



FACULTAD DE  
CIENCIAS ECONÓMICAS  
Y DE ADMINISTRACIÓN



UNIVERSIDAD  
DE LA REPÚBLICA  
URUGUAY

UNIVERSIDAD DE LA REPÚBLICA  
FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

TRABAJO FINAL DE PASANTÍA PARA OBTENER EL TÍTULO DE  
LICENCIADO EN ESTADÍSTICA

Estudio de diferentes metodologías estadísticas para el control de calidad  
de la base de datos diarios de precipitación en Uruguay.

por

Joaquín Amiel Marcora

TUTORES: Madeleine Renom y Marco Scavino

Montevideo

URUGUAY

2012

# Estudio de diferentes metodologías estadísticas para el control de calidad de la base de datos diarios de precipitación en Uruguay.

Pasantía realizada en la Facultad de Ciencias en la Unidad de Ciencias de la Atmósfera- Instituto de Física.

Estudiante: Joaquín Amiel Marcora.

Tutores: Dra. Madeleine Renom, Dr. Marco Scavino.

Setiembre 2012

# Índice general

<b>1. Introducción.</b>	<b>3</b>
1.1. Características de las precipitaciones. . . . .	3
1.2. Proceso de medición de la precipitación. . . . .	4
1.3. Algunos antecedentes. . . . .	5
1.4. Descripción de los problemas a enfrentar. . . . .	7
<b>2. Base de datos.</b>	<b>9</b>
2.1. Descripción de la base de datos. . . . .	9
2.2. Tratamiento y depuración de coordenadas. . . . .	12
2.3. Representación geográfica de las estaciones. . . . .	12
2.3.1. Procedimiento para el trazado del mapa en R. . . . .	12
2.3.2. Representación de las estaciones pluviométricas. . . . .	12
<b>3. Control de calidad de la base de datos.</b>	<b>15</b>
3.1. Lectura de los datos. . . . .	15
3.2. Errores detectados. . . . .	17
3.3. Superposición de series en un mismo punto. . . . .	20
3.4. Características de las estaciones. . . . .	21
3.5. Primer análisis de datos faltantes. . . . .	22
3.6. Conclusiones del capítulo. . . . .	28
<b>4. Reducción de la base de datos y distancias entre las estaciones.</b>	<b>29</b>
4.1. Selección de las estaciones pluviométricas. . . . .	29
4.1.1. Por años de registro. . . . .	29
4.1.2. Por datos faltantes. . . . .	30
4.1.3. Por porcentaje de días sin lluvia sobre el total de registros. . . . .	32
4.1.4. Selección final de estaciones pluviométricas. . . . .	34
4.2. Distancias entre estaciones pluviométricas. . . . .	35
<b>5. Caso de estudio: departamento de Colonia.</b>	<b>37</b>
5.1. ¿Por qué se eligió Colonia?. . . . .	37
5.2. Descripción de las estaciones pluviométricas. . . . .	39
5.3. Datos faltantes en Colonia. . . . .	41
5.4. Estudio de correlaciones y relación con las distancias entre estaciones. . . . .	43
5.4.1. Marco teórico. . . . .	43
5.4.2. Distancias entre estaciones. . . . .	45
5.4.3. Estudio de las correlaciones. . . . .	45
5.5. Estudio de días lluviosos en el mes y lluvia mensual. . . . .	49

5.6.	Densidad estimada de precipitación diaria. . . . .	55
5.7.	Análisis de precipitación. . . . .	57
5.8.	Conclusiones del capítulo. . . . .	61
<b>6.</b>	<b>Modelo para la estimación de ocurrencia de lluvia.</b>	<b>62</b>
6.1.	Selección de la metodología. . . . .	62
6.1.1.	Primer intento para el modelado de la ocurrencia de lluvia. . . . .	62
6.1.2.	Marco teórico. . . . .	63
6.1.3.	Segundo intento para el modelado de la ocurrencia de lluvia. . . . .	65
6.2.	Desarrollo utilizando cadenas de Markov no estacionarias (Markov NE). . . . .	66
6.2.1.	Estimación del modelo para las probabilidades de lluvia. . . . .	66
6.2.2.	Simulación del modelo estimado y comparación con la serie observada. . . . .	69
6.3.	Aplicación del modelo a la estación de La Estanzuela. . . . .	72
6.3.1.	Estimación del modelo para las probabilidades de lluvia. . . . .	72
6.3.2.	Simulación del modelo estimado y comparación con la serie observada. . . . .	73
6.4.	Aplicación del modelo a la estación pluviométrica 2611. . . . .	75
6.4.1.	Estimación del modelo para las probabilidades de lluvia. . . . .	75
6.4.2.	Simulación del modelo estimado y comparación con la serie observada. . . . .	76
6.5.	Conclusiones del capítulo. . . . .	79
<b>7.</b>	<b>Conclusiones.</b>	<b>81</b>
<b>A.</b>	<b>Tablas utilizadas.</b>	<b>85</b>
A.1.	Tabla de la sección 2.1 . . . . .	85
A.2.	Tabla de la sección 5.4.3 . . . . .	86
<b>B.</b>	<b>Código utilizado.</b>	<b>87</b>
B.1.	Depurado de base de datos, análisis y confección inicial. . . . .	87
B.2.	Análisis de datos faltantes. . . . .	91
B.3.	Mapas utilizados. . . . .	94
B.4.	Análisis de estaciones pluviométricas en colonia. . . . .	98
B.5.	Modelado para medir la ocurrencia de lluvia. . . . .	105
B.5.1.	Prueba de independencia. . . . .	105
B.5.2.	Prueba de estacionariedad. . . . .	106
B.5.3.	Modelado para medir la ocurrencia de lluvia. . . . .	107
B.5.4.	Simulación del modelo de ocurrencia de lluvia. . . . .	118

# Capítulo 1

## Introducción.

El estudio de la variabilidad climática y el cambio climático, ha hecho énfasis, en los últimos años, en el análisis de los eventos climáticos extremos, debido al gran impacto que tienen en la sociedad y en la economía de los países. En nuestra región, Sudeste de Sudamérica, los principales eventos climáticos extremos serían olas de calor y frío, sequías, inundaciones, etc. Esto toma una gran importancia en el área de la agricultura, la ganadería así como también en el ámbito energético.

El estudio de los eventos climáticos extremos requiere por lo menos bases de datos diarias lo más extensas y confiables posibles, ya que estas son más valiables y además la confiabilidad en las estimaciones está relacionada con el largo de la misma. Los estudios sobre cambio climático requieren series de por lo menos 40 años.

### 1.1. Características de las precipitaciones.

La precipitación es una de las variables atmosféricas de mayor importancia. Para la *atmósfera*, precipitación indica la liberación de calor latente en la columna de aire; para el *mar*, precipitación representa una fuente de agua dulce; en la *tierra* precipitación es una de las fases del ciclo hidrológico generado a partir del balance energético; para la *ecología* representa un importante factor de control externo.

