ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELOS ESTADÍSTICOS PARA LA DETECCIÓN DE PATRONES EN MEDIO AMBIENTE Y ECONOMIA

APLICACIÓN DE TÉCNICAS BOOTSTRAP PARA ESTABLECER UMBRALES MÁS EXIGENTES AL MOMENTO DE REALIZAR LA DETECCIÓN DE INHOMOGENEIDADES EN SERIES METEOROLÓGICAS.

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PALLASCO CATOTA JONATHAN FERNANDO

jonathan.pallasco@epn.edu.ec

DIRECTOR: PH. D. MIGUEL ALFONSO FLORES SÁNCHEZ

miguel.flores@epn.edu.ec

QUITO D.M., FEBRERO 2022

CERTIFICACIONES

Yo, PALLASCO CATOTA JONATHAN FERNANDO, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

Pallasco Catota Jonathan Fernando

Certifico que el presente trabajo de integración curricular fue desarrollado por Pallasco Catota Jonathan Fernando, bajo mi supervisión.

PH. D. Miguel Alfonso Flores Sánchez
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

Pallasco Catota Jonathan Fernando

PH. D. Miguel Alfonso Flores Sánchez

AGRADECIMIENTOS

A mis padres por ser el apoyo incondicional y por estar pendientes de mí durante mi etapa como estudiante, pues sin sus consejos, regaños y preocupaciones no podría lograr lo que me propuse. Los aprecio demasiado, ya que son la fuente primordial para no rendirme y seguir creciendo como persona para lograr cualquier objetivo que me proponga.

A Britney quién es mi compañera de vida, la que me acompañaba en mis veladas, la que siempre me preguntaba como me fue en una prueba y la que día a día me motiva a ser un ejemplo para mis hermanos y sobrinos.

A mis hermanos y sobrinos por su cariño, apoyo y sobre todo por verme como ejemplo en la familia.

A cada uno de los profesores que han compartido sus conocimientos y experiencias; en especial, agradezco a mi tutor Miguel Flores, por confiar en mi y guíarme profesionalmente, ya que con su apoyo y tiempo, he logrado desarrollar el presente trabajo de integración curricular.

A mis compañeros, que me permitieron compartir momentos inolvidables en mi etapa universitaria.

DEDICATORIA

A mis queridos y valorados padres, Piedad y Kleber, por su esfuerzo y malas noches que tenían que pasar para verme prosperar en la vida. Las personas más motivadoras y guerreras que he conocido.

A mis hermanos: Kristela, Moises, Dayana y Anahi, por su pasciencia y compañia.

A mis sobrinos: Cielo, Erick y Kleber, por el aprecio y respeto.

A mi novia Britney, por sus sabios consejos.

A mis cuñados, por motivarme día a día.

RESUMEN

Las inhomogeneidades en las series de tiempo puede desatar enventos extremos los cuales provocan un impacto negativo en el campo de estudio. En la meteorología estas perturbaciones pueden representar eventos meteorológicos como: olas de calor, sequías, inundaciones, vientos fuertes e incluso incendios forestales y se originan debido a errores en el registro de la información o al cambio gradual de las estaciones. Estas inhomogeneidades deben ser detectadas y corregidas del conjunto de datos mediante el proceso de homogenización. Con el paquete R *Climatol* se puede realizar todo este proceso mediante código R, pero se presenta un problema al momento de elegir el umbral de corrección de inhomogeneidades. La elección del umbral se realiza de forma subjetiva, solamente observando los gráficos de los histogramas máximos SNHT, pero no es lo más apropiado pues existe gran probabilidad de que no se detecten ciertas inhomogeneidades en las series. Por esta razón el proyecto se centra en aplicar técnicas de remuestreo Bootstrap por bloques móviles (MBB) y Bootstrap Estacionario (SB) para obtener umbrales más exigentes con el 95 % de confianza, especificando de manera muy detallada la estructura y la metodología que se utiliza para llevar a cabo este objetivo. La aplicación se realizó a un conjunto de datos con mediciones diarías de la variable Humedad Relativa de 9 estaciones meteorológicas que monitorea el Grupo de Energías Alternativas y Ambiente (GEAA) en el periodo 2015 al 2017. Los resultados más confiables que permiten minimizar el valor del error cuadrático medio (RMSE) y el valor de la prueba de homogeneidad normal estándar (SNHT) en las series homogéneas se lograron con el umbral obtenido con remuestras SB y se presentan en la sección de Resultados.

Palabras clave: Homogenización, series meteorológicas, inhomogenenidades, Bootstrap, SNHT, GEAA, MBB, SB.

ABSTRACT

The inhomogeneities in the time series can unleash extreme events which cause a negative impact in the field of study. In meteorology, these disturbances can represent meteorological events such as: heat waves, droughts, floods, strong winds, and even forest fires, and they originate due to errors in the recording of information or the gradual change of the seasons. These inhomogeneities must be detected and corrected from the data set through the homogenization process. With the R Climatol package, this entire process can be carried out using R code, but there is a problem when choosing the inhomogeneity correction threshold. The choice of the threshold is made subjectively, only observing the graphs of the maximum SNHT histograms, but it is not the most appropriate since there is a high probability that certain inhomogeneities in the series will not be detected. For this reason, the project focuses on applying Moving Block Bootstrap (MBB) and Stationary Bootstrap (SB) resampling techniques to obtain more demanding thresholds with 95% confidence, specifying in a very detailed way the structure and methodology to be applied. used to accomplish this goal. The application was made to a data set with daily measurements of the Relative Humidity variable of 9 meteorological stations that the Alternative Energy and Environment Group (GEAA) monitors in the period 2015 to 2017. The most reliable results that allow minimizing the value of the root mean square error (RMSE) and the value of the standard normal homogeneity test (SNHT) in the homogeneous series were achieved with the threshold obtained with SB resamples and are presented in the Results section.

Keywords: Homogenization, meteorological series, inhomogeneities, Bootstrap, SNHT, GEAA, MBB, SB.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

El estudio de la detección y corrección de inhomogeneidades en series climáticas es de suma importancia, por que nos permitirá tomar precauciones cuando vaya a ocurrir un evento meteorológicos provocado por el cambio climático. Este proceso se lo puede realizar de manera directa mediante varios paquetes como: MASH, CLIMATOL, ACMANT, HOMER, etc. Últimamente la homogenización de las series se ha realizado con ayuda del paquete R *Climatol* por que es el más actualizado y todo su proceso se lleva a cabo con código R. Sin embargo, la elección de los umbrales de detección de inhomogeneidades se obtienen de manera subjetiva, solamente observando los histogramas máximos SNHT que se obtienen al realizar un primer análisis exploratorio de los datos diarios. Estos histogramas tienen sus barras demasiado separadas y por ende obtener el valor SNHT de manera visual no es apropiado. Así, para solucionar este inconveniente se propone aplicar técnicas de remuestreo Bootstrap. Esta metodología utiliza la prueba de homogenidad normal estándar (SNHT) para detectar inhomogeneidades en las series. Una vez elegido el estadístico de la SNHT se obtiene las estimaciónes de los estadísticos de todas las remuestras Bootstrap construidas para posteriormente obtener el punto crítico de esta distribución que es representa como el umbral de detección de inhomogeneidades. Este umbral se obtiene de forma individual para cada serie, pero necesitamos que sea único para todas las series, por tanto se propone estimarlo mediante la media de todos estos umbra-

les. Por último se homogeniza las series fijando este umbral con el fin de detectar más inhomogeneidades en todas las series de estudio.

1.1. Objetivo general

Aplicar las técnicas de remuestreo Bootstrap por Bloques Móviles y Bootstrap Estacionario para establecer los umbrales de detección de inhomogeneidades en series meteorológicas que contienen información de mediciones diarias de 9 estaciones que monitorean el GEAA en el periodo 2015 al 2017 de la variable humedad relativa, de tal manera que estos umbrales nos permitan detectar y corregir de manera más confiable las anomalías al momento de homogenizar las series.

1.2. Objetivos específicos

- Recopilar información de temas relacionados con el problema que se plantea en el proyecto de investigación.
- Realizar un tratamiento previo a la base de datos que va a ser proporcionada por el Grupo de Energías Alternativas y Ambiente(GEAA) para obtener una nueva base con información de las mediciones diarias de la variable Humedad Relativa de las 9 estaciones.
- Obtener series homogéneas mediante el uso del paquete R Climatol, que nos servirán como hipótesis nulas al momento de aplicar la prueba de homogeneidad normal estándar (SNHT) a las series individuales.
- Aplicar las técnicas de remuestreo Bootstrap por Bloques Móviles y Bootstrap Estacionario a las series meteorológicas homogéneas para obtener umbrales de detección de inhomogeneidades más exigentes y nuevas series homogéneas más eficientes .

1.3. Alcance

En este proyecto, se especifica de manera completa el funcionamiento del paquete R *Climatol*, así como la metodología que usa para homogenizar las series (creación de series de referencia, la detección de inhomogeneidades). Además, la implementación y la teoría elemental acerca de las técnicas de remuestreo Bootstrap por Bloques Móviles y Bootstrap Estacionario que van a ser de utilidad para obtener los umbrales de detección de inhomogeneidades con alta confiabilidad. Asimismo se presenta el algoritmo que nos permite calcular estos umbrales y por último se homogeniza las series añadiendo este umbral como un paramétro de entrada fijo para obtener resultados confiables los cuales se validan con el criterio propuesto por autor [8].

1.4. Marco Teórico

En esta parte se presenta la teoría que se necesita para llevar a cabo los objetivos planteados en el Trabajo de Integración Curricular.

1.4.1. Homogenización de series climatológicas con el paquete Climatol del sofware R Studio

Antecedentes

Los errores en la recolección de mediciones es inherente dependiendo del campo en el que nos encontremos y en la meteorología no es la excepción. Por ello, poder detectar y corregir estas inhomogeneidades en las series de tiempo de variables climáticas es muy antigua como la climatología misma. En la antigüedad, el estudio climatológico se realizaba de manera manual, los métodos gráficos para la detección de estas inhomogeneidades arrojaba resultados útiles, como el de las dobles masas de [9], y la mayoria de veces el objetivo se limitaba a la obtención de promedios referidos a un periodo común de observaciones, para lo que bastaban procedimientos como los de las diferencias o las proporciones [5], estos métodos se los realizaba a bases con una cantidad mínima de

observaciones.

Hoy en día existen paquetes estadísticos con los cuales es posible realizar la construcción de series climatológicas completas, a partir de una estimación de los datos ausentes o faltantes, estos paquetes trabajan con cantidades extensas de datos, por ejemplo se conoce el caso del paquete *MASH*, en este paquete esta implementado el método MASH (Análisis múltiple de series para homogeneización), fue desarrollado en el Servicio Meteorológico de Hungría como un método de homogenización relativa para detectar y corregir inhomogeneidades al no asumir que las series de referencia son homogéneas. Los posibles puntos de ruptura se pueden detectar y ajustar mediante comparaciones mutuas de series dentro de la misma zona climática. Este método trabaja bien, pero presenta el inconveniente de tener restringido su funcionamiento a plataformas y sistema operativo concreto (DOS).

En cambio el sofware estadístico R, es una multiplataforma (hay versiones para otras máquinas además de para los PC's) y funciona con distintos sistemas operativos (GNU-Linux, Solaris, Windows, etc), lo que permite su uso en una amplia gama de entornos de trabajo. En este software se encuentra implementado el paquete *Climatol*, el cual nos permite de igual manera realizar la construcción de series climatológicas, a partir de una estimación de los datos faltantes.

Paquete Climatol

Nos enfocamos en el paquete *Climatol* del software estadístico R, por que es uno de las más actualizados y su metodología es sencilla. Todo se lleva a cabo con código R y muestra de manera gráfica todo el procesamiento que se le esta dando a la información. También da la posibilidad a la persona que esta ejecutando el código, modificar ciertos parámetros de acuerdo al conocimiento o experiencia que tiene o simplemente de acuerdo a comportamiento de las variables en su región de estudio.

El paquete R *Climatol* tiene funciones para homogenizar, tiene funciones para llevar el control de calidad y tiene funciones para rellenar datos ausentes en diferentes series de variable climática.

Metodología

- **Creación de las series de referencia**

  El relleno de datos faltantes inicialmente el paquete lo realizaba con ayuda de series de referencia, las cuales eran construidas a partir de series de distancias más cercanas. Sin embargo, debido a la gra importancia que se le a dado a este tema y dado que el paquete cada vez se actualiza. Hoy en día para rellenar datos faltantes se adaptó el método propuesto por [12] que permite reconstruir series diarias de referencia usando la media de los valores de los vecinos más cercanos normalizados mediante división por sus medias. Este método no es el único pues el paquete R *Climatol* ofrece hacerlo también restando las medias o incluso estandarizando completamente las series. Así, dado m_{X_t} y s_{X_t} la media y desviación típica de la serie diaria X_t, los métodos para llevar a cabo este proceso se presentan a continuación:

  1. Dividir por la media: $x = \frac{X_t}{m_{X_t}}$
  2. Restando la media: $x = X_t - m_{X_t}$
  3. Estandarización completa: $x = \frac{X_t - m_{X_t}}{s_{X_t}}$

  Sin embargo, de las 3 opciones anteriores, la más utilizada es la estandarización completa de la información, pero el problema al aplicar está metodología es que las series durante el periodo de estudio no son completas y por tanto no se conoce la media y desviación estandár de las series originales. Por tanto, *Climatol* lo que hace es, primero calcular estos dos parámetros solamente con la información disponible sin tomar en cuenta los valores faltantes, luego rellena estos datos ausentes utilizando los parámetros m_x y s_x que se obtuvieron anteriormente. Con estas nuevas series, se procede nuevamente a calcular los parámetros m_x y s_x para rellenar los datos faltantes de la base original. Este proceso se realiza de manera repetida hasta conseguir que las medías obtenidas en las últimas iteraciones no difieran al redondearla con la precisión inicial de los datos. Luego se estandariza toda la información de las series, me-

diante la expresión

$$\hat{y} = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j}$$

donde, $\hat{y}$ es una medición estimada usando n mediciones x_j más cercanas disponibles en paso temporal, y w_j es el peso de cada medición estimada. Los pesos w_j de cada una de las mediciones estimadas para las series, dependen de las distancias (d_j) y de la distancia en la que el peso se reduce a la mitad (h).

$$w_j = \frac{1}{1 + \frac{d_j^2}{h^2}}$$

Por otro lado, para ajustar las distancias de las mediciones más cercanas, se utiliza el modelo de Regresión Ortogonal $\hat{y}_i = x_i$ para minimizar las distancias perpendiculares de cada punto y la misma (regresión de tipo II), en lugar de ajustar en dirección vertical con el modelo lineal $\hat{y}_i = r \cdot x_i$ (regresión de tipo I) como se realiza normalmente, siendo r el coeficiente de correlación entre las series x e y. En la Figura 1.1, se observa un gráfico que describe cada una de las regresiones de tipo I y de tipo II respectivamente.

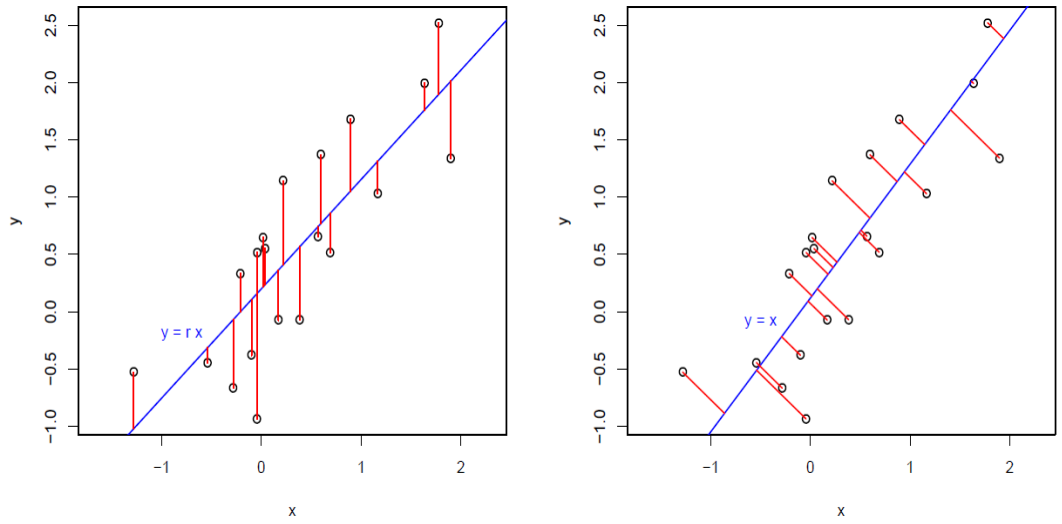


Figura 1.1: Descripción gráfica de los modelos para minimizar la distancia entre las mediciones y la recta. De tipo I (izquierda) y II (derecha).

De manera general, la serie de referencia es construida mediante

varias series de estaciones que se encuentran próximas o que estén bien correlacionadas [7]; es decir, mediante un promedio del valor de las estaciones más cercanas, ponderado por el inverso de la distancia con la estación de análisis, asignando mayor peso a las estaciones mejor correlacionadas [8], mediante la siguiente expresión:

$$V_r = \frac{\sum_{j=1}^{n} V_i \left(1 + \frac{di^2}{h^2}\right)}{\sum_{j=1}^{n} \left(1 + \frac{di^2}{h^2}\right)}$$

donde, V_r es la medición calculada para la serie de referencia, V_i son las mediciones de las estaciones de estudio, d es la distancia en km de la estación de análisis y la estación del registro considerado y h es la distancia en km donde el peso de ponderación es la mitad del peso de una segunda estación ubicada en la misma localidad. En la Figura 1.2, se presenta gráficamente como se toman los parámetros d y h usando Interpolación mediante Distancia Inversa Ponderada (IDIP), la cual determina los valores a través de una combinación ponderada linealmente de un conjunto de puntos de muestra.

