto que la varianza de dos puntos separados por 0 metros (la intercepción con el eje Y) debería ser cero. Es por ello que esta varianza está normalmente indicando variabilidad a una escala inferior a la muestreada (figura 3).

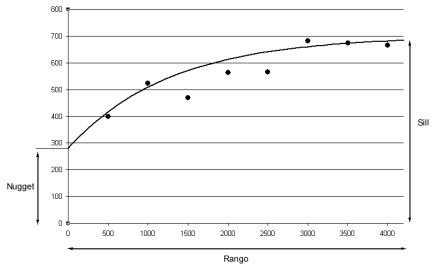


Fig. 3: Resumen de la nube del variograma en un variograma experimental con sus elementos definitorios Fuente: Olaya (2012).

2) Ajuste del semivariograma experimental: se utilizaron los criterios planteados por Olaya (2012) para ajuste del semivariograma experimental; una función apta para este propósito debe cumplir los siguientes requisitos: a) Ser monótona creciente; b) Tener un máximo constante o asintótico, es decir, un valor definido del sill y c) $(0) > 0$, el nugget debe ser positivo. Se utilizó un modelo de Mínimos Cuadrados Ordinarios (OLS) para el ajuste del semivariograma utlizando la librería gstat de R Bivand y col. (2008).

Ya con las selecciones realizadas para el análisis espacial de los valores faltantes, se utilizó la metodología sugerida por Lu y col. (2003) y Shekhar y col. (2003) la cual consiste en la detección de *Outliers*, para ello se sugiere tanto el uso de técnicas no espaciales como espaciales (métodos de estadística espacial sean aproximaciones Lattice o Geoestadística). Barnett y col. (1994) definen formalmente los *Outliers* (Valores Extremos) como observaciones dentro de un conjunto de datos que no parecen guardar consistencia con el resto del conjunto de los datos. Dentro de un contexto espacial Breunig y col. (2011) modifican el concepto original e introduce la noción de *Outlier* local, con el se plantea que el grado para que un objeto espacial sea catalogado como *Outlier* local debe está definido por la estructura de sus vecinos espaciales y su estructura de aglomeramiento espacial o «Clustering».

Los Lattice es una de las aproximaciones sugeridas por Shekhar y col. (2003) para la detección de *Outlier* locales, éste autor sugiere la aplicación de técnicas multidimensionales, como lo es el índice definido por Moran (1950), ampliamente descrito por diversos autores como Cressie (1993), Lloyd (2010), Olaya (2012), Toral (2001) y que introduce una variante del coeficiente de correlación lineal de Pearson, denominada I de Moran y está definido como:

$$I = \frac{N}{S_o} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{W_{ij}(X_i - \mu)(X_i - \mu)}{\sum_{i=1}^{N}(X_i - \mu)^2} \tag{4}$$

Donde: μ es la media de la variable x, $w_i j = d_{ij}^{-k}$ es la matriz de ponderación espacial, tal que $d_{i,j}$, corresponde a la distancia ponderada entre las muestras (Euclidiana o Manhattan) elevado inversamente proporcional a la potencia k (existen otros métodos para estimar la matriz de ponderación espacial, los cuales están descritos por Getis y col. (2004), S_0 es un factor de normalización igual a la suma de todos los elementos de la matriz de ponderación espacial $S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$, y N es el número de muestras tomadas. Es importante aclarar que la metodología sugerida por Shekhar y col. (2003) no corresponde con un test de hipótesis como tradicionalmente se maneja, sino se aplica como una técnica gráfica que utilizando el diagrama de Moran para identificar los valores *Outliers*.

La detección espacial de los *Outliers* es un proceso clave ya que la tasa de no respuesta y su distribución espacial puede ser variable Medina y col. (2010). Por ello, para la detección espacial de los *Outliers* se aplicó una aproximación modificada de la metodología «FindOut» sugerida por Yu y col. (2002), esta consiste en la remoción de los Clúster para la

identificación de los valores *Outliers* de los datos originales. Sin embargo, para el caso de estudio se desean conservar los valores extremos de mínimos de datos faltantes, por ello se tomó el concepto sugerido por Yu y col. (2002) y se aplicó el indicador local LISA, que corresponden con variantes de las medidas globales y tienen como objeto caracterizar la AE y AET para cada uno de los datos, en ellas se definen «Clúster» o conglomerados de datos, en los que los valores que poseen AE y AET positiva o negativa. Entre las medidas locales se destaca el indicador local de asociación espacial (LISA por sus siglas en inglés), desarrollado por Anselin (1995) y descrito en Anselin (2005), que estima los valores de la siguiente manera:

$$I_i = \frac{x_i - \mu}{S_i^2} w_{ij}(x_j - \mu); S_i^2 = \frac{\sum_{j=1;j\neq 1}^{n} w_{ij}(x_j - \mu)^2}{n-1} - \mu$$

Donde: μ es la media de la variable x, $w_i j = d_{ij}^{-k}$ es la matriz de ponderación espacial, tal que d_{ij} corresponde a la distancia ponderada entre las muestras (Euclidiana o Manhattan) elevado inversamente proporcional a la potencia k.

Siendo el problema analizado uno de característica espacial que pretende seleccionar aquellas estaciones con bajas cantidades de datos faltantes, en cuanto a la autocorrelación espacial positiva, no se seleccionaron aquellas estaciones con la tipología HH, es decir, valores altos de datos faltantes rodeados de valores altos y se conservaron aquellas con la condición, es decir, los valores LL, valores bajos de valores faltantes rodeados de valores bajos. El caso de la autocorrelación espacial negativa se conservaron las estaciones LH (valores bajos de faltantes rodeados de valores altos) y no se seleccionaron aquellas con valores HL (valores altos de faltantes rodeados de valores bajos), en caso de poseer autocorrelación espacial nula sería indiferente el criterio espacial de selección o no de las estaciones. Hay que destacar que tanto el I de Moran como el indicador LISA suponen que los datos se distribuyen de forma normal, es decir $X \sim N(X\beta, \sigma^2 I)$, Bivand y col. (2008). Sin embargo, no se procedió a transformar los datos, ya que no se está caracterizando la Autocorrelación Espacial, se están es identificando *Outliers* espaciales, por lo que una transformación puede enmascarar y/ó diluir los mismos dentro de la distribución de frecuencia.