La precipitación se produce por una combinación de procesos dinámicos y termodinámicos. Los mecanismos disparadores de la precipitación son muy complejos y aún no han sido comprendidos del todo.

Una clasificación básica de las precipitaciones comprende dos tipos de precipitación: 1) “*la precipitación convectiva*”, que depende de condiciones de inestabilidad termodinámicas y dinámicas. Las lluvias convectivas por lo general son precipitaciones abundantes en un corto período de tiempo. 2) “*la precipitación frontal*”, la cual está asociada al pasaje de frentes fríos o cálidos y comprende una escala espacial mayor asociada con la escala sinóptica. La precipitación frontal generalmente tiene una intensidad menor que la convectiva y puede extenderse por varias horas o días.

Es importante mencionar que en una misma situación pueden ocurrir ambos tipos de precipitación.

Hay una serie de parámetros relevantes que caracterizan a la precipitación:

- Cantidad de precipitación: depende del tiempo de acumulación. Para tiempos de acumulación cortos como por ejemplo días y horas, la distribución de la precipitación generalmente no es simétrica, lo que indica que estas acumulaciones no están distribuidas en forma normal. Para la acumulación anual sí se puede llegar a ver una simetría, por lo que se puede asumir, en dicha escala temporal, una distribución normal de los datos de precipitación [13]
- Número de días lluviosos por mes/año: el número de días lluviosos por mes es a menudo independiente de la cantidad de precipitación.
- Rachas de días secos: días consecutivos donde la precipitación es 0 o por debajo de cierto umbral muy bajo.

## 1.2. Proceso de medición de la precipitación.

### Medición de la precipitación.

La medición de la precipitación se realiza para obtener información sobre sus características espaciales y temporales, como ser intensidad, frecuencia, cantidad, etc.

El problema de la representatividad, si bien es general para todas las mediciones climáticas, es particularmente importante en la precipitación, ya que esta variable presenta gran variabilidad espacial y temporal. Las mediciones de precipitación son espacialmente sensibles a la exposición, viento y a la topografía.

El diseño de una red pluviométrica no es un trabajo trivial. Rodríguez-Iturbe y Mejía [10] desarrollaron una metodología para el diseño óptimo, haciendo hincapié en la utilización de información previa para el diseño de sistemas de medición más específicos. Para esto es importante considerar la correlación espacial, temporal, el número de estaciones y la geometría de la red. Al diseñar un sistema de recolección de datos, dos preguntas básicas han de estar involucradas:

1. ¿Cuánto dinero se debe gastar en la red?, y ¿cómo se deben localizar las estaciones para obtener la mayor información?.
2. ¿Qué inferencias pueden ser extraídas de la colección de datos?, y ¿cuál es la incertidumbre de esas inferencias?.

La optimización económica se basa en el hecho de que el beneficio marginal del dato adicional debe ser menor que el costo de obtener el dato. La forma de implementarlo es a través del enfoque Bayesiano, donde la información es valiosa solo si el dato adicional puede causar un cambio de decisión o acción, y si ese valor medido produce una ganancia asociada con el cambio de acción.

El diseño se puede realizar de dos formas. Por un lado una red que opera con un presupuesto fijo debe ser diseñada para minimizar el error de estimación de las variables hidrológicas involucradas. Por otro lado una red que opera con un criterio de un mínimo de precisión requerido debe ser diseñada para minimizar costos.

La unidad en la que se mide la precipitación es en milímetros de agua. Por definición un milímetro de agua de lluvia equivale a 1 litro de agua por  $m^2$ . En Uruguay los instrumentales más utilizados son el pluviómetro y el pluviógrafo, que se encuentran dentro del instrumental denominado como convencional. Al día de hoy no existe en Uruguay una red de medición automática.

El pluviómetro está compuesto por tres secciones principales. La sección *receptora*, en su interior tiene un embudo con orificio de salida del agua y la parte superior está soldada a las paredes del pluviómetro, 10 cm por debajo de la boca. Esta sección receptora se acopla a la parte de *retención* y ambas forman un cilindro de 45.5 cm de altura. En el interior de la parte colectora hay un recipiente llamado *colector* que sirve para traspasar a la probeta el agua recogida para su medición. La sección de retención depende del tipo de pluviómetro, y puede variar desde los 50mm a los 500mm o más.

La lectura de la probeta es realizada por el observador, generalmente una vez al día en la red pluviométrica y cada seis horas en las estaciones meteorológicas.

El pluviógrafo es un instrumento registrador cuyo mecanismo es similar al pluviómetro, pero cuenta además con un sistema de relojería que está unido por medio de un tambor a una banda donde se registra la precipitación. El pluviógrafo aporta información sobre la intensidad de precipitación.

### **Ubicación y cuidado del pluviómetro.**

El instrumento debe colocarse en un lugar alejado de obstáculos, como ser árboles, edificios, etc., para que estos no eviten la entrada de lluvia al pluviómetro. La distancia horizontal entre el pluviómetro y los obstáculos debe ser por lo menos el doble de la altura de los mismos, además la boca del pluviómetro debe estar a una altura de un metro y medio sobre el piso.

Debe tenerse especial cuidado que la boca del pluviómetro esté horizontal, lo cual se comprueba con un simple nivel de carpintero.

En caso de que no haya jardín o lugar abierto sin obstáculos para colocar el pluviómetro en las condiciones requeridas y que sea necesario situarlo sobre un edificio, etc., se instalará de manera que la boca del pluviómetro sobrepase un metro la parte más alta del techo en el que ha sido colocado. Se colocará sobre un poste vertical, de manera que la cabeza del poste esté por lo menos 15 cm más baja que la boca del pluviómetro.

El pluviómetro deberá estar en lugares resguardados de gente con mala intención y donde no puedan llegar animales que lo muevan y/o dañen. Debe vigilarse siempre que en el interior del pluviómetro no haya hojas secas, tierra u otros objetos que impidan la entrada del agua. También debe cuidarse que el aro de la boca no sufra golpes que alteren su forma, para evitar que el dato registrado sea erróneo.

### **1.3. Algunos antecedentes.**

Paulhus y Kohler [8] realizaron el estudio de 59 series pluviométricas que tenían intervalos de datos faltantes en la misma, las mismas provienen de Estados Unidos de América. Para estas series antes de 1948 los totales anuales y mensuales eran omitidos cuando no se contaba con una porción de los registros. Un muestreo en 1950 de datos climatológicos anuales reveló que cerca del 25 % de las estaciones seleccionadas no tenían totales anuales, más aún los totales anuales para esos registros incompletos fueron también omitidos. En este estudio se investigaron métodos simples de interpolación diaria para el año 1950, los cuales pueden ser aplicados por personal administrativo o adecuados a procedimientos computacionales, estos métodos son:

1. *Método de las 3 estaciones* consiste en estimar la precipitación de la estación deseada ( $P_x$ ) utilizando la precipitación de tres estaciones cercanas para la interpolación

$(P_1, P_2, P_3)$ , la misma se realiza planteando la siguiente ecuación:

$$P_x = P_1 \cdot b_1 + P_2 \cdot b_2 + P_3 \cdot b_3$$

donde  $b_1, b_2$  y  $b_3$ , se estimaron por el método de mínimos cuadrados.