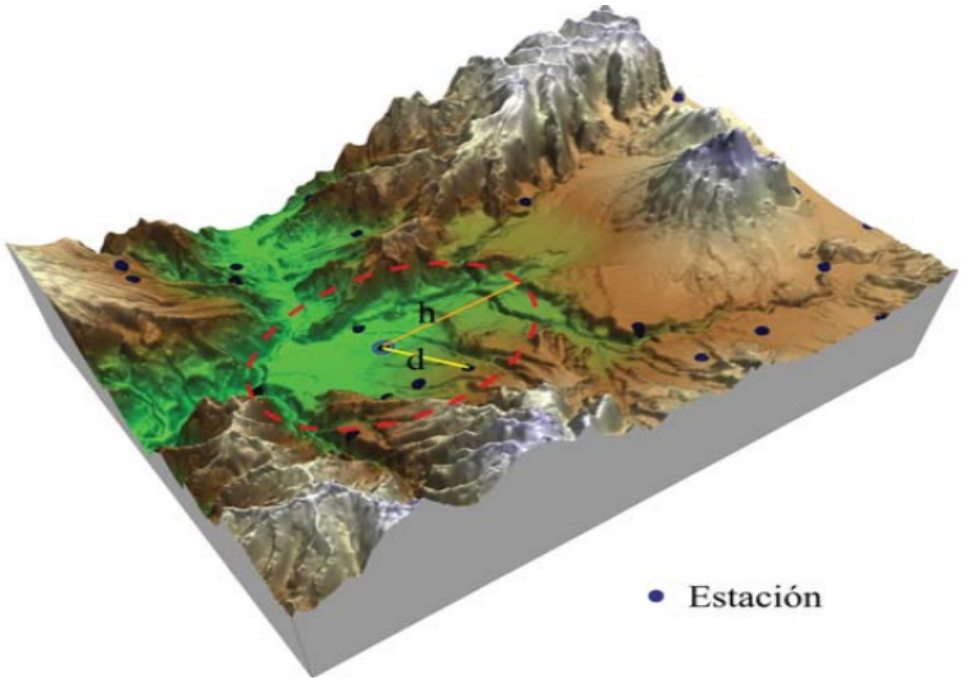


Figura 1.2: Representación gráfica para la elección de los parámetros d y h mediante IDIP

- **Detección de inhomogeneidades**

  La detección de inhomogeneidades se realiza con la prueba de homogeneidad normal estándar (SNHT) para tendencias de [1], esta prueba se encuentra implementada como función del paquete Climatol para detectar inhomogeneidades. Cabe recalcar que SNHT para iniciar el proceso toma como hipótesis nula series homogéneas normalizadas, Z_t, definida como

$$Z_t = \frac{(X_t - \bar{X}_t)}{\sigma_{X_t}}$$

  donde X_t es la serie meteorológica homogénea con mediciones diarias de la humedad relativa, μ_X y σ_X son la media y desviación típica de la serie respectivamente.

  Las hipótesis correspondientes se dan a continuación:

  H_0 : Caso trivial (Se asume homogeneidad en toda la serie)

  H_1 : Hay un cambio gradual en la media (una ruptura).

  o de manera matemática puede representarse como

$$H_0 : Z_i \in N(0,1) \text{ con } i \in \{1, 2, \ldots, n\}$$
$$H_1 : \begin{cases} Z_i \in N(\mu_1, 1) \text{ con } i \in \{1, 2, 3, \ldots, a\} \\ Z_i \in N(\mu_2, 1) \text{ con } i \in \{1 + a, \ldots, n\} \end{cases}$$

  La estadística de prueba se define como

$$T^s_{max} = \max_{1 \le a \le n-1} \{T^s_a\} = \max_{1 \le a \le n-1} \{a\bar{z}_1^2 + (n-a)\bar{z}_2^2\}$$

  donde $\bar{z}_1^2$ y $\bar{z}_2^2$ son los valores medios antes y después del cambio, n es la longitud de la serie. El valor medio correspondiente de a es el punto de ruptura más probable. La hipótesis nula puede rechazarse si T^s_{max} está por encima del nivel de significancia seleccionado y además, T^s_a compara el promedio de las mediciones de los primeros a días con los últimos (n-a) días registrados.

1.4.2. Técnicas de remuestreo Bootstrap

Motivación del principio Bootstrap

El uso del término Bootstrap se deriva de la frase para levantarse por los propios medios (*to pull oneself up by one's Bootstraps*), que se cree que está basada en una de las aventuras del barón Munchausen del siglo XVIII, de Rudolph Erich Raspe. El Barón había caído al fondo de un lago profundo. Justo cuando parecía que todo estaba perdido, pensó en levantarse por sus propios medios.

Introducción

La técnica Bootstrap es un método de remuestreo, introducido inicialmente por [15] para variables independientes y luego ampliado para tratar con variables dependientes más complejas por varios autores, es una clase de métodos no paramétricos que permiten al estadístico realizar inferencias estadísticas sobre una amplia gama de problemas sin imponer muchos supuestos estructurales sobre el proceso aleatorio subyacente de generación de datos. Actualmente, existen varios trabajos, que describen diferentes aspectos de la metodología bootstrap en diferentes niveles de sofisticación y generalidad, por ejemplo, Hall (1992), [15], Shao y Tu (1995), Davison y Hinkley (1997) y [10], entre otros, Además, varios artículos en la literatura brindan una descripción general de varios aspectos del método Bootstrap para series temporales, por ejemplo: [6], [17], Kreiss y Lahiri (2012), Berkowitz y Kilian (2000), Bose y Politis (1995). Los trabajos de Paparoditis y Politis (2009) y de Ruiz y Pascual (2002) se centran especialmente en series temporales financieras, mientras que McMurry y Politis consideran la metodología de remuestreo Bootstrap para datos funcionales.

El Bootstrap es un método estadístico que propone una soluciones a problemas matemáticamente intratables. Se utiliza para estimar errores estándar, sesgos o construir intervalos de confianza si no se cumplen los supuestos, no se conoce la distribución o no existe una solución teórica. Para ello procede mediante remuestreo, es decir, obteniendo muestras mediante algún procedimiento aleatorio que utilice la muestra original

de igual dimensión. Su principal ventaja es que no requiere hipótesis sobre el mecanismo generador de los datos. Sí las requiere, aunque suelen ser más relajadas, para obtener propiedades asintóticas del mismo. Por otra parte, su implementación en ordenador suele ser sencilla, en comparación con otros métodos. Su principal inconveniente es la necesidad de computación intensiva, debido a la fuerza bruta del método de Monte Carlo. Sin embargo, con la capacidad computacional actual, esta mayor carga computacional del Bootstrap no suele ser un problema hoy en día.

En este proyecto, nuestro objetivo es aplicar la tecnica Bootstrap para obtener umbrales de corrección de inhomogeneidades en las series meteorológicas, especialmente para la variable humedad relativa

Método Bootstrap

El método bootstrap se basa en tratar el conjunto de datos como una población y extraer muestrar con reemplazo de él B veces, posteriormente se elige el estadístico de prueba que se va a utilizar (media, mediana, varianza, desviación estándar u otros en específico) para evaluar cada muestra, generando así B estadísticos de prueba para finalmente calcular el error estándar de los B valores obtenidos con anterioridad. A continuación, en la Figura 1.3, se presenta el esquema del método Bootstrap [16].

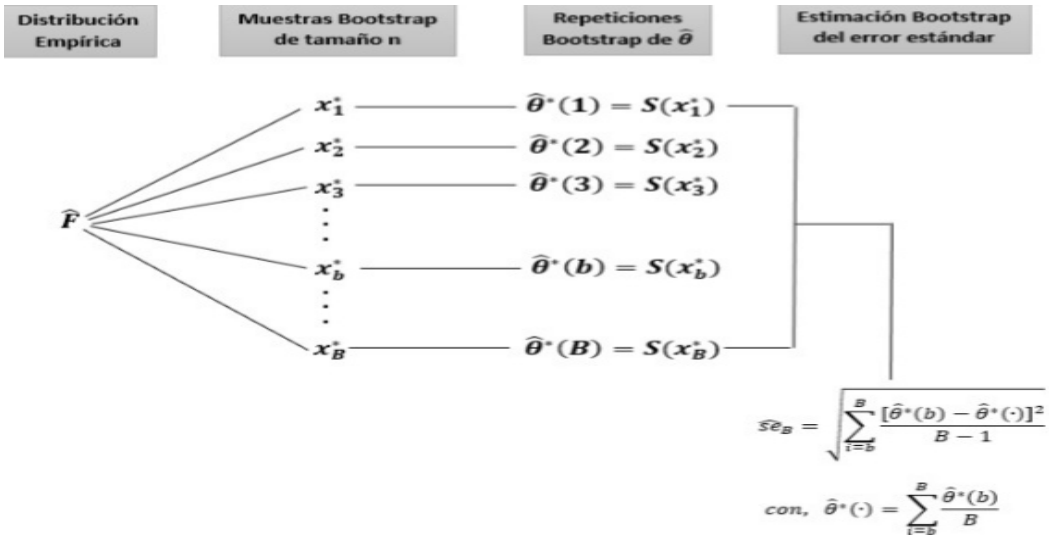


Figura 1.3: Esquema del método Bootstrap

Por otro lado, se presenta el algoritmo que se puede implementar para el método Bootstrap.

Algoritmo 1: Bootstrap simple.

1: Dada $X_i = (x_1, x_2, \ldots, x_n)$ una muestra aleatoria simple de cualquier variable de estudio.

2: Generar B muestras aleatorias con repetición $X_1^*, X_2^*, \ldots, X_B^*$ de igual dimensión que la base original.

3: Se obtienen B valores del Bootstrap $\hat{\theta}^*$ donde,

$$\hat{\theta}^*(b) = S(X_b^*), \qquad \forall\ b \in \{1, 2, \ldots, B\}$$

donde $S(X_b^*)$ es la estadística a usar (media, mediana, varianza, desviación estándar, otros estadísticos).

- Media: $S(X_b^*) = \bar{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- Mediana:
  1. Ordenar los valores de $x_i,$ con $i \in \{1, 2, \ldots, n\}$
     - Si n es par, entonces $S(X_b^*) = x_{\lceil \frac{n}{2} \rceil + 1}$
     - Si n es impar, entonces $S(X_b^*) = \frac{x_{\lceil \frac{n}{2} \rceil} + x_{\lceil \frac{n}{2} \rceil + 1}}{2}$
- Varianza: $S(X_b^*) = S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$
- Desviación estándar: $S(X_b^*) = S = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$
- Otros.

4: Se calcula el error estándar del bootstrap

$$\hat{se}_B = \sqrt{\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}}$$

donde, $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \dfrac{\hat{\theta}^*(b)}{B}$

1.4.3. Bootstrap para series de tiempo

Los métodos Bootstrap para series de tiempo, han tenido una gran acogida en el estudio de la estadística, dado que se han propuesto varios métodos de remuestreo para series temporales, unos basados en determinados modelos (AR(p), MA(q), ARMA(p, q), autorregresiones no paramétricas, entre otros) [2]. Sin embargo, aveces la estructura de dependencia de la serie temporal no permite reconocer el modelo "correcto". Es por eso que propusieron el método Bootstrap para métodos no basados en modelos, principalmente en el método Bootstrap por Bloques. Existen 4 métodos de Bootstrap por bloque [3]: los métodos Bootstrap de bloques no superpuestos, Bootstrap de bloques móviles, Bootstrap de bloques circulares y Bootstrap estacionario, sin embargo, nos centramos en los métodos Bootstrap por bloques móviles (MBB) y el Bootstrap estacionario, con el fin de detectar de manera más confiables los umbrales de corrección de inhomogeneidades en las series.

Bootstrap por bloques móviles (MBB)

El Bootstrap por bloques móviles (moving blocks bootstrap o MBB), es propuesto en [11], a diferencia del método Bootstrap de bloques no superpuestos, este método hace que los bloques se superpongan. Esto permite más bloques que si no se les permitiera superponerse. Así, para un conjunto de datos de n observaciones, el MBB divide los datos en $N = n - l + 1$ bloques superpuestos de longitud l donde, por simplicidad, suponemos que l divide a n. Entonces podemos definir los bloques MBB de X_n como [3]:

$$\begin{aligned}
B_1 &= (x_1, x_2, \ldots, x_l) \\
B_2 &= \quad (x_2, x_3, \ldots, x_{l+1}) \\
B_1 &= \quad\quad (x_3, x_4, \ldots, x_{l+2}) \\
&\vdots \quad\quad\quad\quad \vdots \\
B_N &= \quad\quad\quad\quad (x_{n-l+1}, x_{n-l+2}, \ldots, x_n)
\end{aligned}$$

De estos N bloques, se extraerán $b = \frac{n}{l}$ bloques al azar con reemplazo. La muestra de arranque se obtiene uniendo los bloques b. Una ventaja de

este método en comparación con el método NBB es que al permitir la superposición de bloques, tenemos una gama más amplia de bloques para muestrear. Esto es especialmente útil cuando la muestra es pequeña.

La longuitud óptima del bloque se obtuvo como en [14], el cual utiliza un algoritmo de validación cruzada que minimiza el error cuadrático medio (MSE) analizando varias longitudes de bloques en cada una de las series de estudio.

A continucación se presenta el algoritmo MBB.

Algoritmo2: Bootstrap por bloques moviles MBB.

1: Fijar un entero positivo, b, el tamaño del bloque, y tomar $k = \lceil \frac{n}{b} \rceil$.
2: Definir los bloques (o submuestras):

$$B_{i,b} = (X_i, X_{i+1} \ldots, X_{i+b-1}) \quad \text{ó} \quad B_i, \text{ para todo } i \in \{1, 2, \ldots, n-b+1\}$$

3: Arrojar k observaciones (bloques), $\xi_1, \xi_2, \ldots, \xi_k$, con distribución equiprobable sobre el conjunto de posibles bloques: $B_1, B_2, \ldots, B_{n-b+1}$. Cada ξ_i es un vector b-dimensional
4: Definir X^* como el vector formado por las n primeras componentes de: $\{\xi_{1,1}, \ldots, \xi_{1,b}, \xi_{2,1}, \ldots, \xi_{2,b}, \ldots, \xi_{k,1}, \ldots, \xi_{k,b}\}$

Si tomamos $b = 1$, entonces $k = n$ y se obtiene el bootstrap ordinario. Por otra parte, si $b = n$, tenemos $k = 1$ y se obtiene el remuestreo degenerado, ya que todas las réplicas bootstrap coincidirían con la muestra original.

Bootstrap estacionario

El bootstrap estacionario, de Politis y Romano [13] , difiere de los otros tres métodos mencionados anteriormente en el sentido de que la longitud del bloque no es fija sino una variable aleatoria distribuida geométricamente con un valor esperado l. Debido a la longitud aleatoria del bloque, el número de bloques también es aleatorio. El método necesita de la elección de un número $p \in [0, 1]$ y puede presentarse como:

Algoritmo3: Bootstrap estacionario.

1: Definir los bloques: $B_{i,b} = (X_i, X_{i+1} \ldots, X_{i+b-1})$ para todo $n \in \mathbf{N}$, $i \in \{1, 2, \ldots, n\}$, y $X_t = X_{((t-1) \ mod \ n)+1}$, si $t \leq n$,
2: Arrojar realizaciones iid, $L_1, L_2, \ldots$, con distribución geométrica de parámetro p; es decir,

$$P(L_1 = m) = p(1-p)^{m-1}, \quad \forall m \in \{1, 2, 3, \ldots\}$$

3: Obtener enteros aleatorios, $I_1, I_2, \ldots$, con distribución equiprobable sobre el conjunto $\{1, 2, \ldots\}$.
4: Definir $X_1^*, X_2^*, \ldots, X_n^*$, con los n primeros valores obtenidos al unir los bloques $B_{I_1,L_1}, B_{I_2,L_2}, \ldots$.

1.4.4. Obtención de umbrales de detección de inhomogeneidades usando técnicas de remuestreo Bootstrap

Una vez planteada la teoria y funcionamiento de las técnicas de remuetreo MBB y SB, se procede a plantear un algoritmo que nos permite detectar inhomogeneidades en las series usando las remuestras MBB y SB respectivamente. El algoritmo se plantea de la siguiente manera [4]:

Algoritmo4: Obtención de umbrales de detección de inhomogeneidades usando remuestras MBB

1: Obtener series homogéneas de las 9 estaciones mediante el uso del paquete R Climatol. A partir del paso (2) de realiza el proceso de manera individual para cada una de las estaciones.
2: Obtener el estadístico de prueba T_{max}^s para la muestra observada.
3: Escoger la longitud del bloque óptima para las remuestras MBB.
4: Obtener R muestras MBB, donde la longuitud de cada bloque es fija y calcular el estadístico estimando $T_{max}^{s^*}$ para cada una de las muestras.
5: Aproximar la distribución que siguen los estadísticos de las muestras

MBB.

6: Calcular el umbral de detección de inhomogeneidades, obteniendo el quantil de la distribución del estadístico para el nivel de significacia α.

Se puede modificar el algoritmo para obtener este umbral usando remuestras SB, solamente cambiando paso (4) del anterior algoritmo por:

4: Obtener R muestras SB, donde la longuitud del bloque $l \sim geom(p)$) y calcular el estadístico de prueba $T_{max}^{s^*}$ para cada una de las muestras.

1.4.5. Homogenización de las series fijando el umbral de detección de inhomogeneidades

Una vez obtenidos estos umbrales de corrección de inhomogeneidades para cada una de las estaciones meteorológicas, se procede a estimar un único valor para este umbral. Esto se lleva a cabo estimando este umbral con el valor medio de estos umbrales obtenidos con ayuda del **Algoritmo4**.

$$snht1 = \frac{1}{N_S} \sum_{S=1}^{N_S} T_{0,95}^{S^*}$$

donde $snht1$ es el umbral de corrección de inhomogeneidades estimado para todas las series, N_S es el número de estaciones que se utilizan para el estudio y $T_{0,95}^{S^*}$ es el umbral de corrección de inhomogenidades obtenido con métodos Bootstrap de cada una de las estaciones con una confianza del 95 %

Este proceso se lo realiza debido a que al homogenizar las series usando el paquete R *Climatol* el umbral que se debe ingresar para la corrección de inhomogeneidades es global para todas las series. Con toda esta explicación la homogenización de las series se realiza de manera directa simplemente especificando de manera correcta los parámetros de entrada que utiliza la función *homogen* del paquete R *Climatol*.

Capítulo 2

Metodología

Una vez conocido el marco teórico sobre la aplicación de técnicas de remuestreo para la homogenización de series meteorológicas, en el presente capítulo se aplica la misma sobre datos diarios de la serie meteorológica humedad relativa que monitorea el GEAA, dando una descripción desde el preprocesamiento de las bases de datos que fueron concedidas, también se presenta de manera muy detallada los pasos que se utilizan para homogenizar las series con ayuda del paquete R *Climatol*. Estas series homogéneas sirven como parámetro de entrada de la hipótesis nula de la prueba de homogeneidad normal estándar(SNHT). Asimismo se presentan la aplicabilidad de las técnicas de remuestreo Bootstrap por bloques móviles y Bootstrap estacionario para la obtención de umbrales de corrección de inomogeneidades mediante el estadístico de SNHT. Adicionalmente, en el Anexo, se encuentra todo el código realizado para este capítulo.

2.1. Descripción de la base de datos

Las bases de datos contiene información de series meteorólogicas de 11 estaciones que monitorea el GEAA, especialmente de las variables: humedada relativa, temperatura promedio ambiente, la precipitación, etc. El área de estudio donde se encuentran las estaciones meteorológicas: Alao, Atillo, Cumandá, Espoch, Matus, Multitud, Quimiag, San Juan, Tixan,

Tunshi y Urbina, está localizada en la Sierra Central del Ecuador, en la Latitud -1,6647 y Longitud -78,6543. Estas bases contienen información de mediciones tomadas cada minutos a partir del año 2013 al 2021, dicha información se encuentra en una sola columna y con un formato no conveniente[1]. Para nuestro estudio se tomó en cuenta la información a partir del año 2015 al 2017 de la variable humedad relativa, en el Cuadro 2.1, se observa cómo esta estructurado los datos, por lo cual se procede a realizar un tratamiento previo. .