2.4 Pregunta referida al objeto: «Qué» Individual

Adicionalmente, se requirió realizar una prueba que involucrara la caracterización individual de cada una de las estaciones en cuanto a su porcentaje de datos faltantes y evaluar la presencia de los denominados «Block Missing», es decir, la presencia de datos faltantes consecutivos en el tiempo Kondrashov y col. (2006), Lou y col. (2011) esto con el fin de seleccionar individualmente las estaciones que son más adecuadas para la estimación de los parámetros espaciales y temporales. Para ello, se diseñó una prueba en la que se evaluó de forma individual cada estación para el mejor periodo seleccionado; esta prueba se describe a continuación: Para cada serie se realiza una partición anual en segmentos de 12 unidades de medida (meses), se planteó la hipotesis nula de que los datos se distribuyen según una distribución de Poisson (esta es adecuada para modelar situaciones en una variable discreta en la que interesa determinar el número de hechos de cierto tipo que se pueden producir en un intervalo de tiempo o de espacio, bajo el supuestos de aleatoriedad Universitat de Valencia (2014)). Cada una de estas variables aleatorias representa el número total de ocurrencias de un fenómeno durante un periodo de tiempo fijo o en una región fija del espacio. Expresa la probabilidad de un número k de ocurrencias acaecidas en un tiempo fijo, si estos eventos ocurren con una frecuencia media conocida y son independientes del tiempo discurrido desde la última ocurrencia o suceso Benlloch y col. (2008). Esta distribución se define de la siguiente manera:

$$P(k, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{5}$$

Donde: k es el número de ocurrencias del evento o fenómeno (en este caso el número de datos ausentes por segmento), λ es un parámetro positivo que representa el número de veces que se espera que ocurra el fenómeno, la presencia de ausentes en un año, se estimó para cada serie de la siguiente manera:

$$\widehat{\lambda} = \frac{Num\ Au}{Num\ de\ A} \tag{6}$$

Donde: $Num\ Au$ es el Número de Datos Ausentes para cada serie y $Num\ de\ A$ corresponde con el Número de Años Evaluados, es decir, la proporción de datos ausentes por año (en este caso se colocó como condición de rechazo todos aquellos $\lambda > 2$, es decir, que la densidad anual de datos ausentes es menor a dos por año) y e es la base de los logaritmos naturales ($e = 2,71828$) Martínez (2010). En este caso, la densidad de eventos por segmento es equivalente a la proporción de total de datos faltantes multiplicado por la longitud del segmento, o sea 12. Para cada segmento se calcula la probabilidad bajo la hipotesis nula de aleatoriedad planteada anteriormente, dado el número de eventos o valores faltantes en dicho segmento.

Como se asume que lo ocurrido en un segmento es independiente a lo ocurrido en otros, se multiplican las probabilidades de todos los segmentos de una misma serie. La probabilidad conjunta de ocurrencia de una realización particular bajo la hipotesis nula de distribución de Poisson depende mucho del valor estimado como parámetro, penalizando con mayor frecuencia las menores desviaciones a la hipotesis nula cuando el parámetro se acerca a 6. Por esta razón, fue necesario hallar una probabilidad límite en base a simulaciones realizadas con el parámetro estimado

para cada serie, y no seleccionar a priori una probabilidad límite para todas las series. Como criterio de selección, se simularon 1000 series de tiempo discretas con distribucion de Poisson con el mismo parámetro estimado en la serie original. Se realiza el mismo procedimiento para hallar la probabilidad conjunta en cada simulación, se ordena de menor a mayor probabilidad, se selecciona el primer decil (que se consideraría como un intervalo e confianza unilateral del 10 %), y se compara la probabilidad conjunta de la serie simulada seleccionada con la original, si la probabilidad de la serie simulada es mayor que la probabilidad de la serie original, se rechaza la serie original (se rechaza la hipótesis de aleatoriedad de la distribución de los datos ausentes), en el caso contrario se acepta.

3 Discusión y Resultados

Los mapas de datos faltantes confirman que desde un punto de vista existencial que la mayoría de las estaciones de precipitación mensual poseen datos entre el periodo enero de 1949 a diciembre de 2008 (valores ausentes se presentan en color claro y valores presentes en color oscuro), cabe destacar que en la figura 4 se presentan los mapas de datos faltantes para el año 1900-2014, para el estado Lara y los once estados de trabajo, respectivamente.

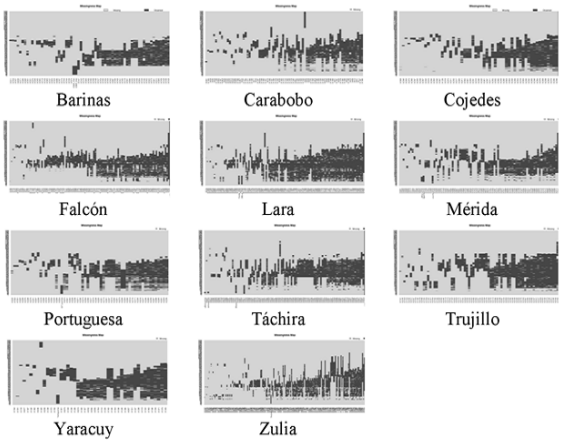


Fig. 4: Mapas de datos faltantes periodo 1900-2014.