2. *Método de la razón normal* consiste en estimar la precipitación de la estación deseada ( $P_x$ ) utilizando la precipitación de tres estaciones cercanas para la interpolación ( $P_1, P_2, P_3$ ), y considerando los pesos como la normalización de la precipitación anual ( $N_x$ ), se llega a la siguiente ecuación:

$$P_x = \frac{1}{3} [P_1 \cdot \frac{N_x}{N_1} + P_2 \cdot \frac{N_x}{N_2} + P_3 \cdot \frac{N_x}{N_3}]$$

El estudio reveló que el método de la razón normal es más eficiente que el método de las 3 estaciones. El método de la razón normal será usado cuando la precipitación normal anual de alguna de las estaciones fijadas difiera de la estación interpolada en más de un 10 %. Si no sucede esto se puede usar cualquiera de los dos métodos .

Federico, Pascualoni, Avolio y Bellecci [3] trabajaron con 61 pluviómetros, desde el primero de enero de 1970 hasta el 31 de diciembre del 2006, pertenecientes a la región de Calabria, Italia. En dicho trabajo los autores llevaron a cabo un control de calidad de la base de datos de precipitaciones, y presentaron la tendencia para el período considerado. También estudiaron correlaciones entre los índices de teleconexión y las precipitaciones diarias.

La correlación usada para medir la asociación entre la precipitación y patrones a gran escala fue la de Spearman ordenada por rango (Spearman rank order o no paramétrica). La correlación anterior es elegida porque es tanto robusta como resistente estadísticamente hablando. Además, para evaluar el significado de la correlación no es necesario asumir distribución.

A partir de la correlación entre precipitación diaria y patrones de gran escala los autores concluyeron que el Índice de Oscilación Mediterránea (MOI) es un buen predictor para la precipitación sobre la región de Calabria.

Finalmente Vicente-Serrano, et al. [12], analizaron 3106 series pluviométricas durante el período que va desde 1902 hasta 2002 pertenecientes a la región del noreste de España, su estudio tuvo tres partes:

1. Se procedió a la reconstrucción de la series pluviométricas, ya que muchas de ellas tenían períodos sin registro. La falta de datos fue llenada con información de observatorios cercanos, fijando un umbral de distancia para evitar la introducción de información falsa en la serie, se evaluaron tres métodos para la reconstrucción de las series:
  - *Vecino más cercano*, este método se basa en llenar los vacíos directamente con datos del observatorio más cercano con información.
  - *Interpolación con ponderaciones inversas a la distancia*, este método está basado en utilizar estaciones cercanas ponderándolas con pesos inversos a la distancia a partir de la siguiente ecuación:

$$z(X_i) = \frac{\sum_{i=1}^n d_{ij}^{-r} \cdot z(x_i)}{\sum_{i=1}^n d_{ij}^{-r}}$$

donde  $z(x_j)$  es el valor predicho de acuerdo a los datos en los puntos  $z(x_1), z(x_2), \dots, z(x_n)$ , para este caso  $r = 2$ , finalmente la distancia máxima permitida en este caso fue de 15 Km.

- *Método de regresión lineal.*

En este estudio al analizar los 3 métodos propuestos, el método que dio mejores resultados fue el del vecino más cercano.

2. Se realizó un control de calidad de la serie. Para ello la serie original fue convertida a percentiles luego de eliminar los datos cero, luego de la transformación a los días cero les corresponde el percentil cero.

En un siguiente paso los percentiles empíricos correspondientes a cada observación fueron comparados con percentiles correspondientes a observatorios vecinos, un umbral de diferencia fue fijado para identificar extremos cuestionables.

3. En un tercer paso realizaron estudios de la homogeneidad de la serie usando el Standard Normal Homogeneity Test (SNHT). Esta prueba permitió la detección de inconsistencias presentadas en la base de datos original o reconstruidas. En este paso fueron estudiados cuatro parámetros en un nivel mensual estos son cantidad de lluvia, número de días lluviosos, máxima precipitación, número de días por encima del percentil 99.

Finalmente de las 3106 estaciones que poseían inicialmente, se logró una base de datos con 828 series largas y confiables.

## 1.4. Descripción de los problemas a enfrentar.

- En Uruguay la información proveniente de la Red Pluviométrica Nacional es informada telefónicamente a la Dirección Nacional de Meteorología (DNM) en forma diaria. Dicha información es transmitida solamente cuando hay registro de precipitación, ingresándola en un archivo no secuencial. El problema detectado es que no se distingue entre lo que se puede calificar como “dato faltante 0 (“no lluvia”), información que es conceptualmente muy distinta. Esto genera una complicación importante para el estudio, por ejemplo, de las sequías, secuencia de días secos, etc.
- Otro problema que se puede presentar es la asignación de varios días de lluvia a tan solo uno. Esto afecta el estudio de precipitaciones intensas y posibles atribuciones a inundaciones ya que lo que genera es un falso evento de lluvia intensa; repercutiendo también en el estudio de rachas secas.
- La distribución de pluviómetros no es uniforme en todo el país y las distancias entre los mismos varía, es decir, hay algunos departamentos que cuentan con una alta densidad de pluviómetros, mientras que hay otros que cuentan con muy poca. Es muy importante para interpolar un dato poder contar con información confiable acerca de estaciones cercanas.

El objetivo planteado en esta tesis es el de poder determinar una metodología estadística óptima que logre detectar errores, para luego poder mejorar la calidad de las series de precipitación diaria en la mayor cantidad de puntos para Uruguay.

Contar con una serie de precipitación de buena calidad es de vital importancia para poder caracterizar eventos de una manera confiable y de esta manera poder determinar patrones de circulación atmosférica asociados para poder predecirlos, así como ser una fuente de información certera para los tomadores de decisiones y poder realizar políticas acorde a la realidad.

## Capítulo 2

# Base de datos.

### 2.1. Descripción de la base de datos.

Al comienzo de este estudio se analizó información diaria proveniente de 163 pluviómetros durante el período 1960-2008. Del total de puntos; 5 son pluviómetros de estaciones agro-meteorológicas pertenecientes al Instituto Nacional de Investigación Agropecuaria (INIA), 15 pertenecientes a Usinas y Trasmisiones Eléctricas (UTE) y los 143 pluviómetros restantes son pertenecientes a la Dirección Nacional de Meteorología (DNM).

De la información brindada por la DNM, ocho series provienen de estaciones meteorológicas, las demás corresponden a estaciones de la Red Pluviométrica Nacional. La diferencia entre las estaciones meteorológicas y las de la red pluviométrica es que en las estaciones meteorológicas se miden, además de la precipitación, otras variables atmosféricas como son temperatura, presión, humedad etc., y por otro lado el personal cuenta con capacitación específica. Mientras que en la red pluviométrica, debido a un convenio estipulado entre el Ministerio de Defensa y el Ministerio de Interior, los pluviómetros se encuentran ubicados en diferentes secciones policiales del interior del país operados por el personal policial. Se constató también que dicha red no cuenta con un mantenimiento periódico del instrumento lo cual podría influir en la calidad de las series analizadas.