$X1$,$X40$,$Stat\_TA\_1m$
date,time ,Avg
1/1/15,12:00:10 AM,81.2342
1/1/15,12:01:08 AM,81.2112
1/1/15,12:02:08 AM,81.1233
1/1/15,12:03:08 AM,81.4322
1/1/15,12:04:08 AM,81.1123
1/1/15,12:05:08 AM,81.2231
1/1/15,12:06:08 AM,81.1534
1/1/15,12:07:08 AM,81.1242
1/1/15,12:08:08 AM,81.3282
1/1/15,12:09:08 AM,81.1432

Cuadro 2.1: Extracto de la información recuperada de la estación Matus de los primeros 10 minutos del año 2015 de la variable humedad relativa

2.2. Tratamiento de los datos

De la información compartida acerca de los datos, se tuvo que unir la información del año 2015 al 2018, debido a que las mediciones aún se encontraban en UTM (*Universal Transverse Mercator*); es decir, a las variables **tiempo** (fecha y hora local) se le debe restar 5 horas, así, los 300 primeros datos del año 2015 se convertian en datos finales del año 2014, de igual manera los primeros 300 datos del año 2018 completaban los datos del año 2017, es por esta razón que se unio las bases para luego filtrar solo datos a partir del año 2015-2017, en el cuadro 2.2 se puede

[1]Los datos están disponibles en la siguiente página web: https://liveespochedu-my.sharepoint.com/:f:/g/personal/estaciones_espoch_espoch_edu_ec/EmR4sgWPDvNGshM9JBqpuj4B_lR9KH1eV1J-imO8uV3M_A?e=7RH77c

observar la cantidad de datos tomados por minuto que deben existir cada año y como se sabe que el año 2016 es un año bisiesto, entonces contiene 366 días.

Año	**Mediciones**
2015	525600
2016	527040
2017	525600

Cuadro 2.2: Cantidad de mediciones por año.

Por otro lado, se necesita datos diarios para realizar la homogenización de las series, y por ende, se procedio a realizar una imputación previa, solamente a datos que les faltaban observaciones durante el día, por ejemplo supongamos que no existian datos a partir de las 20h00 del 14 de Febrero del 2015 hasta las 17h00 del 25 de Febrero del 2015, entonces en el Cuadro 2.3, se resume como se realizó la imputación, para luego transformar a datos diarios, es evidente que si no existen datos durante un día completo, entonces se tomo a estos como valores faltantes (NA's).

Año	**Imputación**
2015/02/14	SI
2015/02/15	NO
⋮	⋮
2015/02/24	NO
2015/02/25	SI

Cuadro 2.3: Imputación previa de datos faltantes.

2.2.1. Variable Meteorólogica Humedad Relativa

Los datos que se obtuvieron, para la variable meteorológica Humedad relativa en los primeros 7 días del año 2015 para aquellas estaciones que disponen de al menos 2 años de registros se resume en los Cuadros 2.4. Cabe recalcar que para esta variable se eliminó las estaciones San Juan y Quimiag, debido a que no existía información previa del periodo que se tomaron los datos.

Alao	Atillo	Cumandá	Espoch	Matus
75,2157888	88,4292673	97,1973152	70,6672604	87,2018236
78,9020569	93,2748909	88,7420541	66,6043104	80,6472819
77,2041611	92,7746187	84,3051722	67,3664618	84,6999694
85,3706951	93,2678	93,4782833	75,7044145	91,9532
81,4878583	93,8159243	83,3958680	69,1385694	87,0672104
77,0498138	89,5083847	88,6003326	69,8465534	86,9380506
77,2665472	87,6447277	87,4169451	66,8316236	90,6055479

Multitud	Tixan	Tunshi	Urbina
97,65418333	83,06317708	78,78169583	87,51107708
99,83755694	74,54481875	73,94785694	92,06249722
98,68595694	78,03698472	75,01530694	92,15003125
99,50768264	80,4758375	83,93023125	95,57171875
98,11963125	81,364025	78,03392847	94,18236319
97,08517431	72,71409653	75,97014861	89,44647292
99,83450139	73,80359167	74,15720556	88,97026181

Cuadro 2.4: Base de datos de la variable meteorológica Humedad Relativa en el periodo 2015-2017

Analizando la base de datos, se pudo corroborar que existen valores faltantes, en la Figura 2.1 se observa la cantidad de datos; es decir que las estaciones meteorológicas con más valores faltantes son: Espoch (27,21 %), Cumandá (13,52 %) y Multitud (11,69 %), además las estaciones meteorologicas con menor cantidad de valores valtantes son: Atillo (1,57 %) y Tunshi (0,27 %), dando un total del 5,4 % de datos faltantes en toda la base.

2.3. Proceso de homogenización de la serie meteorlógica humedad relativa mediantes el paquete de R Climatol

Una vez presentada la metodología que utiliza el paquete R *Climatol* para homogenizar las series en el **Capítulo 1**, en esta sección se presenta su aplicación práctica para la base de datos obtenida con el tratamiento previo, esta base contiene mediciones diarias de la variable Humedad

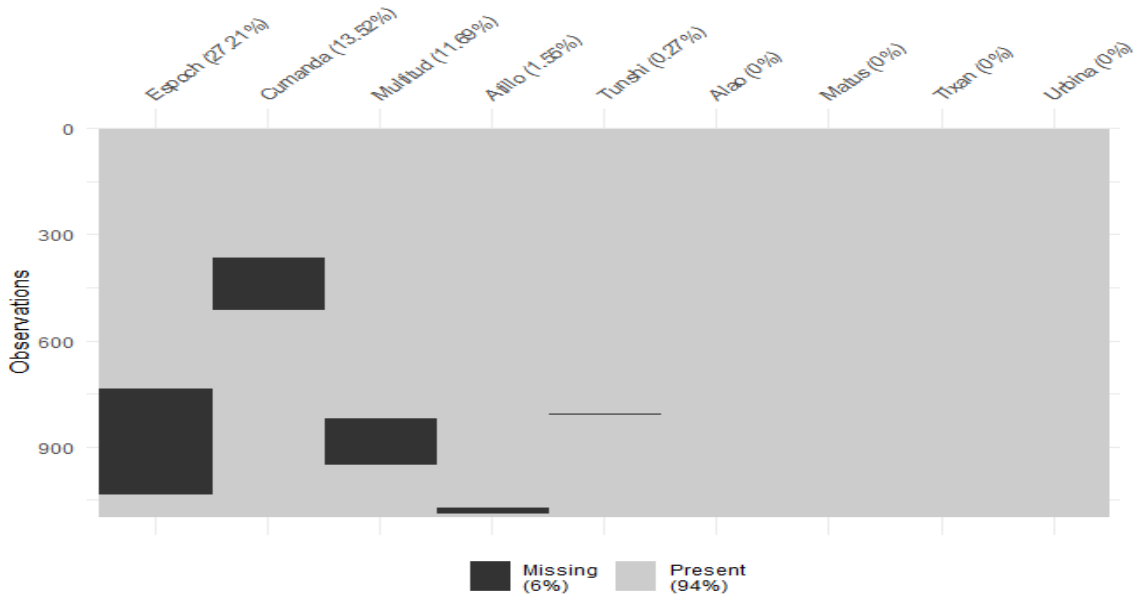


Figura 2.1: Valores NA's de la variable Humedad Relativa en el periodo 2015-2017

Relativa en el periodo 2015-2017 , para ello, empezamos preparando los ficheros de entrada.

2.3.1. Ficheros de Entrada

En paquete *Climatol*, para realizar o empezar su proceso de homogenización, solo requiere dos ficheros de entrada, uno que contiene información de las coordenadas X, Y y Z, donde X e Y es la longuitud y latitud respectivamente y Z es la altitud de cada estación. También se requiere los códigos y nombres de las estaciones meteorológicas que se toman en cuenta para el estudio, en el Cuadro 2.5 se puede observar los datos antes mencionados de cada una de las 9 estaciones. El otro fichero contiene la información de la humedad relativa de cada estación y se la puede ver en los Cuadros 2.4. Los dos ficheros de entrada deben compartir el mismo nombre *VAR-aaa-AAA*, donde *VAR* es las siglas que se utiliza para identificar la variable de estudio, *aaa* es el año inicial y *AAA* es el año final del periodo que se tomaron las mediciones. Por otro lado, para no tener confuciones al momento de leer estos dos ficheros en el software estadístico R, es necesario que tengan diferentes extensiones; es decir, el fichero que contiene información acerca de las estaciones deben tener extensión *.est* y el fichero que contiene información de la humedad relativa de todas

las estaciones debe tener extensión *.dat*. Por último los nombres de los ficheros son: *HR-2015-2017.dat* y *HR-2015-2017.est*, para los datos de la humedad relativa y para los datos de las estaciones respectivamente.

X	Y	Z	Codigo	Estación
-1,86948947	-78,54159151	3064	E05	Alao
-2,186991002	-78,54909957	3467	E09	Atillo
-2,210169027	-79,14532635	330,95	E06	Cumanda
-1,654601771	-78,67751976	2754	E01	Espoch
-1,555626471	-78,50548147	2471	E08	Matus
-2,609667821	-78,99700563	1483	E07	Multitud
-2,157635038	-78,76036938	3546	E04	Tixan
-1,747529386	-78,6263025	2840	E03	Tunshi
-1,488365	-78,712055	3642	E02	Urbina

Cuadro 2.5: Fichero de entrada de las estaciones con extensión *HR-2015-2017.est*

2.3.2. Análisis exploratorio de los datos

La función que permite llevar a cabo el proceso de homogenización de las series con el paquete R *Climatol* se llama *homogen*, para ejecutar esta función de manera trivial solo se debe especificar 3 parámetros de entrada: las siglas de la variable, el años inicial y el año final del periodo de estudio (*homogen('HR', 2015, 2017)*). Esta función se puede aplicar a datos, diarios, mensuales, bimestrales, trimestrales, semestrales o anuales, pero los umbrales para el rechazo de valores atípicos y umbrales corrección de inhomogeneidades es diferente dependiendo de la correlación cruzada entre series. Dado esto, se aconseja realizar un analisis exploratorio de los datos diarios, para ello únicamente se agrega el comando *expl=TRUE* (*homogen('HR', 2015, 2017, expl=TRUE)*). Este resumen se guarda en un archivo pdf (*HR-2015-2017.pdf*).

Como es de conocimientos el análisis exploratorio de datos hace referencia a un proceso de investigación en el que se usan estadísticas de resumen y gráficas para llegar a conocer los datos y comprender lo que se puede averiguar de ellos. Con el EDA, se pueden hallar anomalías en los datos, como valores atípicos u observaciones inusuales, revelar patrones, comprender posibles relaciones entre variables y generar preguntas o hi-

pótesis interesantes que se pueden comprobar más adelante mediante métodos estadísticos más formales. A continuación se presentan los gráficos de diagnósticos que nos arrojan el climatol, al aplicar a la variable metorológica humedad relativa a 9 estación que monitorea el GEEA.

Gráfico de la disponibilidad de datos

El primer gráfico hace referencia a la disponibilidad de datos en la base de estudio, en la Figura 2.2 se puede observar, que los datos disponibles se encuentran de color celeste y los faltantes de color blanco, esto ya se evidenció en la Figura 2.1, donde las estaciones que presentan mayor porcentaje de datos faltantes son: Espoch, Cumandá y Multitud.

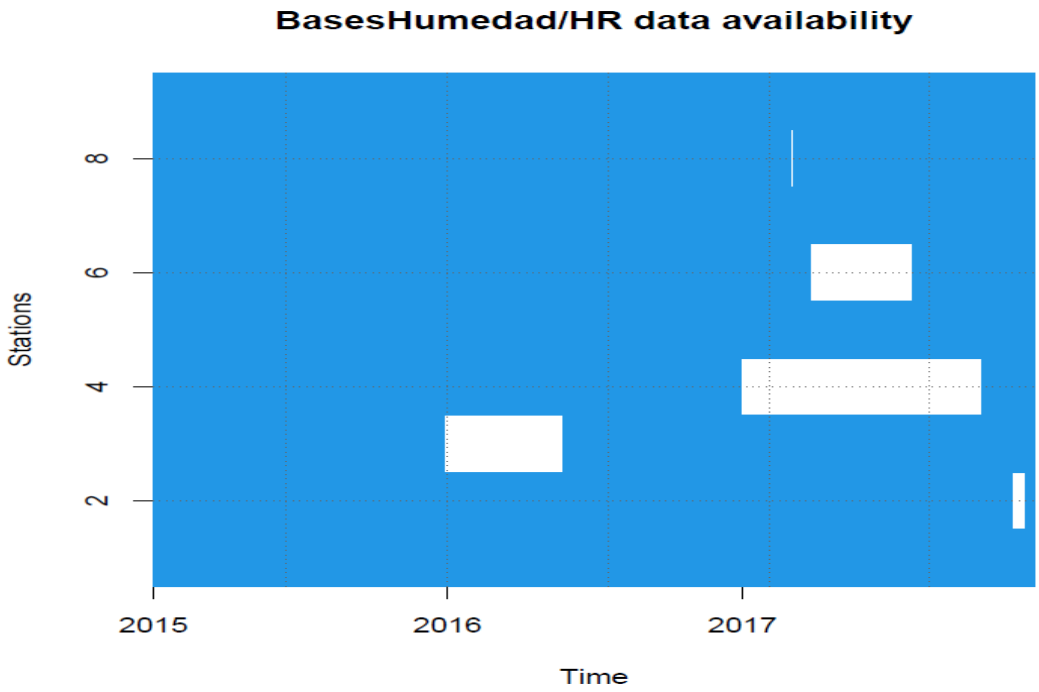


Figura 2.2: Cantidad de datos disponibles de la variable humedad relativa de las 9 estaciones en el periodo 2015-2017

Gráfico de cajas

Se presenta el diagramas de cajas de manera individual para las 9 estaciones. Como se envidencia en la Figura 2.3, en la estación 6 (Multitud) se observa que existen gran cantidad de mediciones menores al 20 %. Estos registros considerados como extremos son muy inusuales debido a

que en particular esta estación durante el periodo de estudio mantiene una media de 89,70 % aproximadamente. Además se observa que todas las estaciones presentan valores atípicos por debajo del umbral inferior, sin embargo existe una estación 2 (Atillo) que presenta valores atípicos por encima del umbral superior. La presencia de valores muy anómalos sería evidente en estos gráficos, lo que permitiría al usuario tomar medidas correctivas en los datos.

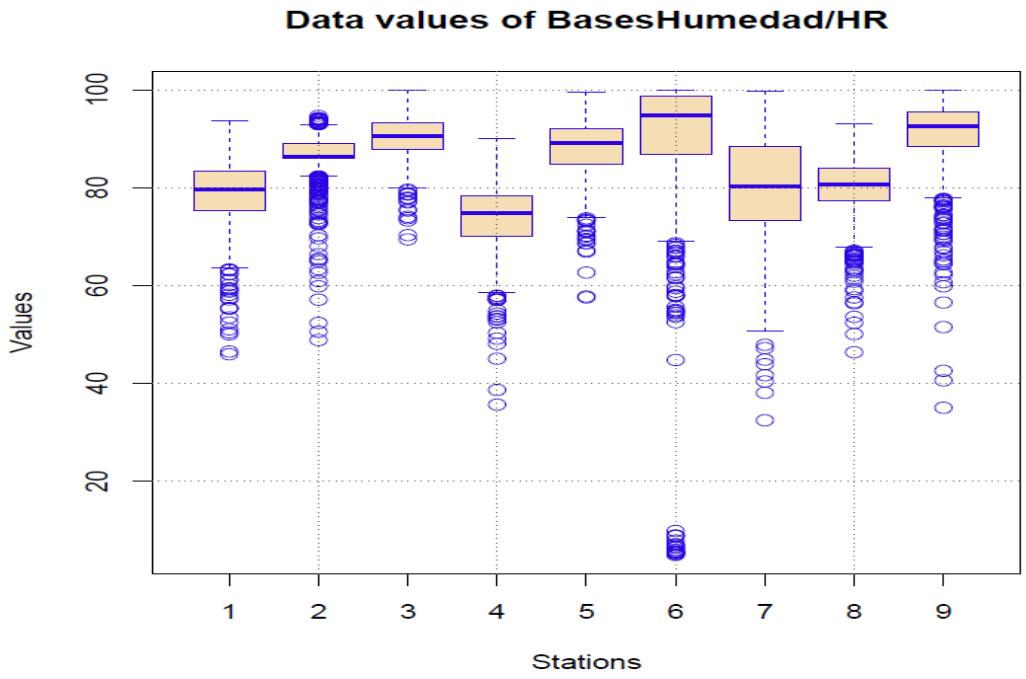


Figura 2.3: Diagrama de cajas de la variable humedad relativa en el periodo 2015-2017 de todas las estaciones

Histograma

En este caso, las mediciones se toman de manera global y se realiza un histograma con toda la información disponible, como se puede observar en la Figura 2.4, la mayor cantidad de mediciones se encuentran entre 60 % y 100 %, se puede además evidenciar que existe un sesgo hacia la izquierda, lo que conlleva a la existencia de valores átipicos. Además su distribución se asemeja a una Log-normal con un sesgo demasiado pronunciado.

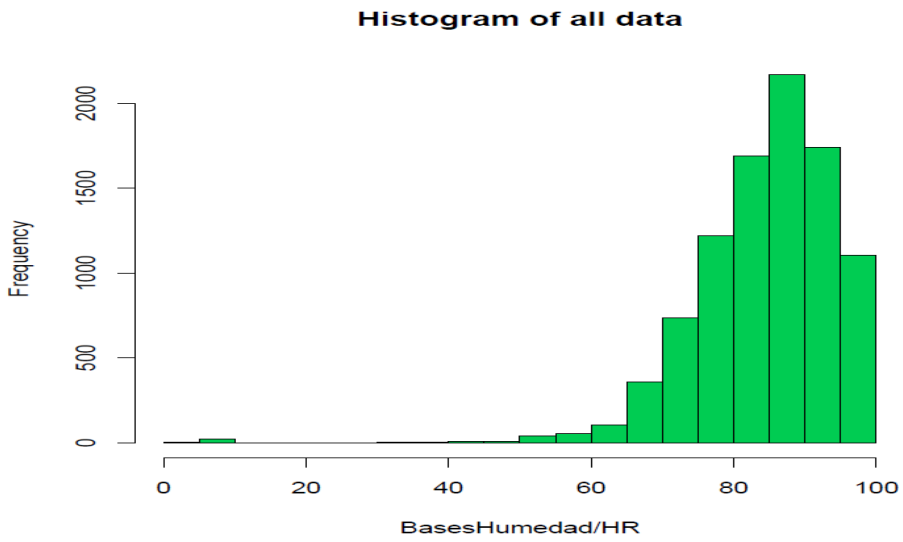


Figura 2.4: Histograma de las mediciones disponibles de todas las estaciones meteorológicas

Correlograma

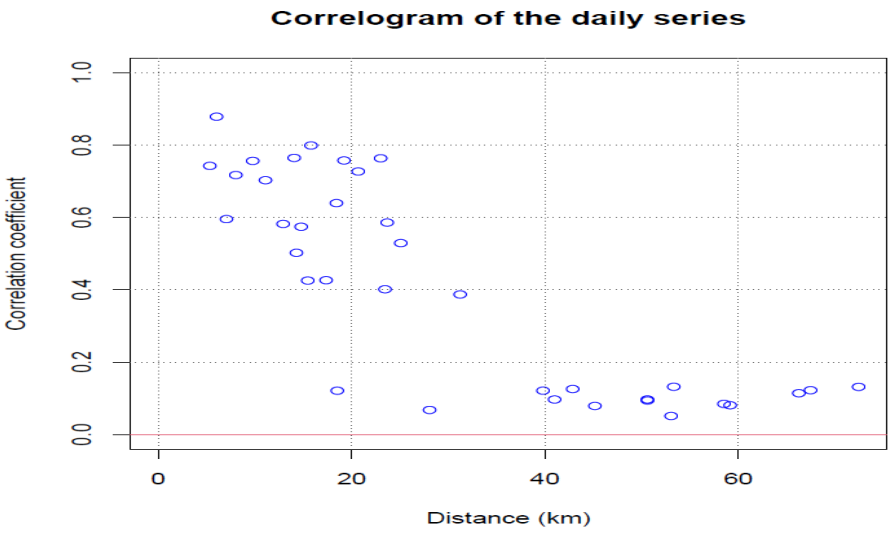


Figura 2.5: Correlograma de las series diarias

El correlograma representa las correlaciones entre las series y se construye a partir de los coeficientes de correlación en función de la distancia. El coeficiente de correlación ayuda a diferenciar las estaciones que a pesar de su proximidad, responden de forma semejante a las variaciones del factor climático en la estación de estudio por compartir una zona fisiográficamente similar. Las correlaciones disminuyen cuando la distancia entre estaciones aumenta, como se observa en la Figura 2.5. Cuanto

más altas sean las correlaciones, mayor será la fiabilidad de la homogeneización y el relleno de datos ausentes. En particular, las correlaciones deben ser siempre positivas, al menos dentro de un rango de distancias razonables.