Los semivariogramas experimentales ajustados corresponden al modelo Esférico, Modelo: Rango s y silla, Comportamiento lineal en el origen, Pendiente igual a 1,5 s/a de la forma:

$$\gamma_{(h)} = \begin{cases} s(\frac{3}{2}\frac{|h|}{a} - \frac{1}{2}\frac{|h^8|}{a^8}) \\ s,\ si|h| > a \end{cases} \quad (7)$$

Este representa fenómenos continuos, pero no diferenciables López (2005), todos con un Nugget elevado, y con un rango ligeramente superior a 100.000 m. El Nugget como se mencionó, es indicador de que el comportamiento o variabilidad que no puede explicarse mediante la estructura espacial Gallardo (2006) (figura 5). Es importante destacar que con la presencia de autocorrelación espacial, se demuestra que el mecanismo de los datos perdidos no responde a un Mecanismo Completamente Aleatorio de Valores Faltantes, desde el punto de vista espacial, ya que según Qu y col. (2009) los valores perdidos tienen un cierto patrón, en este caso espacial.

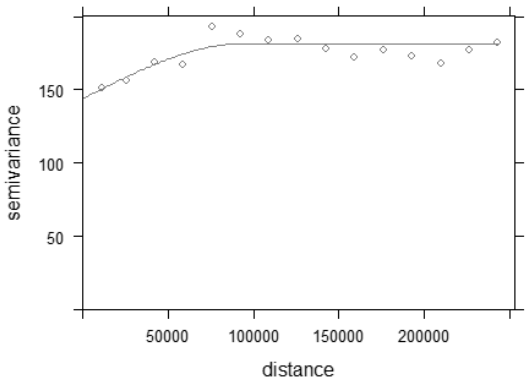


Fig. 5: Semivariograma empírico y modelo ajustado Período 1949-2000.

Al analizar los resultados del Kriging Indicador, estos muestran que los mejores datos son los de los estados Trujillo y Lara para todos los periodos de trabajo; sin embargo, preocupa sustancialmente la situación del norte de los estados de Venezuela, que son los que poseen la cantidad mayor de datos a estimar. Si el usuario decide realizar dicha estimación sobre estas estaciones correrá el riesgo a que el producto pueda tener altas varianzas desviadas de la realidad espacio temporal de los datos. Es importante resaltar que el Kriging indicador tiende a mejorar al recortar el periodo a 1949-2000, siendo este el mejor desde el punto de vista espacial (figura 6).

Bajo los criterios del máximo de datos faltantes tolerados para el período 1949-2008, 135 estaciones cumplen el criterio de correspondiente al 30 %, 86 el 25 % y 60 el 20 %; para el periodo 1949-2005, 151 estaciones cumplen

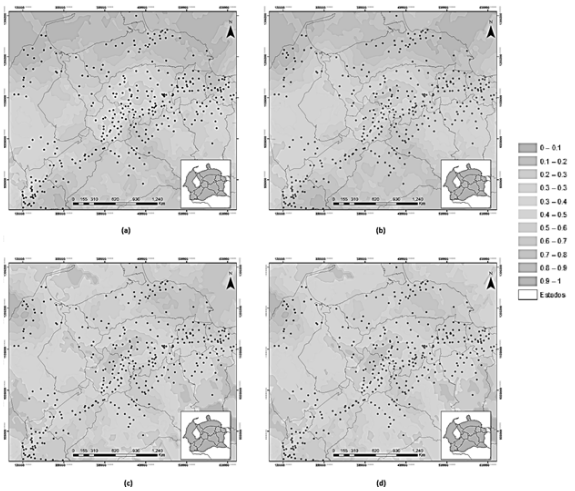


Fig. 6: Kriging Indicador umbral superior al 30 % de datos faltantes (a) Período 1949-2008; (b) Período 1949-2005; (c) Período 1949-2000 y (d) Período 1949-1998.

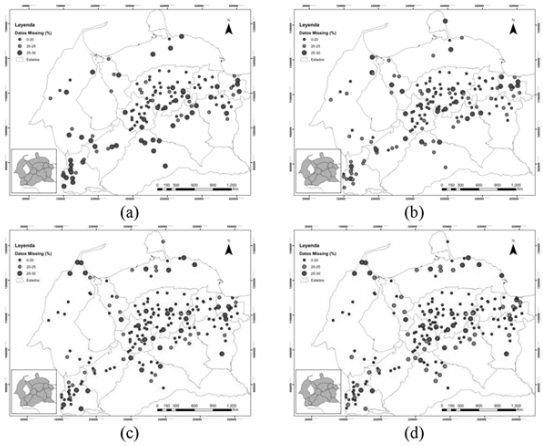


Fig. 7: Datos Faltantes por estación (a) Período 1949-2008; (b) Período 1949-2005; (c) Período 1949-2000 y (d) Período 1949-1998.

el criterio de correspondiente al 30 %, 122 al 25 % y 71 al 20 %; para el periodo 1949-2000, 182 estaciones cumplen el criterio de correspondiente al 30 %, 150 el 25 % y 118 el 20 %; y finalmente para el periodo 1949-1998, 187 estaciones cumplen con el máximo de 30 %, 159 el criterio del 25 % y 124 el 20 % de datos faltantes (figura 7). Al observar el porcentaje de datos a imputar por periodo se observa que no es resaltante el porcentaje total de datos faltantes, dado que este es relativamente bajo, incluso para el periodo 1949-2008, sin embargo, estados con gran variabilidad de la precipitación mensual causada por el relieve, como lo son: Barinas, Portuguesa, Mérida y Táchira, poseen la mayor cantidad de datos a imputar y los estados Lara y Trujillo de las menores cantidad de datos a imputar (tabla 2).