La figura 2.1 muestra la distribución geográfica de las estaciones pluviométricas, con registros en al menos un año en el período de estudio, clasificadas según la fuente de información. Se observa el bajo número de estaciones de la DNM en la zona centro del país, especialmente en los departamentos de Durazno, Florida, Rivera y Treinta y Tres. Por otro lado las estaciones de UTE se concentran en las cuencas del Río Negro y la zona norte de la cuenca del Río Uruguay, debido a la alta necesidad de monitorear las mismas.

Se debe dejar en claro que no se obtuvo la totalidad de puntos pertenecientes a la DNM, sino que los datos provenientes de la DNM son los brindados por ellos. Lo mismo ocurrió para las estaciones de UTE.

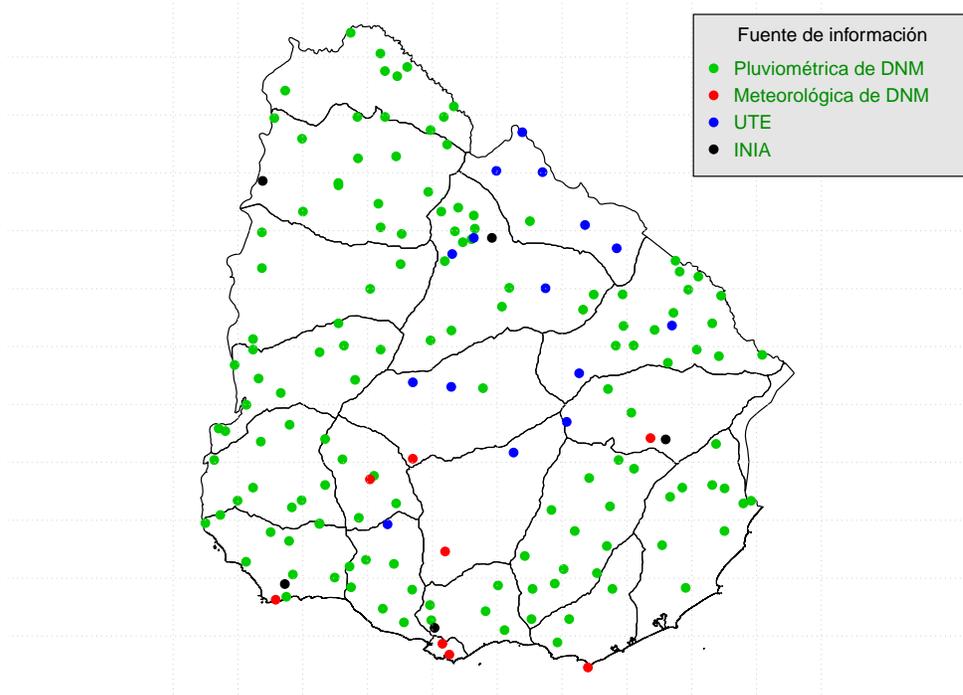


Figura 2.1: Distribución geográfica de las estaciones en el mapa de Uruguay según fuente de información.

Al solicitar la información a la DNM, parte de la misma se encontraba en formato papel la cual a partir de nuestra solicitud se digitalizó. En dicho pedido se solicitó información que cubriera por lo menos el período que va desde el 1960 hasta el 2008 inclusive. Se costató que algunas de las series presentaban períodos de registro menor a lo solicitado, pero comprendidos dentro del mismo.

La información proveniente de INIA se encuentra on-line y es de acceso público a través de la página <http://www.inia.org.uy>. De las cinco estaciones INIA solamente La Estanzuela, ubicada en el departamento de Colonia, cubre el período 1965-2008, en las demás la fecha de comienzo de registro es posterior.

Las series provenientes de UTE en general comienzan en las primeras décadas del Siglo XX, finalizando todas en el año 2004. Debido a que en este estudio se consideró un período común para todas las fuentes de información, en estas estaciones se considerarán solamente los registros a partir del año 1960.

Para cualquiera de las tres fuentes de información los registros son a escala diaria. El “día” de precipitación no es el día calendario, es decir no se considera de 0 a 24 horas. Entre las distintas fuentes existen diferencias en la consideración de “día” de precipitación.

La red pluviométrica de DNM considera “día como la precipitación acumulada en el período de 24 horas comprendido entre las 07:00 AM hora local de un día y las 07:00 del día siguiente, pero el registro se adjudica a la fecha del día anterior. Por ejemplo, si en la

medición de las 07:00 AM del día 15/5 se registran 10 mm, se codifica que la precipitación del día 14/5 fue de 10 mm. Por otro lado en las estaciones meteorológicas de DNM si bien se realizan mediciones cada seis horas (09:00, 15:00, 21:00 y 03:00 hora local), el día es considerado como el acumulado entre las 09:00 AM hora local del día al cual se quiere asignar la precipitación y las 09:00 AM del día siguiente.

NOTA: las mediciones cada seis horas en las estaciones meteorológicas, es un requerimiento del código meteorológico SYNOP regulado por la OMM.

En las estaciones de INIA se procede igual que en la Red pluviométrica de DNM pero la medición es a las 9 AM hora local. Es importante mencionar que en la estación de Las Brujas (ubicada en el departamento de Canelones) desde fines del año 2000 los registros provienen de una estación automática.

No se obtuvo información acerca de la metodología de medición en las estaciones de UTE.

Dada las diferentes formas de definir el dato diario de precipitación, se observó la necesidad de utilizar un protocolo nacional de medición de esta variable, ya que la no consideración de este aspecto repercute negativamente en el análisis diario de las series.

Los pluviómetros provenientes de la DNM poseen un código identificador de cuatro cifras (por ejemplo 1001). Mientras que las estaciones meteorológicas cumplen con una codificación internacional dictaminada por la Organización Meteorológica Mundial (OMM) la cual consta de cinco cifras, donde las dos primeras representan al país y las últimas tres representan a la estación meteorológica, por ejemplo el código OMM de la estación meteorológica de Treinta y Tres es 86500 en donde 86 es el código asignado a Uruguay, y el 500 corresponde a la estación Meteorológica mencionada. Con el propósito de homogeneizar la codificación de las estaciones se le reasignó una compatible con la red pluviométrica, de la siguiente manera:

- La estación 86560 de Colonia pasa a ser la estación Nro. 2774.
- La estación 86530 de Durazno pasa a ser la estación Nro. 2206.
- La estación 86532 de Flores pasa a ser la estación Nro. 2297.
- La estación 86545 de Florida pasa a ser la estación Nro. 2583.
- La estación 86595 de Punta del Este pasa a ser la estación Nro. 2909.
- La estación 86585 de Montevideo (Carrasco) pasa a ser la estación Nro. 2887.
- La estación 86575 de Montevideo (Melilla) pasa a ser la estación Nro. 2867.
- La estación 86500 de Treinta y Tres pasa a ser la estación Nro. 2179.