Dendograma

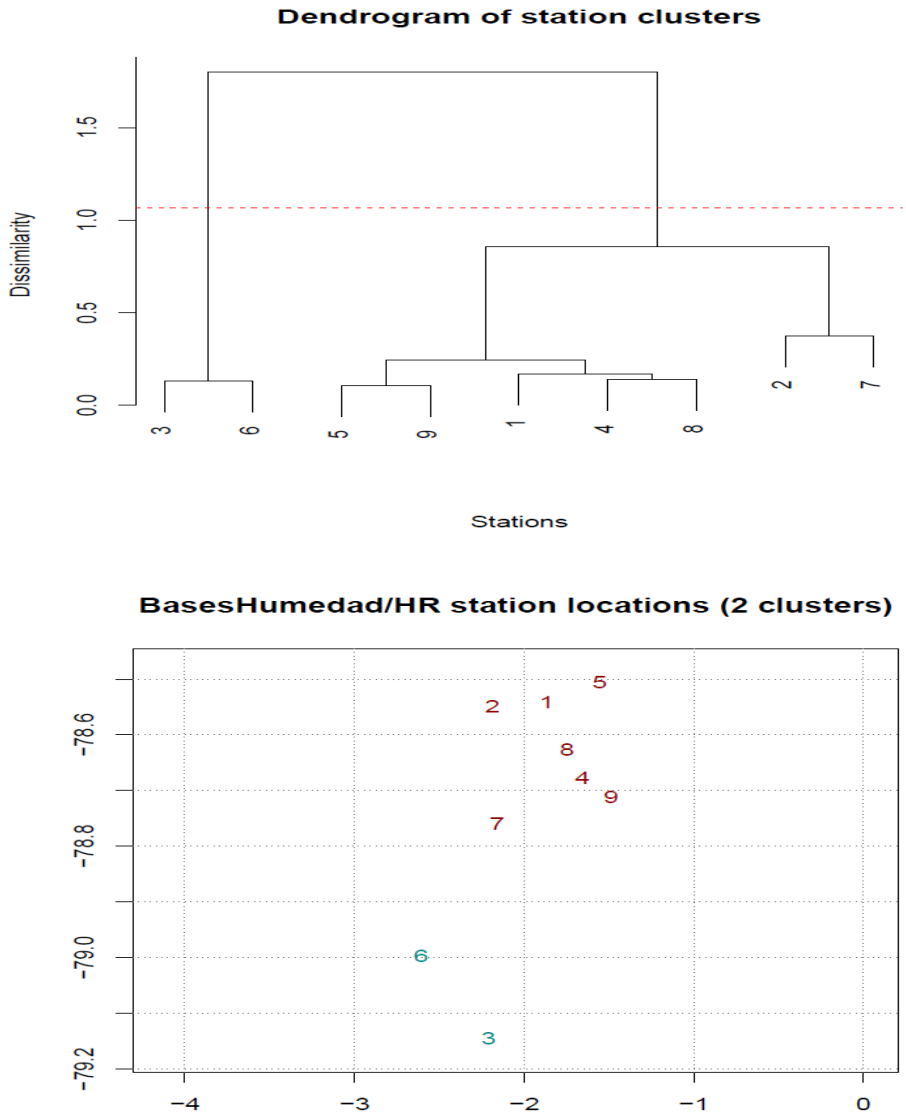


Figura 2.6: Grupos de estaciones con variabilidad similar

El dendograma es un gráfico de agrupación de estaciones, esta agrupación se realiza en base a la matriz de correlaciones para poder calcular la disimilaridad. En la Figura 2.6, se observa los grupos que se construyeron, como se puede observar el dendograma realizó 2 grupos, en el

primer grupo se encuentran las estaciones 1, 2, 4, 5, 7, 8 y 9; es decir, las estaciones meteorológicas: Alao, Atillo, Espoch, Matus, Tixan, Tunshi y Urbina, por otro lado en el segundo grupo se encuentran las estaciones 3 y 6; es decir, las estaciones meteorológicas Cumandá y Multitud.

Se envidencia que el primer grupo se encuentra ubicado al norte y el segundo grupo se encuentra al sur de la región de estudio, en este caso en la provincia de Chimborazo, pues es la provincia en donde se encuentran ubicadas todas las 9 estaciones.

Anomalías estandarizadas

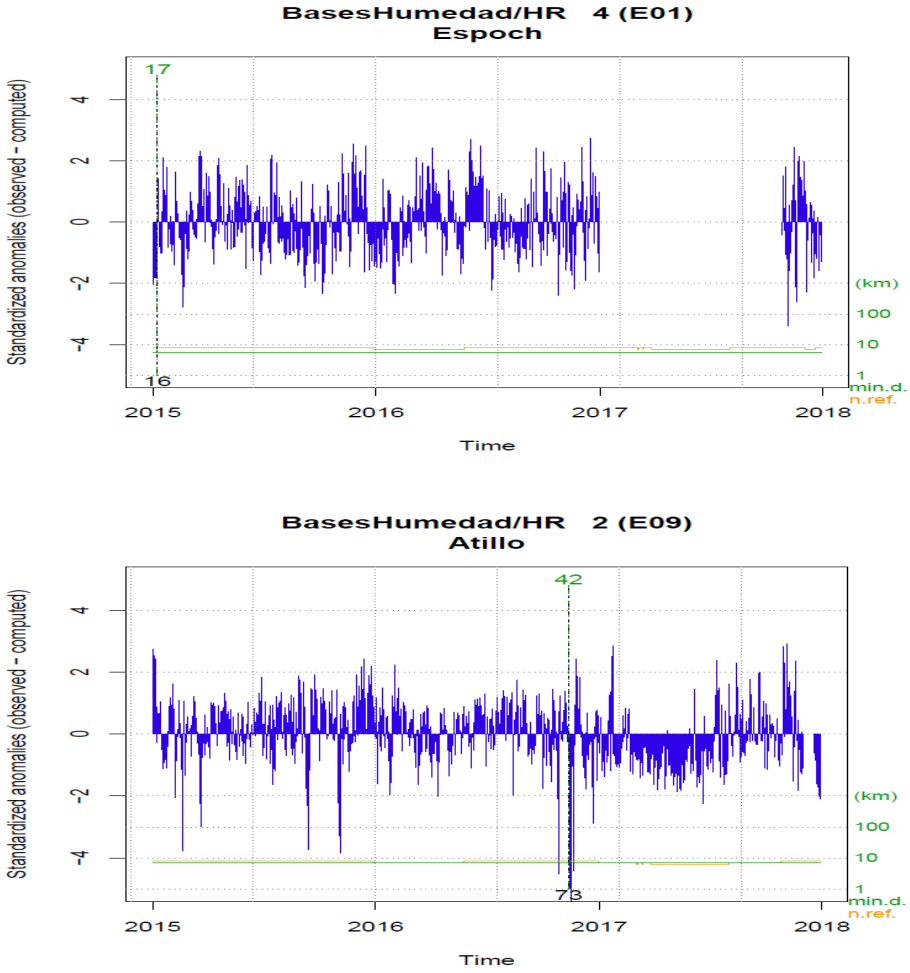


Figura 2.7: Gráfico de las anomalías estandarizadas de la estación Espoch y Atillo respectivamente

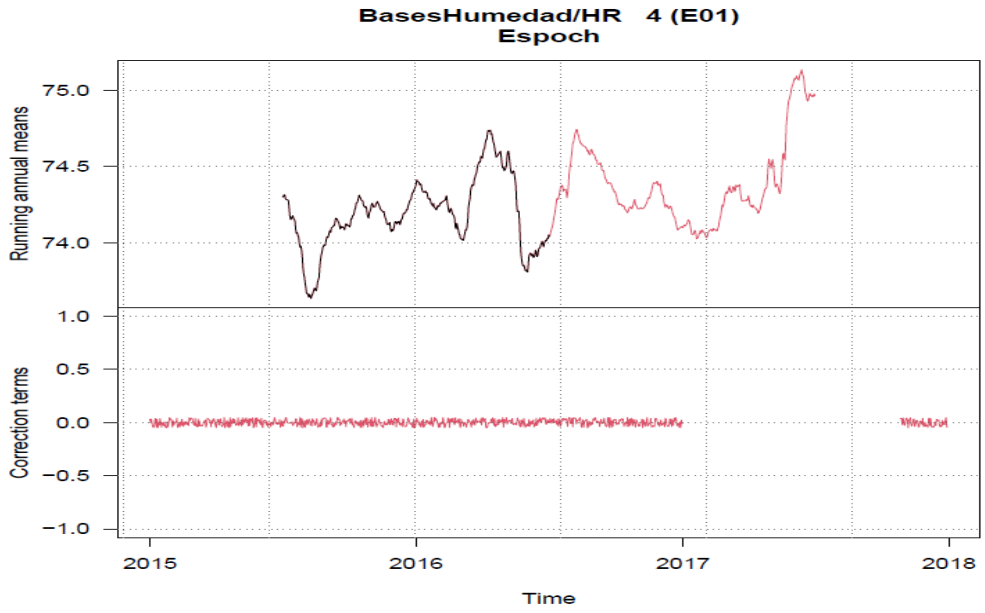
Una vez analizados los gráficos iniciales, en adelante se tienen gráficos para cada estación. Para interpretarlos solo tomamos en cuenta 2

estaciones como referencia: la estación Espoch por que tiene más valores faltantes y la estación Atillo por que no tienen muchos valores faltantes. En la Figura 2.10 se presentan los gráficos de las anomalías estandarizadas de las estaciones antes mencionadas.

Está gráfica nos ayuda a identificar en que punto existe una ruptura en de serie o simplemente en que punto la serie deja de ser homogénea. Dada esta explicación la serie de la estación Espoch deja de ser homogénea a inicios del año 2015 con un valor SNHT máximo de 17 aplicada a ventanas escalonadas superpuestas marcadas con una línea vertical color verde y en el mismo punto un SNHT máximo de 16 aplicada a toda la serie. A diferencia que la serie de la estación Atillo deja de ser homogénea a finales del año 2016 con un valor de SNHT máximo de 42 aplicada a ventanas escalonadas superpuestas y en el mismo punto un SNHT máximo de 73 aplicada a toda la serie; es decir, que apartir de las líneas que representan los valores máximos de SNHT las series sufre un cambio en su comportamiento.

Reconstruccíon de las series de las estaciones meteorológicas

La reconstrucción de estas series se realiza apartir del punto en el que se detectó la rutura en la serie. En la Figura 2.8, se presentan las series originales (negro) y reconstruidas (rojo) de las 2 estaciones de análisis.



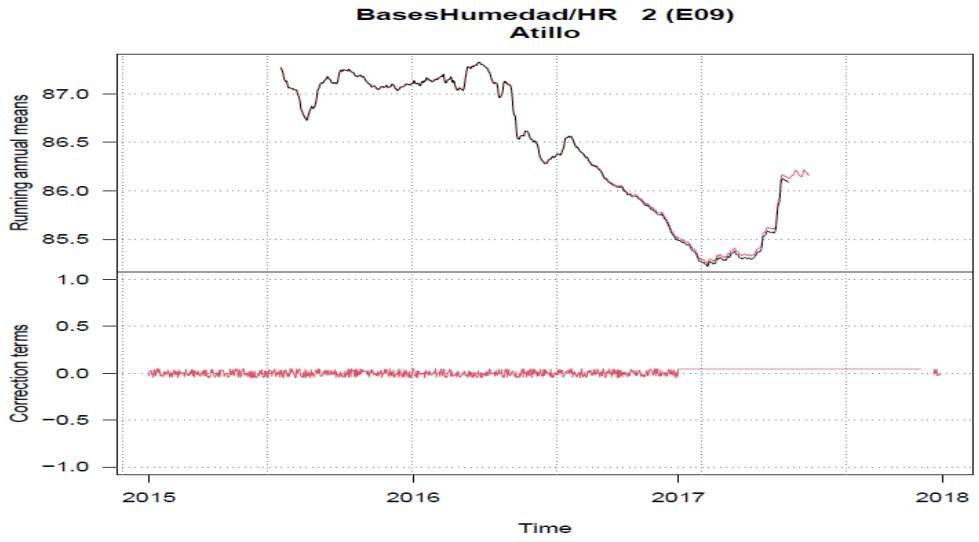


Figura 2.8: Gráfico de las series originales y construidas de las estaciones meteorológicas Espoch y Atillo respectivamente

Además en la segunda parte de la figura se encuenta una linea roja, la cual nos da a conocer que operación realizó el paquete R *Climatol* para reconstruir la serie homogénea, ya sea si a la serie original le aumentó, le disminuyó o simplemente mantuvo el valor original.

Anomalías Normalizadas

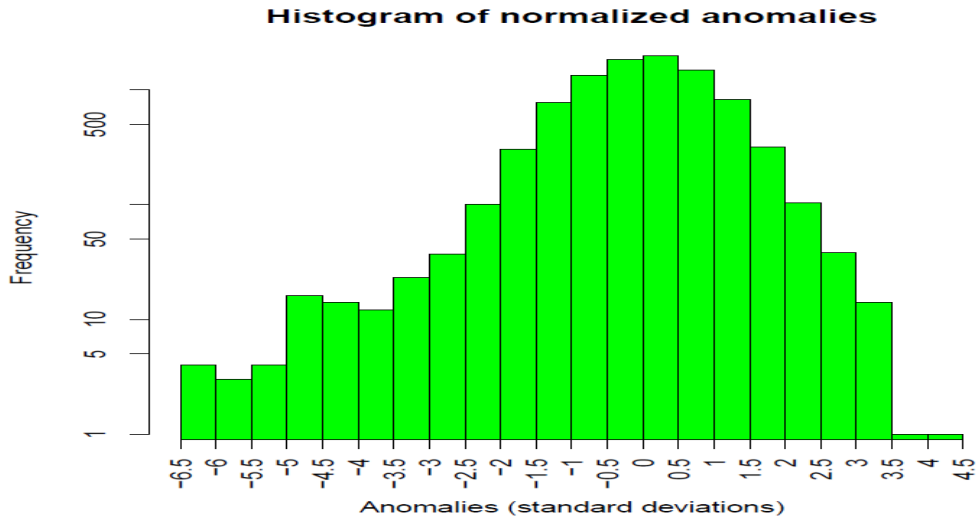


Figura 2.9: Histograma de anomalías normalizadas

El histograma de anomalías normalizadas ayuda a elegir de manera subjetiva los umbrales para rechazar anomalías, suponiendo que son

errores en el registro de la información y pueden eliminarse del conjunto de datos. En la Figura 2.9 se ve que presenta algo de sesgo hacia la izquierda, y por tanto podrían aceptarse todos los valores que se encuentre entre -3.5 y 3.5. Cabe recalcar que se eliminaría los datos con anomalías absolutas superiores a 5 desviaciones típicas si no fijamos estos umbrales.

Histogramas máximos SNHT

Estos histogramas tienen como objetivo elegir los umbrales de corrección de inhomogeneidades en las series.

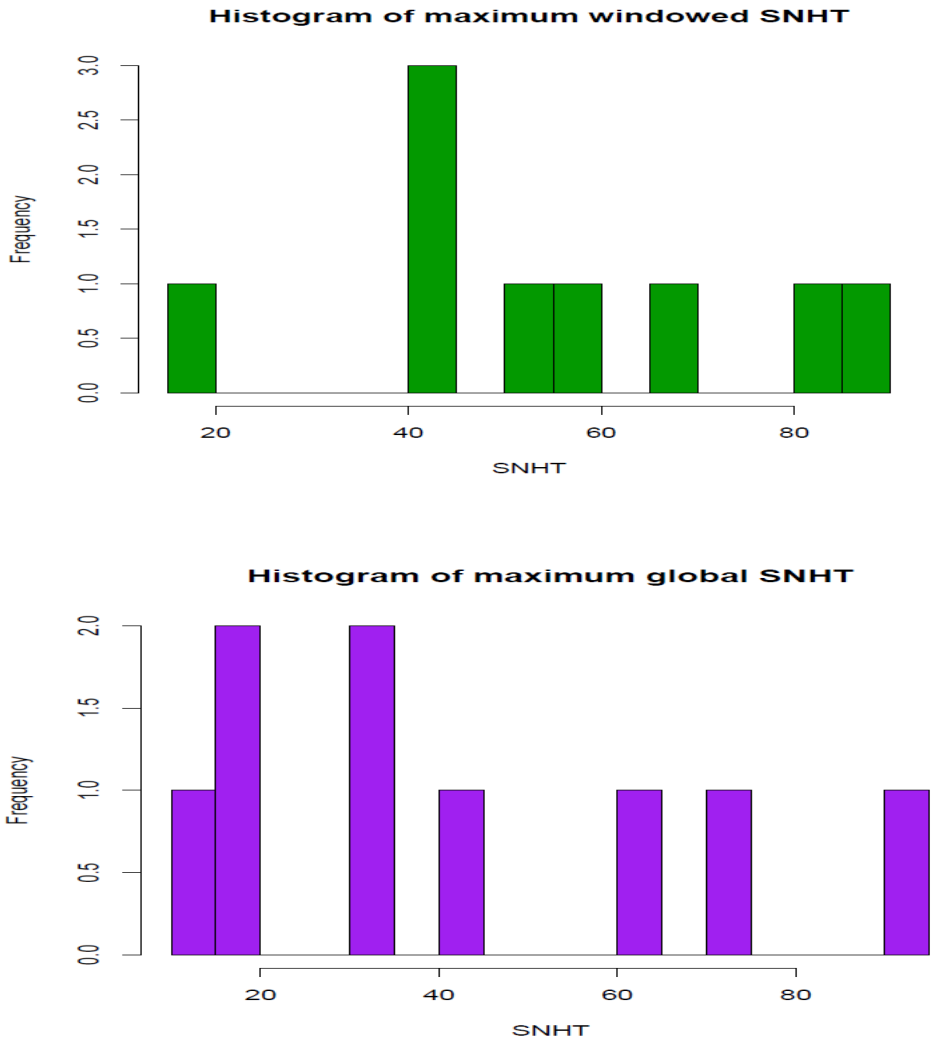


Figura 2.10: Histogramas SNHT máximos para ventanas superpuestas y para toda la serie de datos diarios respectivamente.