El histograma de frecuencia de los datos faltantes muestra un fuerte comportamiento unimodal asimétrico a la izquierda, para el caso de las selecciones 1949-2008 y 1949-2005, indicando que el número de datos faltantes supera el 25 % y el 20 % respectivamente. Para los casos de las selecciones 1949-2000 y 1949-1998, la distribución de frecuencia se transforma en una distribución en campana, esto indica que estas últimas selecciones, desde este punto de vista, son las de mejor calidad ya que la mayoría de las estaciones están por debajo del 20 % de datos faltantes. Este análisis es refrendado con los diagramas de Moran que indican que es más complicado identificar *Outliers* en las selecciones 1949-2008 y 1949-2005, ya que en estos casos la mayoría de los valores están comprimidos cercanos a la recta, caso que se percibe en menor magnitud en las selecciones 1949-2000 y 1949-1998, por lo que es evidente que estas últimas son las estaciones periodos más idóneos (figura 8).

Al analizar los mapas LISA es bastante evidente que los estados Lara y Trujillo poseen los mejores datos, ya que en ellos dominan los valores Clúster LL, y los estados Zulia, Falcón, Barinas y Mérida son los que espacialmente poseen las peores distribuciones; preocupa el estado Mérida, ya que los datos de precipitación mensual poseen una fuerte variabilidad, debido al control que genera el relieve (figura 9). El número de Clúster LISA obtenidos para cada selección indica que la posición espacial es una condición no relevante con respecto a los datos faltantes; sin embargo, estas como se explicó anteriormente, los estados Lara y Trujillo poseen baja cantidad de datos faltantes, siendo los LL los valores los más numerosos, resultando estos para los cuatro períodos de trabajo los estados con mejor calidad de datos (tabla 3).

El análisis de la función K de Ripley a través de su parámetro, $L(d) - d$, indica la presencia de Clúster espacial para todos los períodos, sin embargo, la escala cartográfica de los datos indica que el período 1949-2008 tiene una escala aproximada de 1:118.800, el periodo 1949-2005 1:120.010 y los períodos 1949-2000 y 1949-1998 una escala 1:103.560. Esto se determinó de acuerdo, al identificar el punto máximo del parámetro $L(d) - d$ el cual es indicativo de la escala cartográfica. Sin embargo, esta estimación se realizó sin incluir el efecto de borde, y el área de estudio ocupa parte de la extensión marítima y extensión de Colombia, por lo que al incluir este factor del efecto de borde del área de estudio la escala, probablemente mejore.

Los mapas de datos faltantes de cada una de las selecciones indican que el proceso es tendiente a Mecanismo Completamente Aleatorio de Valores Faltantes en el periodo 1960-1990, sin embargo, el proceso de los datos perdidos se vuelve un claro Proceso de datos No Perdidos de Forma Aleatoria Enders (2010) y Daniels y col. (2008), en los extremos temporales (superior al año 1990 e inferior al año 1960), esto se manifiesta cuando se encuentra lo que se conoce con un «Block Missing» Kondrashov y col. (2006), Lou y col.

Tabla 2: Número de estaciones de precipitación mensual que cumplen el criterio de máximo 30 % del número de observaciones perdidas.

Estado	Periodo1949-2008			**Periodo1949-2005**			**Periodo 1949-2000**			**Periodo 1949-1998**		
	N Est. [1]	**N Dat.**[2]	**Por Dat.** [3]	**N Est.** [1]**.**	**N Dat.**[2]	**Por Dat.** [3]**.**	**N Est.** [1]**.**	**N Dat.**[2]	**Por Dat.** [3]	**N Est.** [1]**.**	**N Dat.**[2]	**Por Dat.** [3]
Barinas	6	1218	28.7	6	1002	24.9	10	1277	20.9	11	1107	17.1
Carabobo	8	1399	24.7	11	1788	24.2	12	1780	24.2	13	1317	17.2
Cojedes	10	1492	21.1	11	1360	18.4	13	1287	16.2	13	1121	14.7
Falcón	4	642	22.7	5	712	21.2	11	1487	22.1	13	1731	22.6
Lara	38	4539	16.9	41	4394	15.9	38	1080	4.6	38	2893	12.9
Mérida	5	922	26.0	7	1148	24.4	9	1079	19.6	9	881	16.6
Portuguesa	8	1418	25.0	8	1182	22.0	11	1418	21.1	11	1307	20.2
Táchira	14	2656	26.8	15	2365	23.5	22	2718	20.2	22	2246	17.4
Trujillo	20	2249	15.9	21	1833	13.0	23	1577	11.2	24	1618	11.5
Yaracuy	8	777	13.7	8	559	10.4	10	730	11.9	10	654	11.1
Zulia	14	2364	23.8	18	2678	22.1	23	2708	19.2	23	2548	18.8
Total	**135**	**19676**	**20.6**	**151**	**19021**	**18.7**	**182**	**17141**	**15.4**	**187**	**17423**	**15.8**

N Est.[1] (Número de Estaciones); N Dat.[2] (Número de datos a imputar); Por Dat.[3] (Porcentaje de Datos a Imputar).

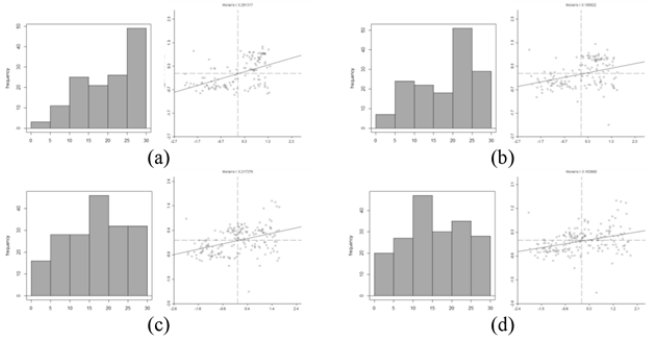


Fig. 8: Histograma y Diagrama de Moran del porcentaje de datos faltantes (a) Período 1949-2008; (b) Período 1949-2005; (c) Período 1949-2000 y (d) Período 1949-1998.