Para los pluviómetros de la DNM se detectó que dos estaciones estaban codificadas con el mismo número 2480 aunque en ubicaciones distintas; uno ubicado en Colonia y el otro en el departamento de Soriano, por ende se procedió a recodificar las mismas como 2480C y 2480S donde la última letra indica su respectivo departamento.

La tabla del anexo A.1 contiene toda la información básica (ubicación, período de registro etc.) acerca de la base de datos analizada en este estudio.

Es importante destacar la ausencia prácticamente de lo que es el metadato, lo cual comprende la historia de la estación, cambios del observador, del instrumental etc. este componente es importante a la hora del análisis de las series.

## 2.2. Tratamiento y depuración de coordenadas.

La DNM nos envió un archivo Excel conteniendo las coordenadas a escala decimal para la ubicación de 274 puntos. Teniendo en cuenta que el archivo que contenía la información pluviométrica contaba con 143 puntos, se constató que la información geográfica proporcionada excede a la real, observándose que dentro de estos 274 puntos se encontraban ubicaciones referentes a estaciones de UTE (15 puntos). Por lo cual se procedió a identificar solamente la ubicación de estos 158 puntos (143+15).

Las estaciones de INIA estaban geolocalizadas en Internet de forma sexagesimal, por lo que no hubo ningún tipo de problema.

Un inconveniente presentado fue que la escala decimal no es la utilizada por el Google Earth, ya que utiliza la escala sexagesimal, es decir grados, minutos, segundos, etc. (de 0 a 60), por lo que se tuvo que cambiar de escala de medición para poder utilizar esta herramienta.

## 2.3. Representación geográfica de las estaciones.

En esta sección se detalla en forma minuciosa cómo se logró, a partir de las coordenadas geográficas, la representación de las mismas.

Es importante mencionar que en esta sección los software utilizados fueron el programa R (el mismo se puede descargar desde la página <http://www.r-project.org/>), Excel y Google Earth (utilizado para verificar y controlar la correcta aplicación de las fórmulas desarrolladas, se puede descargar desde la siguiente página <http://www.google.com/intl/es/earth/download/ge/agree.html>).

### 2.3.1. Procedimiento para el trazado del mapa en R.

En nuestro caso se necesitó obtener un mapa de Uruguay detallado por departamentos en R. Para ello utilizamos el “Global Administrative Areas”(GADM, <http://gadm.org>) es una base de datos gratuita de la localización espacial de la áreas administrativas del mundo (o de los límites administrativos).

Este sitio proporciona información de los mapas como objetos de R que se pueden trazar directamente con la función *plot* o con el uso de la biblioteca *sp*, además para Uruguay esta base de datos hace una partición departamental.

Para utilizar el mapa de Uruguay, primero se debe descargar la biblioteca *sp* que sirve para hacer algunos gráficos por departamento. Luego para cargar los datos de Uruguay en R se ubica directamente el sitio web, con el comando de R:

```
load(url("http://gadm.org/data/rda/URY_adm1.RData"))
```

Esta sentencia almacena en la memoria un objeto de datos llamado *gadm*.

### 2.3.2. Representación de las estaciones pluviométricas.

Las estaciones en el mapa fueron representadas con el número de estación (ver figura 2.2). Se observó que algunas estaciones que fueron asignadas como pertenecientes a un departamento, tenían coordenadas que correspondían a otro; este es el caso de la 2215, que fue enviada como perteneciente al departamento de Durazno, pero sus coordenadas

pertenecen al departamento de Florida. No obstante como atenuante a la situación tenemos que las estaciones en cuestión están muy cerca del límite departamental, lo cual disminuye mucho la gravedad del problema. En general las estaciones presentan una ubicación correcta.

Se observó que la región peor representada es la región central en especial los departamentos de Durazno, Florida y Treinta y Tres. Mientras que la región sur y oeste presentan una buena densidad de puntos en relación a las demás regiones. Se destaca que esta representación se basó solamente en la cantidad de estaciones pluviométricas, todavía no se ha detallado nada acerca de la calidad ni de la cantidad de años de registro de la misma.

Esto fue de suma utilidad ya que nos dió una idea de distribución espacial de los datos. Además la representación geográfica nos permitió colorear los puntos representados según distintos criterios, como pueden ser años de registro, porcentaje de datos faltantes, número de días que no llovió, etc. A partir de esta representación se obtuvo una perspectiva gráfica del problema a enfrentar, cuales son las zonas más dificultosas en el estudio, etc.