Ahora, los histogramas no son muy pronunciados no se puede identificar a simple vista la distribución que sigue, esto es debido a que solo se trabaja con 9 series y por ende las barras de frecuencia están separadas en varios sitios, lo que dificulta elegir de manera subjetiva los umbrales de detección de inhomogeneidades. Para la etapa de ventanas solapadas, *snht1=60* parece justo, pero está lejos de ser clara en el histograma de SNHT aplicado en las series completas. En este caso, la inspección visual no seria una opción adecuada. Es aquí donde surge el problema que queremos resolver en este Trabajo de Integración Curricular, simplemente para la elección confiable de este umbral se va aplicar técnicas Bootstrap.

2.3.3. Conversión de mediciones diarias a mensuales

La función que se utilizó para convertir los datos diarios a mensuales, es la función *dd2m* del paquete de *Climatol*, su aplicación solo requiere de 4 parámetros: nombre de la variable, año de inicio, año final y el valor mensual (*valn=1*) y devuelve dos archivos *HR-m_2015-2015.dat* y *HR-m_2015-2017.est*, la letra m en los nombres hace referencia a que los datos son mensuales. Posteriormente se realizó un exploratorio de los datos mensuales (*homegen(HR-m,2015,2017,expl=TRUE)*), este exploratorio nos arroja una gran variedad de resultados, sin embargo los más importantes son: el gráfico de anomalías normalizadas (Figura 2.11) y los gráficos de los histogramas SNHT máximos tanto para ventanas solapadas y para las series completas (Figura 2.12). El histograma de anomalías normalizadas nos ayuda a escojer adecuadamente los umbrales para rechazar datos muy anómalos, y por ende de este gráfico podemos obtener el límite superior de tolerancia de anomalías (*dz.max*) y el límite inferior de tolerancia de anomalía (*dz.min*), por defecto *Climatol* tiene un valor de 5 desviaciones típicas; es decir, que todas las anomalías absolutas superiores a 5 desviaciones típicas son eliminadas. Ahora los histogramas máximos SNHT nos ayuda a obtener los parámetros *snht1* (SNHT para ventanas solapadas) y *snht1* (SNHT para series completas), por defecto el paquete tiene un snht de 25. Observando estos gráficos podemos tomar: *dz.min*=-3.5, *dz.max*=3, *snht1*=0, *snht2*=8, sd=2.

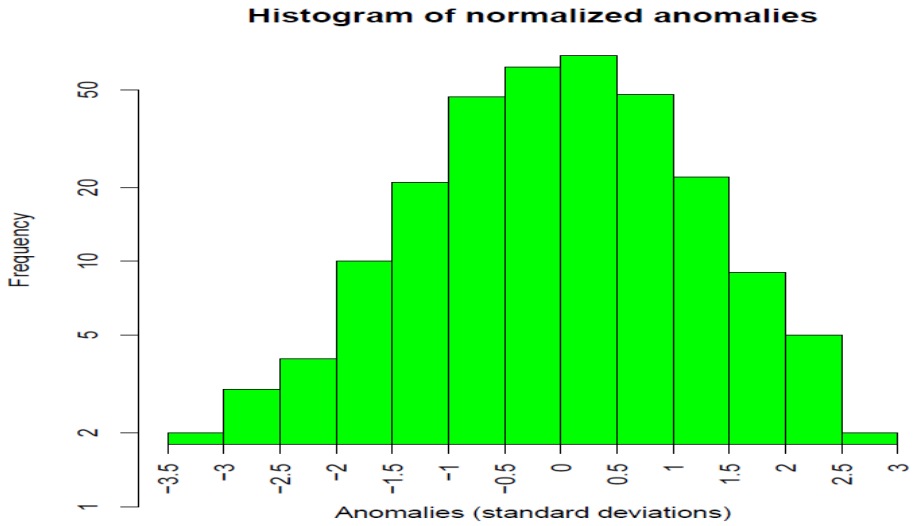


Figura 2.11: Gráfico de las anomalías normalizadas con mediciones mensuales

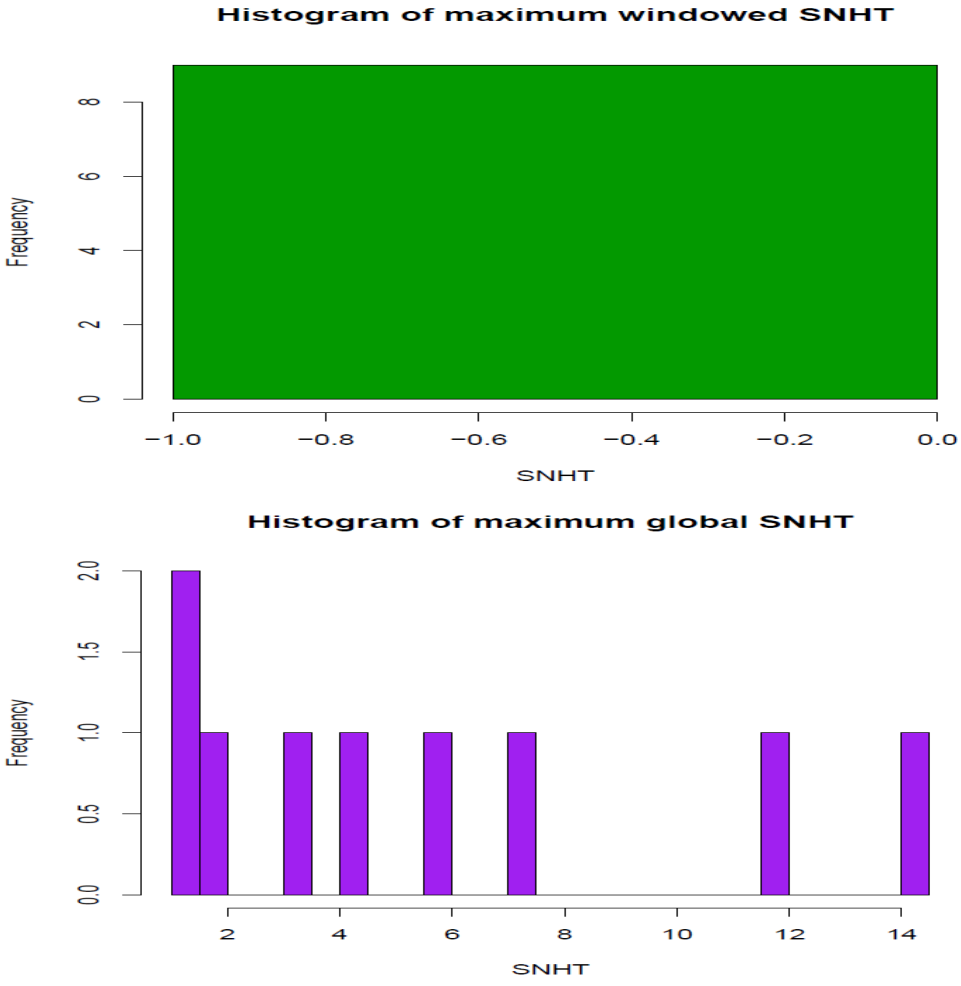


Figura 2.12: Histograma SNHT máximo para ventanas superpuestas y para toda la serie de datos mensuales respectivamente.

2.3.4. Ajustes de los datos mensuales

Una vez obtenidos los parámetros: *dz.min*, *dz.max*, *snht1* y *snht2*, se procede a realizar nuevamente un exploratorio de los datos mensuales. Para ello se utilizó la función *homogen(HR-m,2015,2017,dz.min=-3.5, dz.max=3,snht1=0,snht2=8,sd=2, vmin=0)*, la desviación típica *sd=2*, debido a que estamos hablando de la variable humedad relativa, y el último parámetro es el valor mínimo que toma esta variable, pues sabemos que tiene mediciones entre 0 y 100.

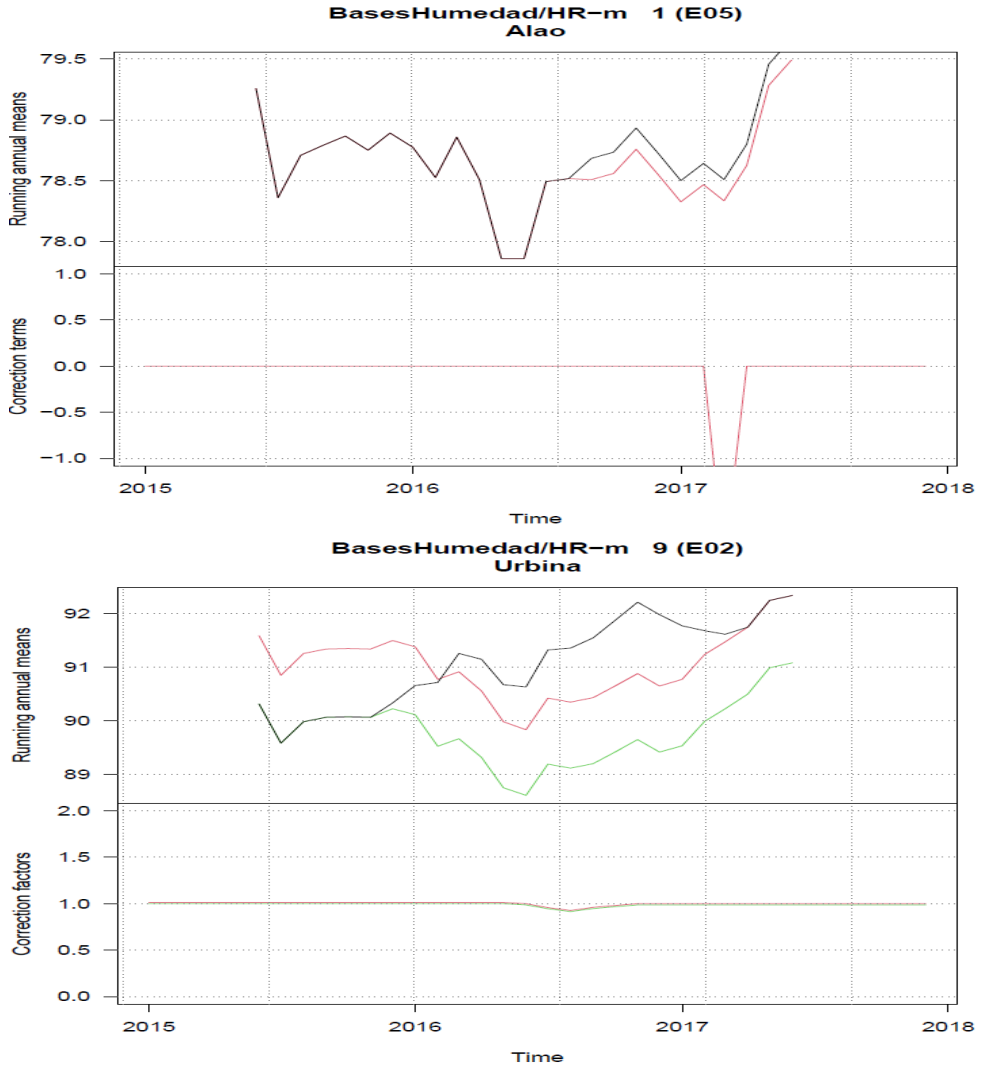


Figura 2.13: Gráfico de las series originales y construidas de la estación meteorológica Alao y Urbina con mediciones mensuales

Los gráficos más importantes que nos arroja en el exploratorio, en este caso son las series reconstruidas, a continuación se presenta algunos

ejemplos, en la Figura 2.13 se observa algunas de las series originales y reconstruidas, se evidencia que para la serie de la estación meteorológica Urbina el paquete reconstruye 2 series homogéneas una de color rojo y otra de color verde, pero para las otras estaciones solo reconstruye una serie homogéneas. Además en la Figura 2.14, se observa el histograma de anomalías normalizadas, donde se encuentran señaladas las anomalías de color rojo.

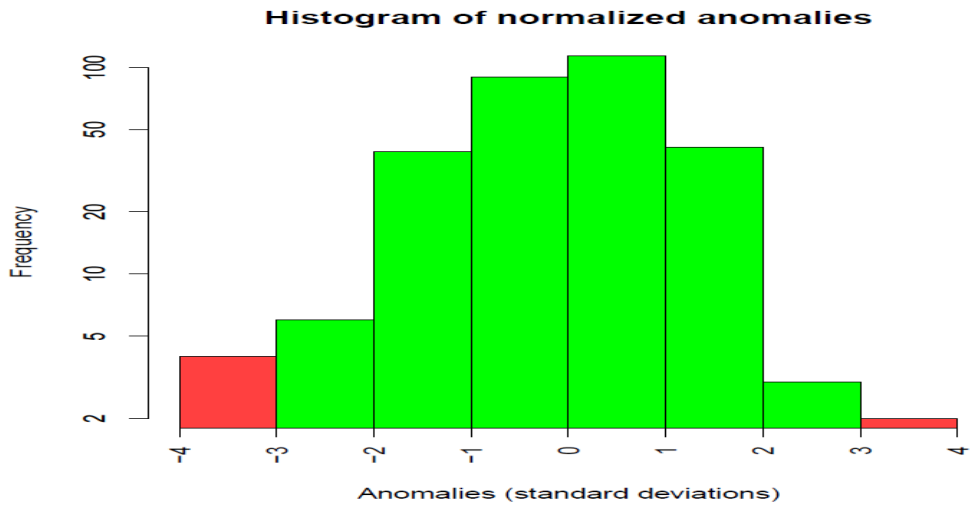


Figura 2.14: Histograma de las anomalías normalizadas de toda la serie con parámetros específicos

2.3.5. Ajustes de los datos diarios con los puntos de corte mensuales

Los datos diarios, se ajustan los puntos de corte detectados en el ajustes de datos mensuales, ahora solo debemos establece como parámetro de entrada el comando *metad=TRUE*, la función se ejecuta de la siguiente manera *homogen(HR,2015,2017,dz.min=-3.5, dz.max=3,vmin=0,metad=TRUE)*, donde los parámetros *dz.min* y *dz.max* se les obtiene del histograma de anomalías normalizadas de datos diarios como se muestra en la Figura 2.9, por otro lado, ya no se toman en cuenta los parámetros *snht1* y *snht2*, debido a que con el comando *metad=TRUE* se da la orden de que se tomen los puntos de corte detectados en el ajustes de datos mensuales los cuales se almacenan en un archivo en excel (*HR-m_2015-2017_brk*). En el cuadro 2.6, se presentan los snht que se generaron al aplicar la

homogenización a los datos mensuales. Aqui se evidencia que para la estación Urbina se detectaron 2 puntos de corte en la serie (SNHT=8.7 y SNHT=9) y es el motivo por el cual se reconstruyeron 2 series homogéneas.

Codigo	Fecha	SNHT
E02	2016-06-01	8.7
E02	2016-11-01	9

Cuadro 2.6: Valores SNHT generados al homogenizar datos mensuales *HR-m_2015-2017_brk.csv*

2.3.6. Resumen Estadístico

Para acceder a los resultados de la homogenización se hace uso de la función *load('HR_1981-2000.rda')*. Esta función genera un archivo *est.c* con los resumenes estadísticos 10 series homogéneas construidas. El resumen obtendio de las series homogéneas de cada estación se encuentra en la sección de **Resultados** en el cuadro 3.1, en este resumen, se presenta la latitud(X), longuitud(Y), altura(Z), código de la estación (Código), nombre de la estación (Nombre), el pod (promedio de datos originales), ios (número de la estación), ope (1:si la estación finaliza con un dato calculado o 0: con un dato original), snht (prueba de homogenización normal estándar) y rmse (error cuadrático medio). Ahora nos surge un problema en elegir la series homogénea más confiable para posteriores análisis.

2.3.7. Series homogenizadas

Con la función *dahstat('Ttest', 1981, 2000, stat='series')*, se crea automáticamente 2 archivos CSV: *HR_2015-2017_series.csv* este archivo contiene las mediciones todas las series homogenéas y *HR_2015-2017_flags.csv* contiene los códigos que indican si los datos son observados (0), rellenados (1, ausentes originalmente) o corregidos (2, por inhomogeneidades o por excesiva anomalía).

2.3.8. Resumen de la homogenización de las series mediante el paquete R Climatol

A continuación en la Figura 2.15, se presenta el diagrama con todos los pasos que realizaron hasta obtener las series homogéneas con el paquete R.

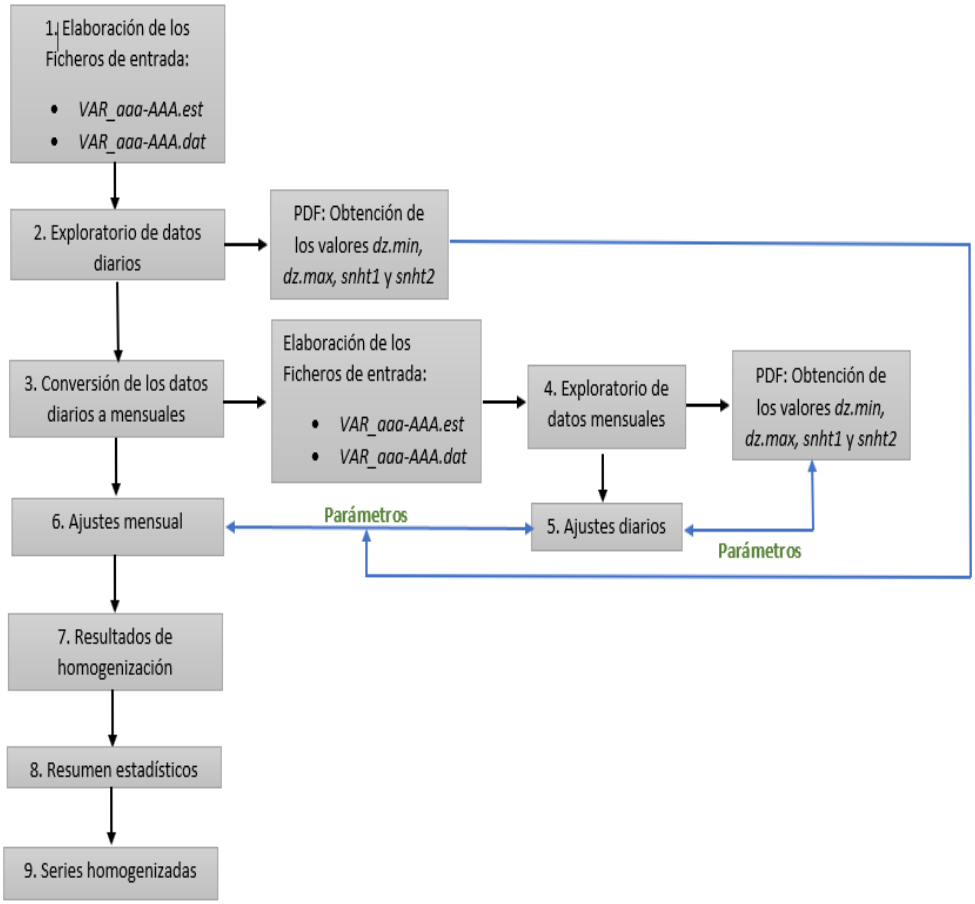


Figura 2.15: Diagrama de funcionamiento del paquete *Climatol*

2.4. Aplicación de la técnica de remuestreo Bootstrap para la obtención de umbrales de detección de inhomogeneidades

Al realizar el proceso de homogenización mediante el paquete *Climatol*, el umbral de detección de inhomogeneidades se establece de manera visual con los gráficos máximos SNHT, lo cual no parece ser apropiado,

es por esto que se puede establecer umbrales más exigentes con alta confiabilidad mediante métodos de remuestreo Bootstrap con el fin de mejor los análisis a posteriori de las series homogéneas.