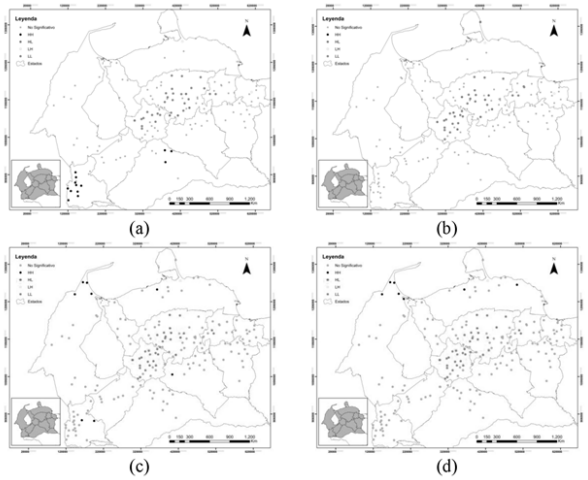


Fig. 9: Clúster LISA del porcentaje de datos faltantes (a) Período 1949-2008; (b) Período 1949-2005; (c) Período 1949-2000 y (d) Período 1949-1998.

(2011) que se produce cuando observaciones consecutivas poseen datos faltantes (figura 10). Los resultados del test de Little rechazan la hipótesis nula de que el proceso de los datos faltantes es un Mecanismo Completamente Aleatorio de Valores Faltantes desde un punto de vista temporal (tabla 4). Esto puede ser causado sustancialmente por la presencia del Block Missing en los extremos temporales de las observaciones, dejando el proceso estocástico como un Mecanismo No aleatorio de Datos Faltantes, pues los valores perdidos tienen un cierto patrón Qu y col. (2009). El producto del análisis Poisson indica que 59 estaciones aprobaron el análisis (figura 11 (a)) estando primordialmente distribuidas en la zona de Trujillo y Lara. La estación que posee el máximo valor de datos faltantes se encuentra en el estado Táchira y posee el valor de máximo de casi 25 % de datos faltantes.

Al generar el mapa de datos faltantes de las estaciones que finalmente aprobaron se encuentra una aproximación más real a un Mecanismo Aleatorio de Datos Faltantes, por lo que la imputación que se genere tendrá menor sesgo. Sin embargo, desde el punto de vista espacial, al generar la función K de Ripley a través del parámetro $L(d) - d$, se obtiene una distribución nuevamente de tipo Clúster con un pico en la escala 1:75.000, por lo que la medida de la intensidad de la correlación espacial utilizada para estos datos será probablemente inadecuada, ya que estos resultados son indicativos de la presencia de zonas mejor representadas que otras.

En efecto, según Olaya (2012) tiene sentido evaluar la función de distribución en valores de distancia pequeña en comparación con el tamaño de la zona de estudio, ya que para otros valores no resulta coherente analizar los efectos de segundo orden (Dependencia Espacial) dentro de dicha zona. Por ello, se procedió a seleccionar los datos de los estados Trujillo, Lara, Yaracuy, Cojedes y Carabobo (Proyecto Piloto) y se reestimó K de Ripley a través del parámetro $L(d) - d$ y el mapa de datos faltantes únicamente, para las estaciones de estos estados. Al realizar el mapa de datos faltantes se observa un mecanismo muy aproximado a un Mecanismo Aleatorio de Datos Faltantes y que aunque se mantiene la distribución Clúster de las estaciones, la distribución espacial mejora aproximándose a una distribución aleatoria, cercano a

Tabla 3: Numero de Clúster por periodo de selección.

Tipología de Clúster	Período1949-2008	Período 1949-2005	Período 1949-2000	Período 1949-1998
No Significativo	88	113	138	143
HH	13	0	8	7
LL	27	27	28	27
LH	2	2	1	2
HL	5	9	7	8

la escala de 1:60.000 (figura 11 (b)).

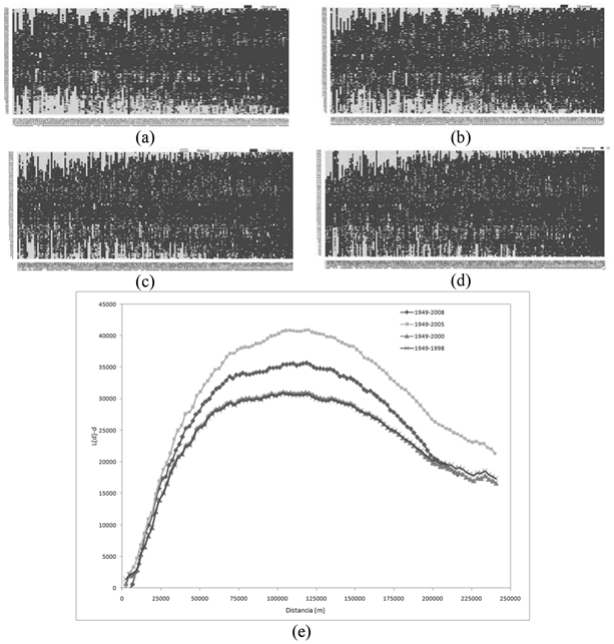


Fig. 10: Mapas de datos faltantes de cada una de las selecciones (a) Período 1949-2008; (b) Período 1949-2005; (c) 1949-2000; (d) Período 1949-1998 y (e) Parámetro $L(d) - d$ para los cuatro períodos analizados.

Tabla 4: Little Test.