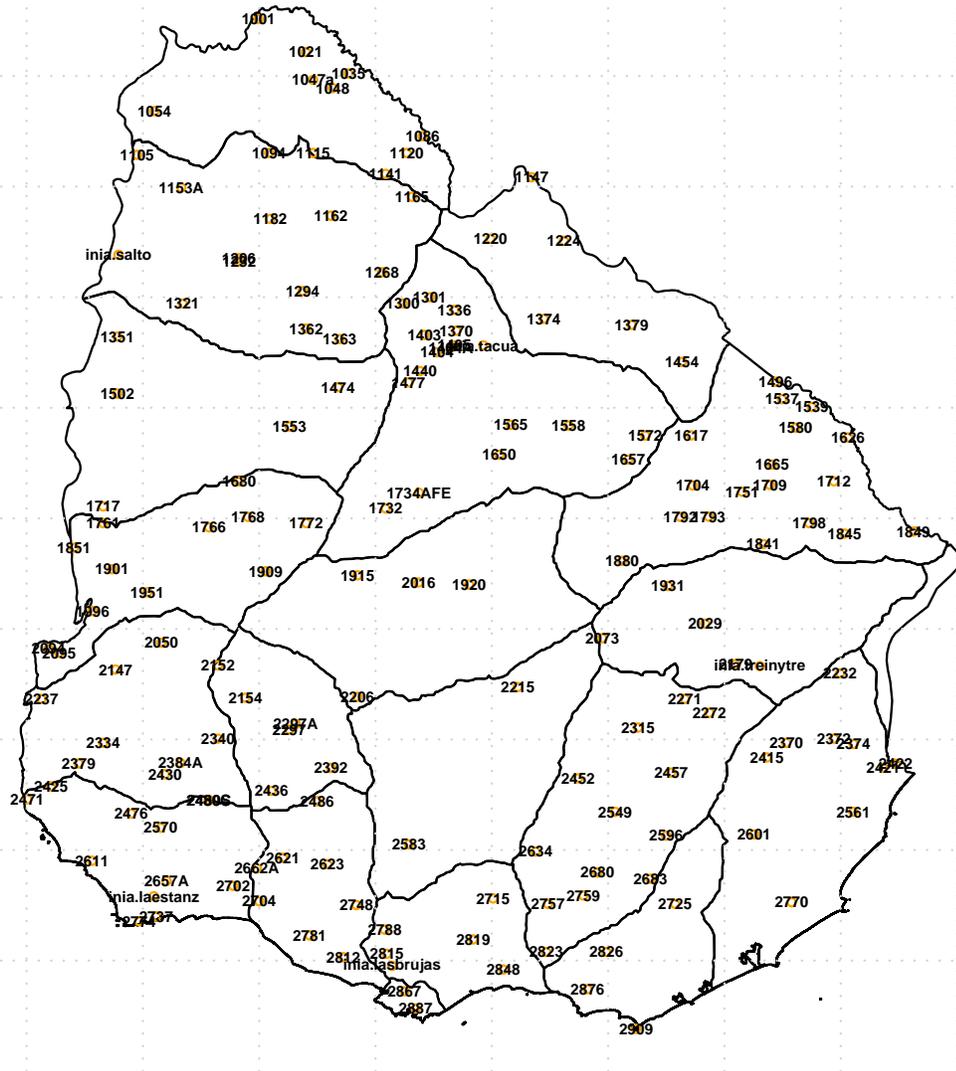


Figura 2.2: Representación geográfica de las estaciones el mapa de Uruguay.

## Capítulo 3

# Control de calidad de la base de datos.

### 3.1. Lectura de los datos.

Los errores de consistencia de la serie son aquellos errores detectados al recibir la base de datos. Un primer análisis de consistencia permite detectar errores indiscutibles, por ejemplo precipitaciones en días inexistentes, precipitaciones negativas o diferentes formas de codificación del dato faltante (en este estudio se utilizó el -99.9 para indicar un dato faltante).

Este tipo de error pudo deberse en principio a dos situaciones: una primer situación derivada de un error de quien tomó la medida de la estación pluviométrica, por citar un ejemplo esté pudo por equivocación haber anotado un día erróneo; una segunda situación se deriva de un error asociado a la digitalización de los datos.

Primeramente fue necesario establecer un formato de lectura básico para asegurar la uniformidad de los datos y por lo tanto su correcto estudio. En un primer estudio de las series y consultando con las fuentes de información, se establecieron como necesarias las siguientes características:

- No debe haber registros de precipitación desordenados cronológicamente.
- Si hay registro de lluvia en el día, no se consideró como posible dato faltante, aunque si puede estar mal ingresado su valor de precipitación (el estudio de la correcta digitalización no fue estudiado aquí a excepción de casos excepcionales).
- Para la DNM todos los meses de todos los años tienen algún registro. En el caso de meses sin datos, la DNM usó distintas codificaciones, nosotros a modo de lograr una uniformidad en los datos tomamos como correcta la codificación -99.9 una única vez en el mes.
- Para los datos de la DNM cuando no llueve en todo el mes se marcó un solo dato mensual que por lo general es en el día 1 con precipitación igual a 0.
- Para los datos de UTE, todos los días hay algún tipo de registro aún cuando no se obtuvo el dato se codificó el dato faltante. Por lo que hay una codificación en todos los días del año.

- Se define traza cuando la precipitación diaria no supera los 0.1mm, por lo general se codificó como TRZ. Cuando se detectaron estos casos se le cambió la codificación a valor real 0.1, con el objetivo de cuantificar el valor de la traza.
- Para el INIA, todos los días hay algún tipo de registro aún cuando no se obtuvo el dato se codificó como -99.9 indicando que el dato es faltante.

En este capítulo se tratan en profundidad las características de los datos, los errores visualizados desde el punto de vista de la codificación de los datos faltantes así como también errores en el tipeo de los días con precipitación, o días cambiados de lugar, etc.

Todo este análisis fue fundamental ya que permitió conocer las estaciones con sus particularidades, se mejoraron las series sin tener que acudir todavía a métodos estadísticos. En esta etapa hubo interacción con los informantes, es decir los organismos que nos otorgaron la base de datos, en la cual ellos fueron nuestro nexo entre el registro electrónico y el registro en el papel y nuestra única forma de acceder al mismo.

Antes de comenzar con este análisis es necesario mostrar la estructura de las series estudiadas a modo descriptivo. En las series de DNM la información está organizada en cinco columnas (de izquierda a derecha): estación, año, mes, día, precipitac (que es la precipitación diaria en mm) si no están exactamente con esos nombres el programa en R modifica el nombre de las columnas distintas al nombre deseado utilizando el comando *colnames*. A modo de ejemplo se muestra como se realiza el registro:

ESTACION	AÑO	MES	DIA	PRECIPITAC
2867	1960	5	22	0.5
2867	1960	6	4	18.0
2867	1960	6	5	20.0
2867	1960	6	7	2.5
2867	1960	6	17	7.2
2867	1960	6	18	8.2
2867	1960	7	6	26.0
2867	1960	7	12	25.0
2867	1960	7	13	12.5
2867	1960	8	8	45.0
2867	1960	8	13	35.0

Note que en los días intermedios no se sabe si es que no llovió o que no se registraron datos. En principio se registró como no lluvia (no como dato faltante) y se intentó poner a prueba la veracidad de la misma.

En UTE, las series están organizadas en tres columnas, la primera marca la fecha, la segunda la precipitación observada, y la tercera nos indica si fue registrado el dato, es decir, en caso de haber registro la tercer columna es igual a 1 y en la segunda columna se indica la precipitación correspondiente, en caso contrario la tercer columna es igual a 0 y en la segunda columna puede ir un 0 o el espacio en blanco. A modo de ejemplo, se muestra el siguiente tramo de una serie de UTE:

Fecha	Cuchilla del Carmen 1880	Si hay valor = 1
06/12/1941	0	1
07/12/1941	0	1
08/12/1941	34	1
09/12/1941	4	1

En este caso los datos corresponden a la estación de Cuchilla del Carmen (Cerro Largo), la segunda columna representa la cantidad de precipitación diaria (en mm).

### 3.2. Errores detectados.

Es importante señalar que no fue tarea fácil la correcta lectura de todas las series, ya que en varios casos algunas filas aparecían como ocultas en los archivos Excel, y como consecuencia de esto el R no alcanzaba a leer todos los datos (esto fue solucionado desocultando las filas ocultas). Otro típico error de lectura apareció al encontrarse palabras o frases fuera de las filas de estudio.

Por otro lado se realizó mediante un script la detección de errores de inconsistencia, como ser datos que se desvían de lo lógico (días que en calendario no existen, o una distinta digitalización de los datos faltantes) o datos no ordenados en forma cronológica.

También el script realiza el llenado de datos donde no hay registro de precipitación con 0 y en caso de saber que es dato faltante se llenó con -99.9 (En una etapa posterior se intentó probar para cada uno de los ceros rellenos, si realmente es un día sin lluvia o es un dato faltante).

A continuación se detallan los errores de inconsistencia detectados en la base de datos:

- Error en el protocolo de codificación: En el siguiente caso uno no sabe que sucedió el día 21. Dado que este no fue el único registro en el mes y se supone que el -99.9 en el caso de la DNM no se usó para datos particulares sino que solo se usó para meses en los cuales fueron todos datos faltantes, se consultó con la fuente de DNM. La decisión tomada fue no considerar ese día como dato faltante, posteriormente se relleno el mismo con un 0.