2.4.1. Acerca de los datos

Este proceso requiere que las series sean homogéneas, por esta razón la de aplicabilidad este proceso se realiza a las series homogéneas que se construyeron de manera subjetiva observando los gráficos máximos SNHT. En la Figura 2.16, se muestran estas series homogéneas reconstruidas anteriormente.

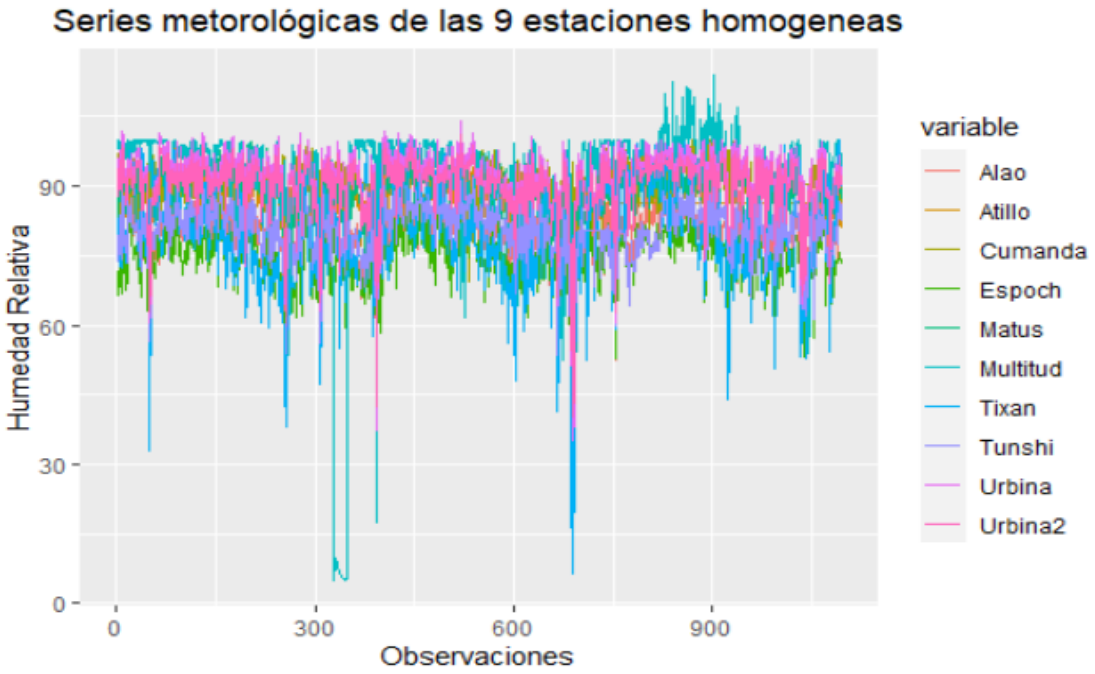


Figura 2.16: Series homogéneas reconstruidas con el paquete R *Climatol*.

Sin embargo, como se reconstruyeron dos series homogéneas para la estación Urbina. Para la elección de la serie homogénea más confiable, se usa el criterio de [8]. Este criterio establece que: *La mejor serie homogénea es la que tiene mayor número de datos promedio originales, menor valor snht y menor valor rmse*. Del cuadro 3.1, podemos elegir a la serie de la estación Urbina-2, pues posee un 47 % de datos originales y su error cuadrático medio es mucho menor que la otra serie.

2.4.2. Obtención del umbral para la detección de inhomogeneidades con el método de remuestreo MBB

Anteriormente se observó en la Figura 2.3 que todas las estaciones contienen inhomogeneidades en las series. En esta sección se presentan los pasos para obtener estos umbrales utilizando remuestras MBB. Este proceso se realiza de forma individual, debido a que las mediciones diarias de la humedad relativa de las estaciones de estudio no son de la misma naturaleza. Para la demostración solo se presenta la metodología de la estación metorológica Alao y para las restantes se procede de igual manera.

Estadístico de prueba para detectar inhomogeneidades en las series.

Para obtener el estadístico observado de la prueba de hipótesis de la homogeneidad normalidad estándar (SNHT), primero es necesario normalizar las series homogéneas de la siguiente manera:

$$Z_t = \frac{X_t - \mu_X}{\sigma_X}$$

donde X_t es la serie meteorológica homogénea con mediciones diarias de la humedad relativa en la estación Alao, μ_X y σ_X son la media y desviación típica de la serie respectivamente. El estadístico de prueba viene dado por:

$$T^s_{max} = \underset{1\leq a\leq n-1}{\text{máx}}\{T^s_a\} = \underset{1\leq a\leq n-1}{\text{máx}}\{a\bar{z}_1^2 + (n-a)\bar{z}_2^2\}$$

donde $\bar{z}_1^2$ y $\bar{z}_2^2$ son los valores medios antes y después del cambio, n es la longitud de la serie. El valor medio correspondiente de a es el punto de ruptura más probable. Así, usando la función **Estadistico** creada en el lenguaje R, se obtiene el valor observado del estadístico de prueba para la serie homogénea de la estación Alao.

$$T^{Alao}_{obs} = 30{,}84223$$

Longitud del bloque óptima para las remuestras MBB

El proceso de elección de la longitud óptima del bloque para las remuestras MBB, se realizó con ayuda de la función *hhj* del paquete R *blocklength*. Este proceso utiliza un algoritmo de validación cruzada que minimiza el error cuadrático medio (MSE) analizando varias longitudes de bloques. En esta función se debe especificar los parámetros de entrada como: la serie de estudio y la longitud de cada submuestra superpuesta: *hhj(Alao,sub_sample=10)*. En el cuadro 2.7 se encuentra el resumen de la ejecución del algoritmo y se verifica que realizó 5 iteraciones hasta encontrar una longitud óptima que minimiza el error cuadrático medio. Además, en la Figura 2.17 se corroborra que la longitud óptima del bloque para las remuestras MBB de la serie homogénea de la estación Alao es 5 ($l = 5$).

Iteración	l	*MSE*	*Iteración*	l	*MSE*
1	3	0.4977186	3	15	0.5115980
1	5	0.5016145	3	18	0.5135123
1	8	0.5114673	3	20	0.5045521
1	20	0.5267976	4	5	0.4730150
1	23	0.5192147	4	8	0.4802371
1	26	0.5102009	4	10	0.4902355
2	3	0.4780808	4	13	0.5015429
2	5	0.4711238	4	15	0.5118855
2	8	0.4794659	4	18	0.5081506
2	10	0.4909562	4	23	0.4972921
2	13	0.5004746	4	26	0.4882392
2	18	0.5077118	5	5	0.4717068
2	20	0.5022806	5	8	0.4792711
2	6	0.4879127	5	13	0.5030532
3	3	0.4754670	5	15	0.5136415
3	5	0.4778495	5	18	0.5066013

Cuadro 2.7: Resumen del algoritmo para la elección de la longitud óptima del Bloque .

Estimación de la distribución del estadístico mediante remuestreo MBB para obtener los umbrales de detección de inhomogeneidades.

Obtenida la longitud óptima del bloque ($l = 5$), debemos usar el método de remuestreo MBB para aproximar la distribución Bootstrap del esta-

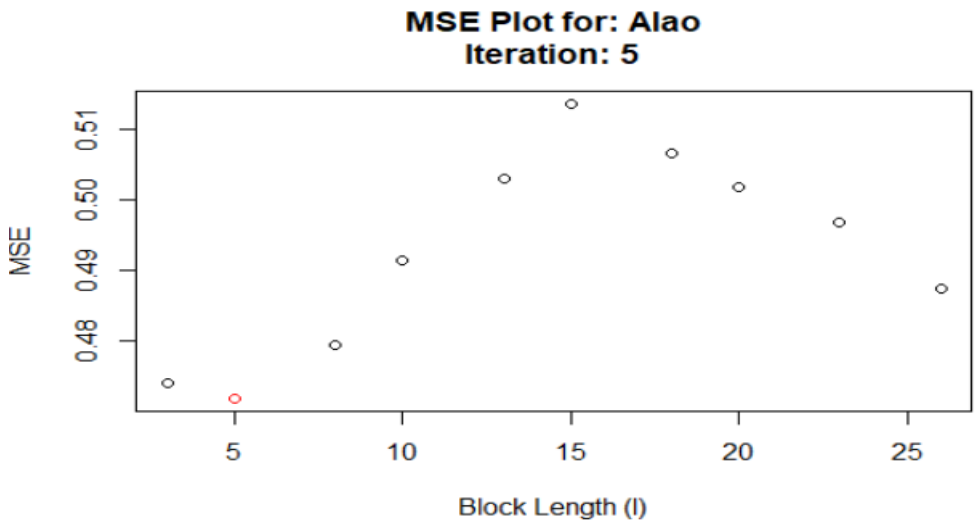


Figura 2.17: Elección de la longuitud del bloque óptima para Alao

dístico $\hat{T}_a^s$. La función *tsboot* del paquete R *boot* es eficiente para llevar a cabo este objetivo, simplemente se debe especificar de manera adecuada los parámetros de entrada, entre ellos los más importantes son: la serie de tiempo, el estadístico de prueba, $l = 5$ (longitud de cada bloque), R (número de remuestras Bootstrap) y *sim* (el método que se utiliza para obtener las remuestras Bootstrap). Cuando trabajamos con MBB se usa el método *sim="fixed"*, por que la longuitud de los bloques debe ser fija. En la Figura 2.18, se presenta la distribución estimada los estadísticos de cada una de las remuestras Bootstrap contruidas.

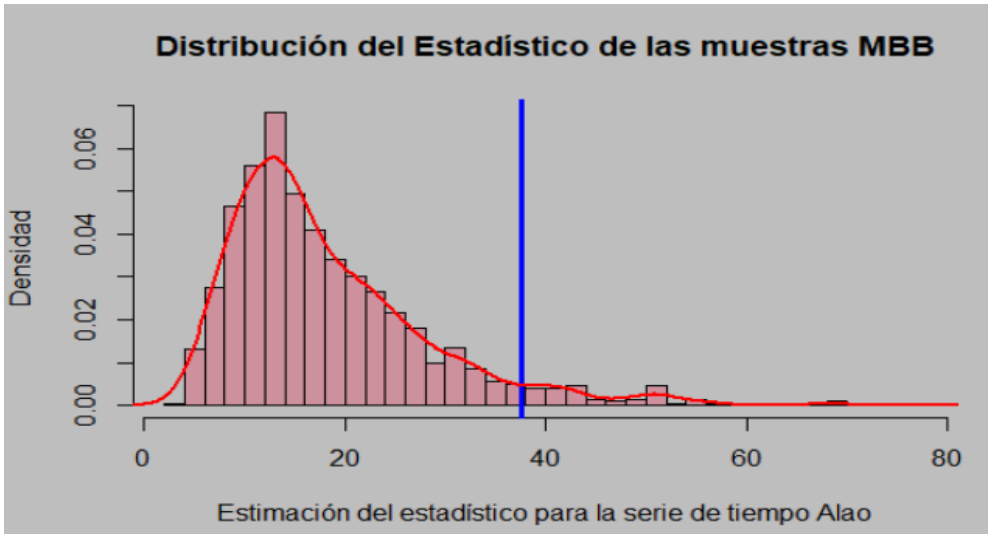


Figura 2.18: Distribución del estadístico con remuestras MBB para la serie Alao

Valor crítico de la distribución estimada del estadístico de prueba

Una vez estimada la distribución del estadístico de prueba se calcula el valor crítico de la misma, para ello se debe obtener el cuantil 95 (95 % de confianza para elegir los umbrales de detección de inhomogeneidades) de la distribución del estadístico con remuestras MBB, obteniendo:

$$T_{0,95}^{Alao^*} = 37{,}61672$$

Este valor también se pueden observar en Figura 2.18 que se representa con una linea vertical de color azul y nos ayudará a detectar de manera más eficiente las inhomogeneidades al momento de homogenizar las serie. Ahora, debemos notar que si T_{obs}^{Alao} se encuentra dentro del umbral elegido ($T_{0,95}^{Alao}$), este valor no se considera como una inhomogeneidad. Caso contrario si.

Resumen de los umbrales obtenidos para la detección de inhomogeneidades de todas las series con remuestras MBB

Usando el mismo proceso se obtiene cada uno de los umbrales de detección de inhomogeneidades de las 9 series meteorológicas y su resumen se presenta en el cuadro 2.8.

Puntos críticos usando remuestreo MBB			
Long. Bloque	Estadístico obs T_a^S	Estación	$T_{0,95}^{S*}$
5	30.84223	Alao	37.61672
10	64.16025	Atillo	82.92276
5	38.32451	Cumanda	29.10615
5	21.21852	Espoch	57.35205
5	44.49179	Matus	38.77264
9	45.93964	Multitud	183.0562
5	30.40769	Tixan	41.379
7	37.08077	Tunshi	50.3491
5	14.17178	Urbina2	42.33739

Cuadro 2.8: Valores críticos para cada estación meteorológica usando MBB

Se observa que para la estación Multitud se obtiene un alto valor SNHT para detectar estas inhomogeneidades, pero homogenizar las series con

este valor seria incorrecto por que existe gran probabilidad de que no se detecten inhomogeneidades en todas las series.

2.4.3. Homogenización de las series meteorológicas fijando como parámetro de entrada el umbral de detección de inhomogeneidades (snnh1) con remuestras MMB

Como se presento en la teoría, el problema que surge en este punto es que la función *homogen* que se utiliza para realizar la homogenización de las series, no acepta estos umbrales de forma individual para cada una de las estaciones. Entonces se propone estimar el valor global del umbral obteniendo el promedio de los umbrales calculados de cada estación dando como resultado

$$snht1_{MBB} = \frac{1}{9}\sum_{S=1}^{9} T_{0,95}^{S^*} = 62{,}54357$$

El proceso de homogenización de las series con el paquete R, se lo realiza de manera directa, únicamente fijando como parámetro de entrada también el comando $snht1 = 62{,}54357$. En este punto ya no es necesario seguir todos los pasos que se explicaron al inicio de la sección. Por otro lado, los resultados se obtuvieron únicamente con la función: *homogen("Propuesta1/HRP1",2015,2017, snht1 = 86.41562).*

Resultados de homogenización fijando el umbral de detección de inhomogeneidades (SNHT) con remuestras MBB

Los resultados de homogenización se presentan en cuadro 2.11, en este caso se detectaron más inhomogeneidades; es decir, se reconstruyeron 15 series homogéneas, en este caso para las estaciones Urbina, Atillo, Matus y Multiud se sugieren dos series homogéneas, sin embargo, para la estación Tunshi se suguiere 3 estaciones homogéneas.

Estos resultados son confiables debido a que se detectaron más inhomogenidades en la mayoria de las series. De igual manera para la elección

Serie	Código	Estación	pod	ios	ope	snht	rmse
1	E05	Alao	98	1	1	49.3	3.0
2	E09	Atillo	30	2	1	82.0	0.0
3	E06	Cumanda	86	3	1	39.5	5.0
4	E01	Espoch	71	4	1	28.2	2.6
5	E08	Matus	42	5	1	10.3	2.6
6	E07	Multitud	40	6	1	22.3	8.6
7	E04	Tixan	100	7	1	30.4	9.8
8	E03	Tunshi	24	8	1	31.9	2.3
9	E02	Urbina	37	9	1	11.9	3.4
10	E02-2	Urbina-2	46	9	0	15.5	3.5
11	E09-2	Atillo-2	62	2	0	23.4	2.5
12	E03-2	Tunshi-2	28	8	0	21.4	2.0
13	E02-2	Matus-2	55	5	0	17.3	2.6
14	E09-2	Multitud-2	43	6	0	36.3	5.3
15	E03-2	Tunshi-3	36	8	0	25.6	2.1

Cuadro 2.9: Resumen estadístico de la homogenización de las series fijando el SNHT con remuestras MBB

de las mejores series homogéneas se usa el criterio del autor [8]. Tomando esto en cuenta, en el cuadro 3.2 se encuentra el resumen de las mejores series homogéneas elegidas para cada estación meteorológica.

2.4.4. Obtención del umbral para la detección de inhomogeneidades con el método de remuestreo SB

De igual manera para este caso se obtienen los umbrales de corrección de inhomogeneidades para cada una de las estaciones de estudio, para ello vamos a utilizar los mismos parámetros antes obtenidos como: las series homogéneas y las longitudes de bloque óptima para poder construir las nuevas remuestras Botstrap Estacionarias (SB). En este proceso la longitud del bloque deja de ser fija, más bien representa a una variable distribuida geométricamente con valor esperado $l = 5$. De igual manera se toma a la estación Alao como referencia para realizar todo este proceso.

Estimación de la distribución del estadístico mediante remuestreo SB para obtener umbrales de detección de inhomogeneidades

La metodología es la misma, sin embargo como la longuitud del bloque $l \sim geom(p)$ para poder obtener las remuestras SB, basta especificar los mismos parámetros de entrada en la función *tsboot*, el único parámetro que se debe reestablecer es el método que se va a utilizar, en este caso se usa el parametro *sim="geom"*, porque ahora necesitamos remuestras Bootstra estacionarias donde la longuitud de los bloques no es fija. En la 2.19 se presenta la distribución estimada los estadísticos de cada una de las remuestras Bootstrap contruidas para la estación Alao.