Período	Chi-Cuadrado	Gl	p-value
1949-2008	74605	72874	0
1949-2005	81699	79701	0
1949-2000	93052	91193	0
1949-1998	92695	91018	0

Para establecer un aproximado de la representatividad de las estaciones del proyecto piloto se calculó la función K de Ripley a través del parámetro $L(d) - d$, utilizando como límite de estimación una capa modificada de una Geometría Convexa, ya que es importante considerar los efectos de borde para saber si los valores calculados dentro de cualquier análisis estadístico son válidos o no; es decir, debe tomarse en cuenta la función de densidad espacial, ya que depende de el número de muestras N y el área de estudio A, tal que: $\lambda = \frac{N}{A}$, por lo que al modificar el área de estudio y su dimensión, la función de densidad espacial cambia Olaya (2012). Se seleccionó un polígono en donde las densidades visualmente fuesen aproximadamente la misma, al estimar el parámetro $L(d)$, los resultados de esta indican la presencia en esa zona de un Mecanismo Aleatorio de Distribución Espacial (Proceso Poisson Homogéneo), por lo que en esta zona la medida de correlación es espacial y temporalmente balanceada (figura 13). Ya que se logró controlar el Mecanismo no Aleatorio de Datos Faltantes, en el contexto espacial las medidas derivadas de esta serán estadísticamente insegadas. Teniendo estos datos una escala aproximadamente constante de 1:70.000.

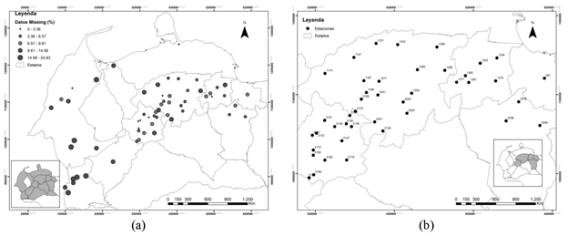


Fig. 11: (a) Estaciones de precipitación mensual del Proyecto que aprobaron la Prueba Poisson y (b) Estaciones de precipitación mensual del Proyecto Piloto que aprobaron la Prueba Poisson.

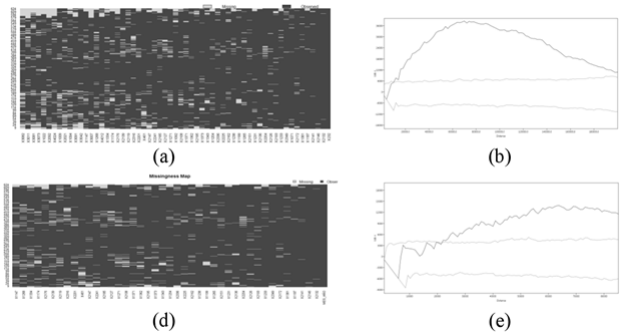
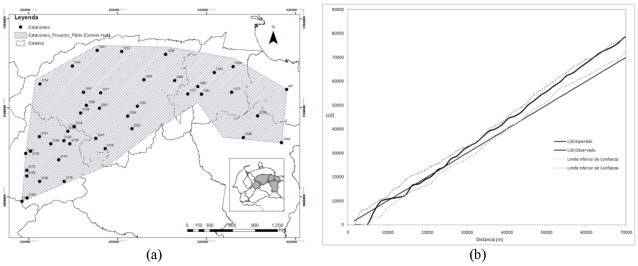


Fig. 12: (a) Mapa de datos faltantes de la estaciones que aprobaron el Test Poisson; (b) Parámetro $L(d) - d$ Estaciones que aprobaron el Test Poisson; (c) Mapa de datos faltantes de la estaciones que aprobaron del Proyecto Piloto y (d) Parámetro $L(d) - d$ Estaciones Proyecto Piloto.



(a) (b)

Fig. 13: (a) Geometría Convexa Modificada del Proyecto Piloto y (b) Parámetro $L(d)$ Estaciones del Proyecto Piloto con limitación del área de estudio a la Geometría Convexa Modificada.

4 Conclusiones

Los datos comprendidos en cualquier periodo de tiempo, para el conjunto de las 961 estaciones, poseen un claro Mecanismo No Aleatorio de Datos Faltantes, estando caracterizado por la presencia de autocorrelación espacial en la ocurrencia de los datos faltantes y la aparición de bloques de datos faltantes consecutivos en el tiempo, que condicionan además los fenómenos de Dependencia Espacial.

Si el mecanismo de generación de datos faltantes fuera aleatorio, las estimaciones realizadas de las medidas de autocorrelacion temporal y espacial fueran insesgadas, a pesar de la pérdida de precisión, debido a la menor disponibilidad de datos. Pero, al ser en el caso del presente estudio, un mecanismo no aleatorio, sino influenciado por los fenómenos mencionados, no hay garantía de la insesgadez de dichas estimaciones, ya que puede producir estimaciones de parámetro distorsionadas, cuando el supuesto del mecanismo no aleatorio no se sostiene Cacabelos (2014). Es decir, para lograr una descomposición correcta de la estructura de los datos, en términos de la autocorrelación espacial y temporal, es imperativo obtener un subconjunto de los datos que cumplan con poseer un mecanismo completamente aleatorio en sus datos faltantes. Al estimar únicamente sobre este subconjunto de datos, se evita que el área mejor representada (con menos cantidad de datos faltantes) tenga un efecto demasiado influyente sobre las estimaciones. Al obtener una estimación a partir de solo un subconjunto de los datos, se debe asumir que los parámetros son constantes en el tiempo para dichas estaciones que hagan parte del subconjunto (estacionariedad temporal), que los parámetros son constantes en el espacio para el periodo de tiempo del subconjunto y aun más complejo, se debe asumir que los parámetros son constantes para las demás estaciones o puntos en el espacio, en un periodo de tiempo distinto al correspondiente al subconjunto.