ESTACION	AÑO	MES	DIA	PRECIPITAC
1035	1995	2	21	-99.9

- Errores de fecha: El siguiente ejemplo muestra un registro el 31 de setiembre, al consultar la fuente se constató que el dato correspondía al día 30. Casos similares se detectaron en los meses de abril, junio y noviembre. En el mes de febrero que muestra las características del año bisiesto se detectaron también errores a fines de mes. A menudo el día no es el correcto, sin embargo a veces no se tiene información al respecto. Al consultar con la fuente el dato es asignado a un día consecutivo o es quitado de los datos según corresponda.

ESTACION	AÑO	MES	DIA	PRECIPITAC
1035	1973	9	31	4

También se detectaron casos con codificación de días mayores a 31, afortunadamente fueron menos frecuentes que los otros.

- Hubo muchos casos en los cuales en el lugar donde debe ir el monto de precipitación apareció S/R; SIN REGISTROS o SIN DATOS. En estos casos quedó la duda si esta

anotación se debió a que no llovió o a un dato faltante. En la amplia mayoría de los casos se debió a un mes faltante, sin embargo hubo excepciones. En consecuencia de esto se sustituyó por -99.9 o 0 según correspondiera. Por ejemplo, en el mes de abril de año 1966 para la estación 1048 era dato faltante, lo mismo pasó en el mes de agosto de 1999.

ESTACION	AÑO	MES	DIA	PRECIPITAC
1048	1966	4	S/R	S/R

ESTACION	AÑO	MES	DIA	PRECIPITAC
1048	1999	8	SIN REGISTROS	

- Se detectó algunos días donde en vez de registrarse -99.9 para el dato faltante se codificó -1, para todos estos casos se sustituyó el -1 por -99.9.
- Se encontraron casos que en donde debería registrarse el mes, no había ningún registro. Debido a que los datos están ordenados en forma cronológica se puede hacer inferencias acerca del posible mes (lo mismo pasó con los días), de todos modos se consultó a la DNM. A modo de ejemplo, en el siguiente caso se consultó a la DNM y se encontró que los espacios en blanco correspondían al mes 7.

ESTACION	AÑO	MES	DIA	PRECIPITAC
1048	2001	6	25	28.0
1048	2001		1	2.0
1048	2001		10	10.0
1048	2001		17	5.0
1048	2001	8	6	11.0

- En algunas fechas, en vez de registrar un mes se registró otro distinto, aquí en el ejemplo, en vez de colocarse mes 9 se colocó por error mes 8, lo cual daba lugar a confusión y fue necesario consultar la fuente.

ESTACION	AÑO	MES	DIA	PRECIPITAC
1537	1989	8	30	37.0
1537	1989	9	5	13.0
1537	1989	8	11	3.0
1537	1989	8	21	10.0
1537	1989	9	22	29.0

- Casos más complejos: Hubo un par de casos donde no se comprendió como se pudo obtener un dato, el siguiente ejemplo es una clara exposición de lo anteriormente expresado.

ESTACION	AÑO	MES	DIA	PRECIPITAC
1704	2004	5	SIN REGISTROS DEL DATO	
1704	2004	6	SIN REGISTROS DEL DATO	
1704	2004	7	SIN REGISTROS DEL DATO	
1704	2004	8	SIN REGISTROS DEL DATO	
1704	2004	9	SIN REGISTROS DEL DATO	
1704	2004	10	23	26.0
1704	2004	11	SIN REGISTROS DEL DATO	
1704	2004	12	SIN REGISTROS DEL DATO	
1704	2005	1	SIN REGISTROS DEL DATO	
1704	2005	2	SIN REGISTROS DEL DATO	
1704	2005	3	SIN REGISTROS DEL DATO	
1704	2005	4	SIN REGISTROS DEL DATO	
1704	2005	5	SIN REGISTROS DEL DATO	

En este caso aunque se consultó a la DNM por el dato y la misma nos respondió que el dato era correcto, se optó por quitarlo de la serie y se registró el mes de octubre como mes con datos faltantes.

- En un par de casos ocurrió que hubieron registros de 999.9 mm. En un principio se pensó que esto era debido a una falta de registros mensuales de precipitación, sin embargo esta teoría se revocó rápido al ver que había más registros en el mes. Para todos estos casos se consultó a la fuente, y se decidió no considerar dicha precipitación dado que no había registros en la planilla original.
- En algunas estaciones había meses o incluso años que no tenían registro, es decir ni como dato faltante, ni las precipitaciones correspondientes. Aquí se muestra un ejemplo de ello.

ESTACION	AÑO	MES	DIA	PRECIPITAC
1537	1999	12	4	2.0
1537	1999	12	12	22.0
1537	1999	12	29	4.0
1537	2001	1	6	35.0
1537	2001	1	8	15.0
1537	2001	1	9	44.0
1537	2001	1	17	32.0

En este caso la DNM nos envió el año que quedó sin registrar ya que sí estaban los registros, sin embargo no en todos los casos ocurrió esto, algunas veces fue directamente un año completo de datos faltantes.

- Se detectaron datos faltantes en partes de los meses, no necesariamente en el mes entero. Por ejemplo, hubo un par de días que en el espacio para precipitación apareció S/C en referencia a que no hubo comunicación con la seccional, estos casos se tomaron como datos faltantes ya que el dato no pudo ser recuperado.

- Para las estaciones de UTE en algunas fechas de la serie no había registro. Por ejemplo, en Cuchilla del Carmen (Cerro Largo) se pasa del 02/01/1997 al 04/01/1997. Para este caso se agregó el día 03/01/1997 pero en donde dice la precipitación va un espacio en blanco, y en la tercer columna que indica si el valor existe se coloca un cero.

Fecha	Cuchilla del Carmen 1880	Si hay valor = 1
01/01/1997	0	1
02/01/1997	0	1
03/01/1997		0
04/01/1997	0	1

### 3.3. Superposición de series en un mismo punto.

Existen algunas estaciones que al momento de recibir la serie tenían una parte que pertenece a la DNM y otra parte que perteneció a la Administración de Ferrocarriles del Estado (AFE). Dado que AFE cerró en 1985, ocurre que desde 1960 hasta 1985 AFE se encargó de los registros pluviométricos para luego pasar a la DNM.

De la ubicación de las mismas solo se tiene información precisa de la DNM, sin embargo la distancia entre las estaciones de AFE y de las de DNM fueron consideradas despreciables.

El número de estaciones de AFE es ocho, de las cuales seis se pudieron ajustar perfectamente con las de DNM, ya que cuando terminó la serie en AFE, empezó en DNM. Sin embargo, hay dos casos donde hay años en común entre una serie y otra.

Para la serie 1047 de Artigas, tenemos por un lado que la serie de AFE abarca desde el año 1960 hasta el año 1985 inclusive (codificada como afe1047) y por otro que la serie de la DNM comenzó en el año 1981 llegando hasta el año 2008 inclusive (codificada como 1047).