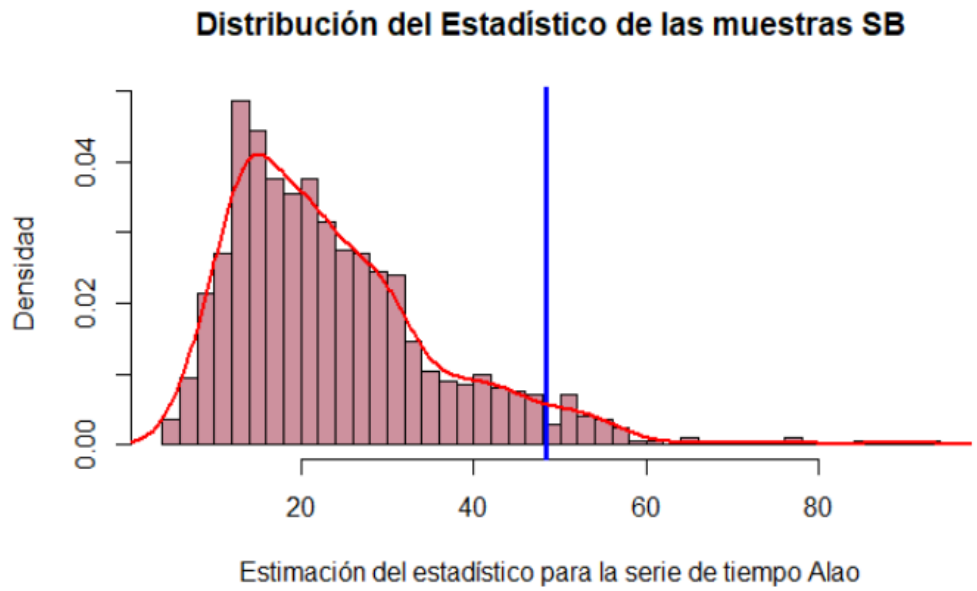


Figura 2.19: Distribución del estadístico para la serie Alao de muestras SB

Valores críticos de la distribución estimada del estadístico de prueba

Nuevamente se calcula el cuantil 95 (95 % de confianza para elegir ese umbral de detección de inhomogeneidades) de la distribución del estadístico con remuestras SB, y se obtuvo el siguiente resultado para la estación Alao

$$T_{0,95}^{Alao^*} = 48{,}41147$$

Este valor también se puede observar en el histograma de la distribución del estadístico, el cual se representa con una linea vertical de color azul y se utiliza el mismo criterio para verificar si el punto analizado representa una inhomogeneidad o no.

Resumen de los umbrales obtenidos para la detección de inhomogeneidades de todas las series usando SB

Los resultados obtenidos para los umbrales de las 9 series, se da a conocer en el cuadro 2.10.

Puntos críticos usando remuestreo SB			
Long. Bloque	Estadístico obs T_a^{Est}	Estación	$T_{0,95}^{Est^*}$
5	30.84223	Alao	48.41147
10	64.16025	Atillo	74.87484
5	38.32451	Cumanda	37.49563
5	21.21852	Espoch	46.50781
5	44.49179	Matus	48.14995
10	45.93964	Multitud	233.4455
5	30.40769	Tixan	53.4522
8	37.08077	Tunshi	67.62659
5	14.17178	Urbina2	56.62166

Cuadro 2.10: Valores críticos para cada estación meteorológica usando remuestreo SB

Al igual que el método anterior se obtuvo un alto valor SNHT para la detección de inhomogenidades.

2.4.5. Homogenización de las series meteorológicas fijando como parámetro de entrada el umbral de detección de inhomogeneidades (snnh1) con remuestras SB

Obtenemos el estimador de la media de los umbrales de detección de inhomogeneidades obtenidas anteriormente.

$$snht1_{SB} = \frac{1}{9}\sum_{S=1}^{9} T_{0,95}^{S^*} = 74,06174$$

El proceso de homogenización, se lo realiza de manera directa fijando como parámetro de entrada el comando $snht1 = 77,90925$.

Resultados de homogenización fijando como parámetro el umbral de detección de inhomogeneidades (SNHT) con remuestras SB

Los resultados de homogenización se presentan en cuadro 2.11, en este caso, se detectaron menos inhomogeneidades a diferencia del anterior metodo, es por esta razón que el total de series homogéneas construidas son 12, en este caso para las estaciones Urbina, Atillo y Tunshi se sugieren dos series homogéneas.

Serie	Código	Estación	pod	ios	ope	snht	rmse
1	E05	Alao	98	1	1	15.2	2.9
2	E09	Atillo	30	2	1	80.0	0.0
3	E06	Cumanda	86	3	1	45.8	4.9
4	E01	Espoch	71	4	1	24.1	2.5
5	E08	Matus	98	5	1	61.4	2.7
6	E07	Multitud	84	6	1	65.3	8.0
7	E04	Tixan	99	7	1	28.7	9.2
8	E03	Tunshi	24	8	1	42.6	2.4
9	E02	Urbina	51	9	1	55.1	3.6
10	E02-2	Urbina-2	46	9	0	15.6	3.2
11	E09-2	Atillo-2	62	2	0	23.4	2.5
12	E03-2	Tunshi2	65	8	0	43.2	2.2

Cuadro 2.11: Resumen estadístico de la homogenización de las series fijando el SNHT con remuestras SB

De manera similar se usa el criterio del autor [8], para la elección de la mejor serie homogénea. En la sección de **Resultados** se presentan el resumen de las mejores series homogéneas de cada estación, ver cuadro 3.3.

Capítulo 3

Resultados, conclusiones y recomendaciones

3.1. Resultados

Ahora, dado que tenemos tres métodos para la homogenización de las series y como el procedimiento es extenso, presentamos un resumen de los resultados obtenidos de las 9 estaciones meteorológicas.

Resumen de la homogenización mediante el uso del paquete R Climatol

Usando el primer método de homogenización de series con el umbral de detección de inhomogeneidades establecido visualizando los gráficos de los histogramas máximos SNHT se detectaron 2 inhomogenidades en la estación Urbina. Dada esta circunstancia la mejor serie homogénea es la estación Urbina-2 pues es la que posee mayor porcentace de datos originales, menor valor SNHT y menor valor del error cuadrático medio como lo menciona [8]. Notemos que el porcentaje de los datos originales que se utilizan para la reconstrucción de las series no necesariamente debe ser el 100 % al homogenizar las series por que es evidente que no se estan detectando inhomogeneidades debido a que el valor del SNHT fijado es demasiado grande y no es el apropiado para ciertas estaciones. Dada esta explicación las mejores series usando el primer método para todas las estaciones se presentan en el cuadro 3.1 con su respectivo resumen

estadístico.

Serie	*Código*	*Estación*	*pod*	*ios*	*ope*	*snht*	*rmse*
1	E05	Alao	99	1	1	31.2	3.1
2	E09	Atillo	98	2	1	43.0	3.2
3	E06	Cumanda	86	3	1	35.1	5.0
4	E01	Espoch	72	4	1	23.7	2.7
5	E08	Matus	99	5	1	14.8	2.9
6	E07	Multitud	88	6	1	22.3	17.2
7	E04	Tixan	99	7	1	59.1	10.0
8	E03	Tunshi	99	8	1	37.6	2.7
9	E02-2	Urbina-2	47	9	0	16.6	3.7

Cuadro 3.1: Resumen estadístico de la homogenización de las series usando el paquete R Climatol

Resumen de la homogenización de las series fijando el umbral de detección de inhomogeneidades SNHT con remuestras MBB

La metodología es la misma, sin embargo, con este método de deben establecer el umbral de detección de inhomogeneidades usando técnicas de remuestro MBB. El proceso se realiza de forma individual para cada serie, de tal manera que el umbral para todas las estaciones se estima mediante la media de todos los umbrales obtenidos de cada serie. Aquí se detectaron 15 inhomogeneidades en el conjunto de datos, por esta razón se reconstruyen 15 series homogéneas de las cuales las mejores se presentan a continuación en el cuadro 3.2.

Serie	**Código**	**Estación**	**pod**	**ios**	**ope**	**snht**	**rmse**
1	E05	Alao	98	1	1	49.3	3.0
2	E09-2	Atillo-2	62	2	0	23.4	2.5
3	E06	Cumanda	86	3	1	39.5	5.0
4	E01	Espoch	71	4	1	28.2	2.6
5	E08-2	Matus-2	55	5	0	17.3	2.6
6	E07-2	Multitud-2	43	6	0	36.3	5.3
7	E04	Tixan	100	7	1	30.4	9.8
8	E03-2	Tunshi-2	28	8	0	21.4	2.0
9	E02-2	Urbina-2	37	9	1	11.9	3.4

Cuadro 3.2: Resumen estadístico de las series homogéneas con SNHT fijo con remuestras MBB

Resumen de la homogenización de las series fijando el umbral de detección de heterogeneidades SNHT con remuestras SB

Se usa la misma metodología que el método 2, lo único que cambia es que los umbrales de detección de innhomogeneidades se ubtuvieron con remuestreo SB. Con este método se detectaron 12 inhomogeneidades y en el cuadro 3.3 se presenta el resumen de las 9 mejores series homogeneas.

Serie	Código	Estación	pod	ios	ope	snht	rmse
1	E05	Alao	98	1	1	15.2	2.9
2	E09-2	Atillo-2	62	2	0	23.4	2.5
3	E06	Cumanda	86	3	1	45.8	4.9
4	E01	Espoch	71	4	1	24.1	2.5
5	E08	Matus	98	5	1	61.4	2.7
6	E07	Multitud	84	6	1	65.3	8.0
7	E04	Tixan	99	7	1	28.7	9.2
8	E03-2	Tunshi-2	65	8	0	43.2	2.2
9	E02-2	Urbina-2	46	9	0	15.6	3.2

Cuadro 3.3: Resumen estadístico de las series homogéneas con SNHT fijo usando remuestras SB

Elección del mejor método

De los 3 métodos establecidos en el Trabajo de Integración Curricular, podemos elegir nuevamente la mejor serie homogénea usando el criterio [8]. Notemos las series en las cuales se detectaron una única inhomogenidad son: Alao, Cumanda, Espoch y Tixan. Es evidente que todas estas series homogéneas obtenidas con el tercer método son las más confiables porque minimiza el valor del SNHT y el valor del rmse, todo esto se corroborra observando los resumenes de los cuadros antes citados. Por otro lado para la estación Atillo-2, se puede elegir cualquier serie reconstruida por el método 2 y el método 3 por que presentan las mismas características. Las series de la estación Matus y Multitud se puede elegir del método 2 pues sus valores SNHT y RMSE son menores por ultimo para las estaciones Tunshi y Urbina se eligen las series homogéneas reconstruidas por el método 3 pues cumple con el criterio establecido.

Una vez elegidas las 9 mejores series homogéneas de cada estación

con los tres métodos, se obtiene que 7 series pertenecen al método 3 y 2 series pertenecen al método 2, dado esto podemos establecer que el mejor método para realizar la homogenización de las series es fijando los umbrales de corrección de inhomogenidades obtenidos con remuestreo SB pues posee la mayor cantidad de series homogéneas.

3.2. Conclusiones y recomendaciones

3.2.1. Conclusiones

El tratamiento previo que se realizó a la base de datos nos ayudó a obtener una base tratable y apta para realizar todo proceso que se plantéa en el Trabajo de Integración Curricular. La homogenización de las series meteorológicas con ayuda del paquete R *Climatol* resulta ser demasiado extensa si no conocemos el umbral de detección de inhomogeneidades y establecerlo de manera visual conlleva a resultados poco fiables; es decir, si establecemos el umbral demasiado alto, hay la posibilidad que se detecte una única inhomogeneidad en todas las series. Sin embargo, en las gráficas que explican a los datos diarios se pueden observar gran cantidad de datos anómalos los cuales deben ser tratados. Este trataimiento de datos anómalos se realiza estableciendo de manera adecuada estos umbrales, para ello se presentan 2 métodos: obtención del umbral de detección de inhomogeneidades usando remuestreo MBB y obtención del umbral de detección de inhomogeneidades usando remuestreo SB. La otención de estos umbrales tienen una metodología similar, pero en el método que utilizan para la reconstrucción de remuestras Bootstrap la longitud del bloque l es fija para el primer método y para el segundo método la longuitud del bloque $l \sim geom(p)$. La elección de la longitud del bloque óptima ayudó a mejorar la velocidad de ejecución del código al momento de estimar la distribución del estadístico de prueba mediante remuestreo y obtener el umbral de corrección de inhomogeneidades para todas las series no resultó difícil.

Al fijar este umbral en el proceso de homogenización se detectan algunas inhomogeneidades en las series. Lo bueno es que podemos elegir las mejores series homogéneas para mejorar los análisis a posteriori que se

les desea realizar a las series. Por otro lado, una buena serie homogénea no es la que se reconstruye con el 100 % de los datos originales si no más bien son las que se reconstruyen con ciertas cantidades de datos basta que el valor del SNHT y del RMSE de reduzca.

Al homogenizar las series fijando el umbral de detección de inhomogeneidades con remuestras Bootstrap al 95 % de confianza, se detectan inhomogeneidades que no fueron tomadas en cuenta al momento de homogenizar las series solo viendo gráficos máximos de SNHT. Del resumen de las series homogéneas obtendidas en la sección de Resultados, podemos concluir que la homogenización de las series estableciendo el umbral con remuestras SB es la mejor elección pues 7 de las 9 series homogéneas obtenidas en este método presenta buenos resultados en lo que hace referencia al criterio establecido; es decir, son las que tienen menor valor SNHT y menor valor RMSE. Para finalizar, se puede asegurar que para análisis futuros se puede homogenizar cualquier variable meteorológica usando el método 3 descrito en este proyecto.

3.2.2. Recomendaciones

Este proyecto se puede extender al proceso de homogenización de series individuales usando el paquete R *Climatol*, esta propuesta se plantea para investigaciones futuras. Para llevar a cabo lo propuesto se debe transformar la serie de la estación que se desea estudiar en varias series con información de mediciones diarias en periodos de un año. Todas estas nuevas series son de la misma naturaleza pues representan a una misma variable. Para el procesos de homogenización de todas estas series mediante el paquete R *Climatol* se debe establecer el umbral de detección de inhomogeneidades obtenido con remuestras Bootstrap para la estación que se esta estudiando. Sin embargo va a surgir un problema debido a que existen años bisiestos y al momento de unir las nuevas bases no van a tener la misma cantidad de datos, una solución para este problema es eliminar este dato, esto queda a decisión del autor.

Capítulo A

Código R de la homogenización de series meteorológicas mediante el paquete Climatol

```
#########################################################################
############ Homogenizacion mediante el paquete R Climatol ############
#########################################################################

# Funcion ipak: Instala y carga varios paquetes R.
ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)}

# Paquetes utilizados
packages <- c("openxlsx", "tidyverse", "climatol", "writexl","readxl",
    "visdat","naniar")
ipak(packages)

# Lectura de la bases de datos consolidada para la Humedad Relativa y
    sus respectivas estaciones
datos <- as.matrix(read.xlsx("BasesHumedad/Humedad.xlsx",na.strings = "
    NA",rowNames = F))
estaciones <- read.xlsx("BasesHumedad/Coordenadas.xlsx")

# Creacion de los ficheros de entrada de la humedad relativa y sus
    respectivas estaciones
write(datos,"BasesHumedad/HR_2015-2017.dat")
```

```
write.table(estaciones, "BasesHumedad/HR_2015-2017.est" ,row.names=
    FALSE, col.names=FALSE)

# Analisis exploratorio de datos diario para la variable HR
homogen('BasesHumedad/HR',2015,2017,expl=T)

# Conversion de datos diarios a mensuales
dd2m(varcli="BasesHumedad/HR",2015,2017)

# Exploratorio de los datos mensules para la variable HR
homogen("BasesHumedad/HR-m",2015,2017,expl = TRUE)

# Exploratorio de los datos mensuales con valores obtenidos de las
    anomalias normalizadas y de los SNHT maximos para ventanas
    solapadas y para series completas.
homogen("BasesHumedad/HR-m",2015,2017,dz.min = -3.5, dz.max = 3
        , snht=0, snht2 = 8,std=2, vmin = 0)

# Ajustes de los datos diarios a partir de los mensuales
homogen("BasesHumedad/HR",2015,2017,dz.min = -6.5, dz.max = 3.5
        , vmin = 0, metad=TRUE)

# Resumen estadistico
load("BasesHumedad/HR_2015-2017.rda")
View(est.c)

# Series homogenizadas
dahstat("BasesHumedad/HR", 2015, 2017, stat="series")
```