La consecuencia de modelar ante la presencia de una distribución Aleatoria, Clúster o Regular en el espacio, la discute Hengl (2006) indicando que un elemento importante como lo es el tamaño del píxel, fundamental en el proceso de interpolación, va a depender de la tipología de la distribución espacial de las muestras. Por ejemplo, si la distribución fuese regular, el tamaño del píxel se estima de la siguiente manera: $\Delta s = 0{,}5\sqrt{\frac{A}{N}}$, donde: Δs es el tamaño del píxel A y N el número de muestras; sin embargo, si la distribución fuese agregada el tamaño del píxel es $\Delta s = 0{,}25\sqrt{\frac{A}{N}}$, es decir, que la distancia media entre puntos, es aproximadamente la mitad de la existente en un patrón regular con el mismo número de puntos, con la salvedad de que será válido, solo en las zonas donde se encuentren los conglomerados de muestras. Este mismo autor también recomienda el uso de la autocorrelación espacial como herramienta de estimación de la escala cartográfica si esta es desconocida. Los resultados descritos en estudios previos, como lo son los de Andrades y López (2015), Andrades-Grassi y cols. (2015), son útiles desde el punto de vista de exploración (objetivos planteados en los mismos); sin embargo, como en ellos se estimaron parámetros de Autocorrelación Espacio Temporal, y de Correlación Espacial en estructuras en distribución de tipo Clúster espacial y en presencia de datos faltantes, estos parámetros no son adecuados para el modelamiento.

Bajo las condiciones actuales y en vista de la disponibilidad de los datos y su distribución, se recomienda trabajar en el Proyecto Piloto la imputación de los datos, descomposición espacial y temporal de la autocorrelación, modelamiento y demás análisis a realizar para el periodo comprendido entre los años 1949-2000. Es importante que otras variables climáticas sean analizadas desde este punto de vista, sobre todo aquellas con importancia en el análisis de cambio climático y disponibilidad de oferta hídrica.

Referencias

Andrades J, López Hernández J, 2015, Caracterización de los procesos espaciales y temporales y sus interrelaciones en estaciones de precipitación mensual en la zona centro occidental de la República Bolivariana de Venezuela, GeoFocus. Revista Internacional de Ciencia y Tecnología de la Información Geográfica, 16 pp. 151-176.

Andrades-Grassi J, Cuesta L, López Hernández J, Goitía A, 2015, Evaluation of normality in the determination of spatio-temporal autocorrelation of monthly precipitation in the central-west region of Venezuela, 3rd Edition of the Integrated Management of Environmental Resources Conference Suceava.

Andrienko N, Andrienko G, Gatalsky P, 2003, Exploratory spatio-temporal visualization: an analytical review, Journal of Visual Languages & Computing, Nro. 6-14,pp. 503-541.

Anselin L, 1988, Spatial Econometrics: Methods and Models, Kluwer Academic.

Anselin L, 1995, Local indicators of spatial association - lisa, Geographical analysis, Nro. 27-2, pp. 93-115.

Anselin L, 1999, Spatial Econometrics, Bruton Center, School of Social Sciences, University of Texas, Dallas.

Anselin L, 2005, Exploring Spatial Data with GeoDa (tm): A Workbook, Center for Spatially Integrated Social Science.
Anselin L, Florax R, Rey S, 2013, Advances in spatial econometrics: methodology, tools and applications, Springer Science & Business Media.
Anselin L, Rey S, 2010, Perspectives on Spatial Data Analysis, Springer, pp. 1-20.
Baddeley A, 2008, Analysing spatial point patterns in R, Citeseer.
Baddeley A, Turner R, 2005, Spatstat: an R package for analyzing spatial point patterns, Journal of statistical software, Nro. 12-6 pp. 1-42.
Barnett V, Lewis, T, 1994, Outliers in Statistical Data, Probability & Mathematical Statistics, Wiley.
Benlloch M, Gómez D, y Martínez M, 2010, La distribución de Poisson.
Bivand RS, Pebesma EJ, Gómez-Rubio V, 2008, Applied spatial data analysis with R, New York: Springer Google Scholar.
Box G, Jenkins G, Reinsel G, Ljung G, 2015, Time series analysis: forecasting and control. John Wiley & Sons.
Breunig M, Kriegel HP, Ng R, Sander J, 1999, Optics-of: Identifying local outliers, European Conference on Principles of Data Mining and Knowledge Discovery, pp.262-270.
Cacabelos M, 2014, Imputación de datos faltantes en un modelo de tiempo de fallo acelerado. Universidad de Vigo.
Cressie N, 1993, Statistics for spatial data: Wiley series in probability and statistics. Wiley-Interscience, New York.
Cressie N, Wikle C, 2015, Statistics for spatio-temporal data. John Wiley & Sons.
Daniels MJ, Hogan JW, 2008, Missing data in longitudinal studies: strategies for Bayesian modeling and sensitivity analysis. Monographs on Statistics and Applied Probability.
Enders C, 2010, Applied missing data analysis. Guilford Press.
Hengl T, 2004, Finding the right pixel size, Computers & Geosciences, 32-9, pp. 1283-1298.
Gallardo A, 2006, Geostadística, Ecosistemas, 15-3, pp. 1-11.
Getis A, Aldstadt J, 2004, On the specification of the spatial weights matrix, Geographical Analysis, 36-2, pp. 90-104.
Gujarati D, Porter D, 2010, Econometría (Quinta edición). México: McGraw-Hill/Interamericana Editores, SS DE CV.
Hair J, Black W, Babin B, Anderson R, Tatham R, 2009, Multivariate data analysis. Bookman Editora.
Honaker J, King G, Blackwell M, 2011, Amelia II: A program for missing data models, Journal of statistical software, Nro. 45-7,pp. 1-47.
Hyndman R, Kostenko A, 2007, Minimum sample size requirements for seasonal forecasting models, Foresight, International Institute of Forecasters, Nro. 6,pp. 12-15.
Kamarianakis Y, Prastacos P, 2003, Spatial time series modeling: A review of the proposed methodologies, The Regional Economics Applications Laboratory.
Kamarianakis Y, Prastacos P, 2005,Space-time modeling of traffic flow, Computers & Geosciences, Nro. 31-2,pp. 119-133.
Kondrashov D, Ghil M, 2006,Spatio-temporal filling of missing points in geophysical data sets, Nonlinear Processes in Geophysics, Nro. 13-2,pp. 151-1159.
Li J, Heap A, 2008,A review of spatial interpolation methods for environmental scientists, Geoscience Australia Canberra.
Li X, Kraak M, 2008,A new method of visual exploration of geo-data in time-space, The Cartographic Journal, Taylor & Francis ,Nro. 45-3,pp. 193-200.
Lloyd C, 2010, Spatial data analysis an indtroduction for GIS users, Oxford University Press.
López F, Chasco C, 2005, Space-time lags: specification strategy in spatial regression models, Contributions in spatial econometrics, Nro. 5, pp. 125-149.
López F, Chasco C, 2007, Time-trend in spatial dependence: Specification strategy in the first-order spatial autoregressive model, Estudios de Economía Aplicada, Nro. 25, pp. 631-650.
Lou Q, Obradovic Z, 2011, Modeling multivariate spatio-temporal remote sensing data with large gaps, Data Mining, IJCAI Proceedings-International Joint Conference on Artificial Intelligence, Nro. 22-1.
López C, 2005,Variogramas.http://www.geo.upm.es/postgrado/CarlosLopez/geoestadistica/VARIOGRAMA.ppt Fecha de consulta: 26 Enero 2016.
Lu C, Chen D, Kou F, 2003, Algorithms for spatial outlier detection, Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, pp. 597-600.
Martínez, J, 2010,Contenedor hipermedia de estadística aplicada a las ciencias económicas y sociales (CEACES). https://www.uv.es/ceaces/base/modelos %20de %20probabilidad/poisson.htm Fecha de consulta: 20 Marzo 2015.
Matheron G, 1970, La teoría de las variables regionalizadas y sus aplicaciones, Los Cuadernos del Centro de Morfología Matemática de Fontainebleau.
Medina F, Galván M, 2007, Imputación de datos: teoría y práctica, United Nations Publications.
Medina RD, Montoya EC, Jaramillo, A, 2008, Estimación estadística de valores faltantes en series históricas de lluvia. http://www.cenicafe.org/es/publications/arc059(03)260-273.pdf. Fecha de consulta: 20 Enero 2015.
Moran P, 1950, Notes on continuous stochastic phenomena, Biometrika, Nro. 37-1/2,pp. 17-23.
Olaya, V, 2012,Sistemas de información geográfica. http://volaya.github.io/libro-sig/ Fecha de consulta: 20 Enero 2015.
Peuquet D, 1994, It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems, Annals of the Association of american Geographers, Wiley Online Library, Nro. 3-84,pp. 441-461.
Pfeifer P, Deutrch S, 1980, A three-stage iterative procedure for space-time modeling phillip, Technometrics, Taylor &