Capítulo B

Homogenización de las series fijando el umbral de detección de inhomogeneidades usando remuestras MBB y SB

```
#########################################################################
### Homo. mediante Climatol con el SNHT calculado calculado con MBB ###
#########################################################################

# Funcion ipak: Instala y carga varios paquetes R.
ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,"Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)}

# Paquetes utilizados
packages <- c("blocklength","openxlsx","readr", "tidyverse", "climatol"
    , "writexl","readxl", "visdat","boot", "reshape")
ipak(packages)


# Lectura de la base de datos con series homogenizadas
HR_2015_2017_series <- read_csv("C:/Users/Hp/Desktop/Tratamientos␣de␣R/
    R/Bootstrap␣estad stico/SerieHomogeneas/HR_2015-2017_series.csv")
HR1 <- rename(HR_2015_2017_series, c(E05="Alao",E09="Atillo",E06="
    Cumanda",E01="Espoch", E08="Matus",E07="Multitud",E04="Tixan",E03="
```

```
    Tunshi", E02="Urbina",`E02-2`="Urbina2"))

# Grafico de las series homogeneas
Numero <- c(1:1095)
HR <- select(HR1,-Date)
df <- cbind(Numero,HR)
df <- melt(df,id.vars = "Numero")
ggplot(df, aes(x = Numero, y = value, color = variable)) + geom_line()+
    labs(x="Observaciones",y="Humedad␣Relativa") + ggtitle("Series␣
    metorol gicas␣de␣las␣9␣estaciones␣homogeneas") + theme(plot.title
    = element_text(hjust = 0.5))

# Creaci n del estad stico de prueba
Estadistico <- function(base){
  n <- length(base)
  A <- vector()
  B <- vector()
  T_a <- vector()
  for (i in 1:(n-1)){
    A[i] <- i*mean(base[1:i])*mean(base[1:i])
    B[i] <- (n-i)*mean(base[(i+1):n])*mean(base[(i+1):n])
    T_a[i] <- A[i]+B[i]
  }
  return(max(T_a))
}

##########################################################################
#################### Bootstrap por bloques moviles ######################
##########################################################################



############################ Estacion Alao ##############################

Alao <- select(HR1,Alao)
Alao <- ts(Alao)

# Normalizaci n de la serie
Alao <- (Alao-mean(Alao))/sd(Alao)

# Longuitud optima del bloque serie Alao
lbAlao <-  hhj(Alao, sub_sample = 10)

# Muestras MBB
```

```
MBBAlao <- tsboot(Alao, Estadistico, R = 1000, l = 5, sim ="fixed" )
par(bg = "gray")
hist(MBBAlao$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del␣
    estad stico␣para␣la␣serie␣de␣tiempo␣Alao",main ="Distribuci n␣del
    ␣Estad stico␣de␣las␣muestras␣MBB",ylab = "Densidad")
abline(v=quantile(MBBAlao$t, c(0.95)), col= "blue",lwd=3)
dz <- density(MBBAlao$t)
lines(dz, col = "red", lwd = 2)
TAlao1 <- quantile(MBBAlao$t,c(0.95))
TAlao1
print(MBBAlao)

############################ Estacion Atillo ############################

Atillo <- select(HR1,Atillo)
Atillo <- ts(Atillo)

# Normalizaci n de la serie
Atillo <- (Atillo-mean(Atillo))/sd(Atillo)

# Longuitud optima del bloque
lbAtillo <- hhj(Atillo,sub_sample = 10)

# Muestras MBB
MBBAtillo <- tsboot(Atillo, Estadistico, R = 1000, l = 10, sim ="fixed"
    )
par(bg = "gray")
hist(MBBAtillo$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣Atillo",main ="Distribuci n␣del␣
    Estad stico␣de␣las␣muestras␣MBB",ylab = "Densidad")
abline(v=quantile(MBBAtillo$t, c(0.95)), col= "blue")
dz1 <- density(MBBAtillo$t)
lines(dz1, col = "red", lwd = 1)
TAtillo1 <- quantile(MBBAtillo$t,c(0.95))
TAtillo1
print(MBBAtillo)

########################### Estacion Cumanda ############################

Cumanda <- select(HR1,Cumanda)
Cumanda <- ts(Cumanda)

# Normalizaci n de la serie
```

```
Cumanda <- (Cumanda-mean(Cumanda))/sd(Cumanda)

# Longuitud optima del bloque
lbCumanda <- hhj(Cumanda,sub_sample = 10)

# Muestras MBB
MBBCumanda<- tsboot(Cumanda, Estadistico, R = 1000, l = 5, sim ="fixed"
    )
hist(MBBCumanda$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣de␣tiempo␣Cumanda",main ="
    Distribuci n␣del␣Estad stico␣de␣las␣muestras␣MBB",ylab = "
    Densidad")
abline(v=quantile(MBBCumanda$t, c(0.95)), col= "blue",lwd=3)
dz2 <- density(MBBCumanda$t)
lines(dz2, col = "red", lwd = 2)
TCumanda1 <- quantile(MBBCumanda$t,c(0.95))
TCumanda1
print(MBBCumanda)



############################ Estacion Espoch ############################

Espoch <- select(HR1,Espoch)
Espoch <- ts(Espoch)

# Normalizaci n de la serie
Espoch <- (Espoch-mean(Espoch))/sd(Espoch)



# Longuitud optima del bloque
lbEspoch <- hhj(Espoch, sub_sample = 10)

# Muestras MBB
MBBEspoch <- tsboot(Espoch, Estadistico, R = 1000, l = 5, sim ="fixed"
    )
hist(MBBEspoch$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣de␣tiempo␣Espoch",main ="
    Distribuci n␣del␣Estad stico␣de␣las␣muestras␣MBB",ylab = "
    Densidad")
abline(v=quantile(MBBEspoch$t, c(0.95)), col= "blue",lwd=3)
dz3 <- density(MBBEspoch$t)
lines(dz3, col = "red", lwd = 2)
TEspoch1 <- quantile(MBBEspoch$t,c(0.95))
```

```
TEspoch1
print(MBBEspoch)

########################## Estacion Matus ############################

Matus <- select(HR1,Matus)
Matus <- ts(Matus)

# Normalizaci n de la serie
Matus <- (Matus-mean(Matus))/sd(Matus)


# Longuitud optima del bloque
lbMatus <- hhj(Matus,sub_sample = 10)

# Muestras MBB
MBBMatus <- tsboot(Matus, Estadistico, R = 1000, l = 5, sim ="fixed" )
hist(MBBMatus$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del
    ␣estad stico␣para␣la␣serie␣de␣tiempo␣Matus",main ="Distribuci n␣
    del␣Estad stico␣de␣las␣muestras␣MBB",ylab = "Densidad")
abline(v=quantile(MBBMatus$t, c(0.95)), col= "blue",lwd=3)
dz4 <- density(MBBMatus$t)
lines(dz4, col = "red", lwd = 2)
TMatus1 <- quantile(MBBMatus$t,c(0.95))
TMatus1
print(MBBMatus)

########################## Estacion Multitud ############################

Multitud <- select(HR1,Multitud)
Multitud <- ts(Multitud)

# Normalizaci n de la serie
Multitud <- (Multitud-mean(Multitud))/sd(Multitud)


# Longuitud optima del bloque
lbMultitud<- hhj(Multitud,sub_sample = 10)

# Muestras MBB
MBBMultitud <- tsboot(Multitud, Estadistico, R = 1000, l = 10, sim ="
    fixed" )
```

```
hist(MBBMultitud$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣de␣tiempo␣Multitud",main ="
    Distribuci n␣del␣Estad stico␣de␣las␣muestras␣MBB",ylab = "
    Densidad")
abline(v=quantile(MBBMultitud$t, c(0.95)), col= "blue",lwd=3)
dz5 <- density(MBBMultitud$t)
lines(dz5, col = "red", lwd = 1)
TMultitud1 <- quantile(MBBMultitud$t,c(0.95))
TMultitud1
print(MBBMultitud)

############################ Estacion Tixan ############################

Tixan <- select(HR1,Tixan)
Tixan <- ts(Tixan)

# Normalizaci n de la serie
Tixan <- (Tixan-mean(Tixan))/sd(Tixan)

# Longuitud optima del bloque
lbTixan <- hhj(Tixan,sub_sample = 10)

# Muestras MBB
MBBTixan <- tsboot(Tixan, Estadistico, R = 1000, l = 5, sim ="fixed" )
hist(MBBTixan$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del
    ␣estad stico␣para␣la␣serie␣de␣tiempo␣Tixan",main ="Distribuci n␣
    del␣Estad stico␣de␣las␣muestras␣MBB",ylab = "Densidad")
abline(v=quantile(MBBTixan$t, c(0.95)), col= "blue",lwd=3)
dz6 <- density(MBBTixan$t)
lines(dz6, col = "red", lwd = 2)
TTixan1 <- quantile(MBBTixan$t,c(0.95))
TTixan1
print(MBBTixan)

############################ Estacion Tunshi ############################

unshi <- select(HR1,Tunshi)
Tunshi <- ts(Tunshi)

# Normalizaci n de la serie
Tunshi <- (Tunshi-mean(Tunshi))/sd(Tunshi)

# Longuitud optima del bloque
```

```
lbTunshi<- hhj(Tunshi, sub_sample = 10)

# Muestras MBB
MBBTunshi <- tsboot(Tunshi, Estadistico, R = 1000, l = 7, sim ="fixed"
    )
hist(MBBTunshi$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣Tunshi",main ="Distribuci n␣del␣
    Estad stico␣de␣las␣muestras␣MBB",ylab = "Densidad")
abline(v=quantile(MBBTunshi$t, c(0.95)), col= "blue",lwd=3)
dz7 <- density(MBBTunshi$t)
lines(dz7, col = "red", lwd = 2)
TTunshi1 <- quantile(MBBTunshi$t,c(0.95))
TTunshi1
print(MBBTunshi)

############################ Estacion Urbina ############################

Urbina2 <- select(HR1,Urbina2)
Urbina2 <- ts(Urbina2)

# Normalizaci n de la serie
Urbina2 <- (Urbina2-mean(Urbina2))/sd(Urbina2)

# Longuitud optima del bloque
lbUrbina2<- hhj(Urbina2, sub_sample = 10)

# Muestras MBB
MBBUrbina2 <- tsboot(Urbina2, Estadistico, R = 1000, l = 5, sim ="fixed
    " )
hist(MBBUrbina2$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣de␣tiempo␣Urbina",main ="
    Distribuci n␣del␣Estad stico␣de␣las␣muestras␣MBB",ylab = "
    Densidad")
abline(v=quantile(MBBUrbina2$t, c(0.95)), col= "blue",lwd=3)
dz6 <- density(MBBUrbina2$t)
lines(dz7, col = "red", lwd = 2)
TUrbina1 <- quantile(MBBUrbina2$t,c(0.95))
TUrbina1
print(MBBUrbina2)


##### Hom. fijando el SNHT obtenido del promedio de las estaciones ####
```

```
# Promedio del umbral de correccion de inhomogeneidades de todas las
    estaciones
SNHT1 <- mean(c(TAlao1,TAtillo1,TCumanda1,TEspoch1,TMatus1,TMultitud1,
    TTixan1,TTunshi1,TUrbina1))
SNHT1

# Lectura y transformaci n de las Base de datos
Datos <- read_xlsx("Propuesta1/Humedad.xlsx")
Datos$Espoch <- as.numeric(Datos$Espoch)
Datos$Cumanda <- as.numeric(Datos$Cumanda)
Datos$Multitud <- as.numeric(Datos$Multitud)
Datos$Tunshi <- as.numeric(Datos$Tunshi)
Base <- as.matrix(Datos)
estaciones <- read_xlsx("Propuesta1/Coordenadas.xlsx")

# Ficheros de entrada
write(Base,'Propuesta1/HRP1_2015-2017.dat')
write.table(estaciones, 'Propuesta1/HRP1_2015-2017.est', row.names=
    FALSE, col.names=FALSE)

# A lisis exploratorio
homogen("Propuesta1/HRP1",2015,2017,expl = TRUE)

# Homogenizacion de las series fijando el umbral e correccion de
    inhomogeneidades SNHT
homogen("Propuesta1/HRP1",2015,2017, dz.min=-3.5, dz.max = 3.5, snht1 =
     SNHT,vmin = 0,vmax = 100)

# Resumen estad stico
load('Propuesta1/HRP1_2015-2017.rda')
View(est.c)

# Series homogeneizadas
dahstat('Series//HR', 2015, 2017, stat='series')


##############################################################################
######################## Bootstrap Estacionario ##############################
##############################################################################

############################## Estacion Alao #################################

# Muestras SB
```

```r
SBAlao <- tsboot(Alao, Estadistico, R = 1000, l = 5, sim ="geom" )
par(bg = "gray")
hist(SBAlao$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del␣
    estad stico␣para␣la␣serie␣de␣tiempo␣Alao",main ="Distribuci n␣del
    ␣Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad")
abline(v=quantile(SBAlao$t, c(0.95)), col= "blue",lwd=3)
dz <- density(SBAlao$t)
lines(dz, col = "red", lwd = 2)
TAlao <- quantile(SBAlao$t,c(0.95))
print(SBAlao)

############################ Estacion Atillo ############################

# Muestras SB
SBAtillo <- tsboot(Atillo, Estadistico, R = 1000, l = 10, sim ="geom" )
par(bg = "gray")
hist(SBAtillo$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del
    ␣estad stico␣para␣la␣serie␣Atillo",main ="Distribuci n␣del␣
    Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad")
abline(v=quantile(SBAtillo$t, c(0.95)), col= "blue",lwd=3)
dz1 <- density(SBAtillo$t)
lines(dz1, col = "red", lwd = 2)
TAtillo <- quantile(SBAtillo$t,c(0.95))
print(SBAtillo)

############################ Estacion Cumanda ############################

# Muestras SB
SBCumanda<- tsboot(Cumanda, Estadistico, R = 1000, l = 5, sim ="geom" )
hist(SBCumanda$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣de␣tiempo␣Cumanda",main ="
    Distribuci n␣del␣Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad
    ")
abline(v=quantile(SBCumanda$t, c(0.95)), col= "blue",lwd=3)
dz2 <- density(SBCumanda$t)
lines(dz2, col = "red", lwd = 2)
TCumanda <- quantile(SBCumanda$t,c(0.95))
print(SBCumanda)

############################ Estacion Espoch ############################

# Muestras SB
SBEspoch <- tsboot(Espoch, Estadistico, R = 1000, l = 5, sim ="geom" )
```

```
hist(SBEspoch$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del
    ␣estad stico␣para␣la␣serie␣de␣tiempo␣Espoch",main ="Distribuci n␣
    del␣Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad")
abline(v=quantile(SBEspoch$t, c(0.95)), col= "blue",lwd=3)
dz3 <- density(SBEspoch$t)
lines(dz3, col = "red", lwd = 2)
TEspoch <- quantile(SBEspoch$t,c(0.95))
print(SBEspoch)

############################ Estacion Matus ############################

# Muestras SB
SBMatus <- tsboot(Matus, Estadistico, R = 1000, l = 5, sim ="geom" )
hist(SBMatus$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del␣
    estad stico␣para␣la␣serie␣de␣tiempo␣Matus",main ="Distribuci n␣
    del␣Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad")
abline(v=quantile(SBMatus$t, c(0.95)), col= "blue",lwd=3)
dz4 <- density(SBMatus$t)
lines(dz4, col = "red", lwd = 2)
TMatus <- quantile(SBMatus$t,c(0.95))
print(SBMatus)

########################### Estacion Multitud ###########################

# Muestras SB
SBMultitud <- tsboot(Multitud, Estadistico, R = 1000, l = 10, sim ="
    geom" )
hist(SBMultitud$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣de␣tiempo␣Multitud",main ="
    Distribuci n␣del␣Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad
    ")
abline(v=quantile(SBMultitud$t, c(0.95)), col= "blue",lwd=3)
dz5 <- density(SBMultitud$t)
lines(dz5, col = "red", lwd = 1)
TMultitud <- quantile(SBMultitud$t,c(0.95))
print(SBMultitud)

############################ Estacion Tixan ############################

# Muestras SB
SBTixan <- tsboot(Tixan, Estadistico, R = 1000, l = 5, sim ="geom" )
hist(SBTixan$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del␣
    estad stico␣para␣la␣serie␣de␣tiempo␣Tixan",main ="Distribuci n␣
```

```
    del␣Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad")
abline(v=quantile(SBTixan$t, c(0.95)), col= "blue",lwd=3)
dz6 <- density(SBTixan$t)
lines(dz6, col = "red", lwd = 2)
TTixan <- quantile(SBTixan$t,c(0.95))
print(SBTixan)

############################ Estacion Tunshi ############################

# Muestras SB
SBTunshi <- tsboot(Tunshi, Estadistico, R = 1000, l = 8, sim ="geom" )
hist(SBTunshi$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣del
    ␣estad stico␣para␣la␣serie␣Tunshi",main ="Distribuci n␣del␣
    Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad")
abline(v=quantile(SBTunshi$t, c(0.95)), col= "blue",lwd=3)
dz7 <- density(SBTunshi$t)
lines(dz7, col = "red", lwd = 2)
TTunshi <- quantile(SBTunshi$t,c(0.95))
print(SBTunshi)

############################ Estacion Urbina ############################

# Muestras SB
SBUrbina2 <- tsboot(Urbina2, Estadistico, R = 1000, l = 8, sim ="geom"
    )
hist(SBUrbina2$t, prob=T,breaks = 50,col = "pink3",xlab="Estimaci n␣
    del␣estad stico␣para␣la␣serie␣de␣tiempo␣Urbina",main ="
    Distribuci n␣del␣Estad stico␣de␣las␣muestras␣SB",ylab = "Densidad
    ")
abline(v=quantile(SBUrbina2$t, c(0.95)), col= "blue",lwd=3)
dz6 <- density(SBUrbina2$t)
lines(dz7, col = "red", lwd = 2)
TUrbina2 <- quantile(SBUrbina2$t,c(0.95))
print(SBUrbina2)

##### Hom. fijando el SNHT obtenido del promedio de las estaciones ####

# Promedio del umbral de correcci n de inhomogeneidades de todas las
    estaciones
SNHT <- mean(c(TAlao,TAtillo,TCumanda,TMatus,TMultitud,TTixan,TTunshi,
    TUrbina2))
SNHT
```

```r
# Lectura y transformaci n de la base de datos
Datos1 <- read_xlsx("Propuesta1SB/Humedad.xlsx")
Datos1$Espoch <- as.numeric(Datos1$Espoch)
Datos1$Cumanda <- as.numeric(Datos1$Cumanda)
Datos1$Multitud <- as.numeric(Datos1$Multitud)
Datos1$Tunshi <- as.numeric(Datos1$Tunshi)
Base1 <- as.matrix(Datos1)
estaciones1 <- read_xlsx("Propuesta1SB/Coordenadas.xlsx")

# Ficheros de entrada
write(Base1,'Propuesta1SB/HRP1_2015-2017.dat')
write.table(estaciones1, 'Propuesta1SB/HRP1_2015-2017.est', row.names=
    FALSE, col.names=FALSE)

# An lisis exploratorio
homogen("Propuesta1SB/HRP1",2015,2017,expl = TRUE)

# Homogenizaci n de las series fijando el umbral de correccion de
    inhomogeneidades SNHT
homogen("Propuesta1SB/HRP1",2015,2017, dz.min=-3.5, dz.max = 3.5, snht1
     = SNHT,vmin = 0,vmax = 100)

# Resumen estadistico
load('Propuesta1/HRP1_2015-2017.rda')
View(est.c)
# Series homogeneizadas y
dahstat('Propuesta1SB/HRP1', 2015, 2017, stat='series')
```

Referencias bibliográficas

[1] Hans Alexandersson and Anders Moberg. Homogeneización de datos de temperatura suecos. parte 1: Prueba de homogeneidad para tendencias lineales. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 17(1):25–34, 1997.

[2] Andrés M Alonso, Daniel Peña, and Juan Romo Urroz. Una revisión de los métodos de remuestreo en series temporales. 2002.

[3] Fanny Bergström. Bootstrap methods in time series analysis. pages 9–12, 2018.

[4] José Roberto Bosano Cueva. Aplicación de una nueva metodología basada en métodos bootstrap para la detección de resultados atípicos, en el estudio interlaboratorio. *Universidad de las Fuezar Armadas ESPE*, pages 28–29, 2019.

[5] V. Conrad and L. W. Pollak. *Methods in Climatology*. Harvard University Press, 2013.

[6] Andrés M Alonso Fernández and Juan Jose Romo Urroz. *Técnicas de remuestreo y datos omitidos en series temporales*. PhD thesis, Universidad Carlos III de Madrid, 2001.

[7] Gabriel Gaona, Emmanuelle Quentin, and Jerko Labus. Homogeneidad y variabilidad espacial de series meteorológicas del área del proyecto “ciudad del conocimiento-yachay”. *ACI Avances en Ciencias e Ingenierías*, 5(2), 2013.

[8] Jose Guijarro. Climatol: Sofware libre para la depuración y homogenización de datos climaticos. 08 2015.

[9] José María Jansá. *Curso de climatología*. Madrid, 1996.

[10] SN Lahiri. Bootstrap methods. In *Resampling Methods for Dependent Data*, pages 17–43. Springer, 2003.

[11] Regina Y Liu, Kesar Singh, et al. Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the limits of bootstrap*, 225:225–2045, 1992.

[12] Joseph Paulhus and Max Kohler. Interpolación de registros de precipitación faltantes. 80:129–133, 1952.

[13] Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

[14] Dimitris N Politis and Halbert White. Automatic block-length selection for the dependent bootstrap. *Econometric reviews*, 23(1):53–70, 2004.

[15] Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436, 1993.

[16] Armando Vargas, Miguel Ángel Vargas, A Martín Estrada, and Rogelio González. Programación del método bootstrap. *Memorias de las Grandes Semanas Nacionales de la Matemática Facultad de Ciencias Físico Matemáticas*, page 73.

[17] Walter Quispe Vargas. *Sieve Bootstrap en series de tiempo de nubosidad en el Caribe*. PhD thesis, University of Puerto Rico, Mayaguez (Puerto Rico), 2007.