Francis, Nro. 1-22,pp. 35-47.
Podestà F, 2002, Recent developments in quantitative comparative methodology: The case of pooled time series cross-section analysis, DSS Papers Soc, Nro.3-2,pp. 5-44.
Qu L, Zhang Y, Hu J, 2009, PPCA-based missing data imputation for traffic flow volume: a systematical approach, IEEE Transactions on intelligent transportation systems, Nro.10-3,pp. 512-522.
R Project, 2016,The Comprehensive R Archive Network. https://cran.r-project.org/ Fecha de consulta: 26 Enero 2016.
Rubin T, 1988, An overview of multiple imputation, Proceedings of the survey research methods section of the American statistical association, pp. 79-84.
Schneider T, 2001, Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, Journal of Climate, Nro. 4-14,pp. 853-871.
Shekhar S, Lu C, Zhang P, 2003, A unified approach to detecting spatial outliers, GeoInformatica, Nro. 7-2,pp. 139-166.
Stadelmann-Steffen I, Bühlmann, M, 2008, Space and Time in Comparative Political Research: Pooled Time-series Cross-section Analysis and Multilevel Designs Compared, Methoden, Daten, Analysen, Nro. 1-2,pp. 29-57.
Toral A, 2001, El factor espacial en la convergencia de las regiones de la Unión Europea: 1980-1996, Universidad Pontificia Comillas de Madrid (España).
Universitat de Valencia, 2014, Distribución Poisson. https://www.uv.es/ceaces/base/modelos %20de %20probabilidad/poisson.htmf Fecha de consulta: 20 Marzo 2015.
Viera M, 2002, Geoestadística aplicada, Notas de curso, Instituto de Geofísica, Universidad Nacional Autónoma de México (UNAM), Instituto de Geofísica y Astronomía, CITMA. Cuba.
Yu D, Sheikholeslami G,Zhang A , 2002, Findout: finding outliers in very large datasets, Knowledge and Information Systems, Nro. 4-4,pp. 387-412.

***Recibido**: 02 de Julio de 2017*

***Aceptado**: 08 de Febrero de 2018*

Andrades Grassi, Jesús Enrique: *Ingeniero Forestal, Msc. Manejo de Cuencas, Profesor Asociado de la Universidad de Los Andes, PhD. Candidato en Ciencias Forestales y Ambientales*

López Hernández, Juan Ygnacio: *Ingeniero Forestal, Msc. Manejo de Bosques, Profesor Titular de la Universidad de Los Andes, PhD. Universidad Albert Ludwigs de Friburgo, Alemania. Departamento de Sensores Remotos y Sistemas de Información de Tierras. (FeLiS) Correo electronico: ferninfo@felis.uni-freiburg.de.; cubarro@gmail.com;jlopez.merida@gmail.com*

Goitía Acosta, Arnaldo: *Licenciado en Matemáticas, Msc. Estadística, Profesor Titular de la Universidad de Los Andes, PhD. en Matemáticas, Universidad de Granada, España. Correo electronico: arnaldogt0@gmail.com;goitia@ula.ve*

Torres Mantilla, Hugo Alexander: *Médico Cirujano, Msc. Estadística, Profesor Universidad de Santander, Colombia. Correo electronico: h.a.torresmantilla@hotmail.com*

Mejías Delgado, Jesús Enrique: *Ingeniero Agronomo, Msc. Recursos Hidraulicos, Profesor Titular de la Universidad de Los Andes, Correo electronico: jesusemejiasd@gmail.com*