En un primer paso se verificó que efectivamente se pudieran unir ambas series, para ello se consideró la correlación de Pearson (cuya fórmula y desarrollo se detallará en la sección 5.4.1) en el período en común entre las dos estaciones. La misma dió 0.943, esta medida es bastante cercana a 1, por lo tanto, debido a las características de la correlación de Pearson, es correcto unir a las series en cuestión.

Para realizar la unión de ambas series, en principio se pensó en calcular un promedio diario de la cantidad de precipitación en los años en común entre las dos estaciones, sin embargo esto no es correcto ya que desembocaría en una mayor cantidad de días lluviosos que la expuesta en la serie (ya que si llueve en alguna de las dos series en la serie resultante va a llover). Por lo tanto, se optó para esos años en común, dejar una de las dos series. Para esto, se seleccionó aquella serie que mejor se correlacione con las series pluviométricas de estaciones cercanas. Esto es correcto ya que se conservó de las dos series, la más parecida a las otras series.

En una primera etapa se consideraron solamente las estaciones que se encuentran a una distancia cercana a la estación de la DNM, para este caso la distancia máxima permitida fue de 20 Km (en la sección 4.2 se detallará la forma de cálculo de las mismas). A partir de este filtro se obtuvieron tres estaciones que cumplen dicha condición, estas son las estaciones 1048, 1021 y 1035 (ver tabla 3.1 para distancia entre estas estaciones y la 1047).

Estación	1048	1021	1035
1047	11.48	15.29	19.57

Cuadro 3.1: Distancia en Km a la estación 1047 de las distintas estaciones cercanas.

Se calcularon dos medidas de interés, la correlación de Pearson y la concordancia, esta última se basa en considerar las respectivas series como binarias (lluvia o no lluvia) y calcular para cada par de estaciones el porcentaje de veces que en ambas estaciones se registra 0 o 1 en ambas. Es decir, se contemplan los casos en que en ninguna de las dos estaciones se haya registrado lluvia y en que en ambas estaciones se registró lluvia.

Es necesario mencionar que la cantidad de datos faltantes no afecta la estructura de correlación, ya que se consideró que el dato faltante es independiente a la presencia de lluvia en ese día.

Tanto la correlación de Pearson como la concordancia se estudiaron para el período en común que existe entre las dos series (período 1981-1985, cinco años), ya que de esta forma los valores en las medidas son comparables.

Para las series involucradas se obtuvieron los siguientes resultados:

Características de las estaciones más cercanas a la estación 1047:

Estación	1048	1021	1035
Período	1960-2008	1960-2008	1960-2008
% de datos faltantes	9.36	1.03	3.04

Concordancia entre estaciones:

Estación	1048	1021	1035
1047	0.945	0.939	0.927
AFE1047	0.943	0.933	0.921

Correlación entre estaciones:

Estación	1048	1021	1035
1047	0.879	0.851	0.813
afe1047	0.881	0.825	0.801

La concordancia es mayor para la estación de la DNM respecto a la de AFE aunque la diferencia es poco significativa. Sin embargo, la diferencia en las correlaciones entre las estaciones con respecto a la 1021 y a la 1035 es bastante más significativa, por lo tanto se eligió la estación perteneciente a la DNM.

La estación 1094 de la DNM tenía cinco años en común con la de AFE1094, en primer lugar vimos que la correlación entre las dos estaciones considerando solamente los años en común es de 0.933, la misma es bastante cercana a 1 por lo que si fue posible unir las series, luego al escoger una de las dos estaciones para cubrir los años en común se eligió la de la DNM ya que la de AFE para ese período de unión tiene más cantidad de datos faltantes.

### 3.4. Características de las estaciones.

Luego de haber homogeneizado la estructura de los datos, se obtuvieron para cada serie algunas medidas de interés. A tal propósito se realizó una tabla con las siguientes columnas:

- *Número de estación y departamento.*
- *Año de inicio y de finalización de la serie con la cantidad de días que cada serie abarca.* Por ejemplo si la serie va desde el primero de enero de 1960 hasta el 31 de diciembre del año 2008 la misma contiene un total de 17898 días independientemente de los datos faltantes de la misma.
- *El número de datos faltantes con su respectivo porcentaje sobre el total de días abarcados por la serie.* Esto fue esencial como forma de clasificar las series a modo de visualizar cuales son las más completas (menos problemáticas) así como también en que regiones del país se encuentran las mismas. Este abarca desde estaciones que no contienen datos faltantes, a estaciones con un alto porcentaje de estos. La estación de UTE número 1147 de Rivera es la que contiene un mayor porcentaje de datos faltantes con un total de 47.00 % de datos faltantes.
- *El número de días sin precipitación con su respectivo porcentaje sobre el total de días abarcados por la serie.* Esta medida sirvió como forma de observar a priori estaciones con posibles datos faltantes debido al relleno de los períodos entre lluvias con cerros. Aquellas estaciones con un alto porcentaje de días sin lluvia durante el período conllevan una cierta incertidumbre acerca de la veracidad de estos. En la estación del INIA ubicada en Treinta y Tres se registró el menor porcentaje de días sin lluvia sobre el total de días abarcados por la serie, con un porcentaje de 69.21 %, por otro lado se encuentra la estación de la red pluviométrica de la DNM número 1558 de Tacuarembó con un porcentaje de 86.69 %, un 17 % más de días sin lluvia sobre el total de días abarcados por la serie respecto a la estación de Treinta y Tres.
- *El máximo de precipitación diaria durante el período y la fecha de ocurrencia.* En la estación de la red pluviométrica de la DNM número 1165 de Salto se registró el máximo registro de precipitación diaria el día 21/12/1997 con una precipitación de 403mm.
- *El máximo acumulado en cinco días durante el período y el año en que ocurrió este fenómeno.* En la estación de la red pluviométrica de la DNM número 1268 de Salto se ha registrado el valor más alto, donde en el año 1987 llovió un total de 517 mm en cinco días.

Una tabla muy similar a la construida con estas medidas, pero solamente para las estaciones del departamento de Colonia se puede ver en la tabla 5.2 de la sección 5.2.

### 3.5. Primer análisis de datos faltantes.

En esta sección se analizó el porcentaje de datos faltantes en las distintas estaciones pluviométricas, discriminado por fuente de información y década. Un análisis análogo se podría hacer por departamento, sin embargo se realizó el mismo al momento de analizar el departamento en cuestión.

Para este análisis, se realizó una tabla donde para cada estación se detallan los períodos donde ocurren los faltantes de datos. En la misma se encuentra el porcentaje de datos faltantes para un mes y una estación cualquiera, esto permitió ver donde ocurrió la mayor parte de estos, ya sea en termino de años o de meses. Se estudió también si en los meses

donde hay vacaciones hay una mayor cantidad de datos faltantes (es decir diciembre, enero, febrero y julio).

Se estudiaron series pluviométricas con distintas longitudes, es decir que recorren períodos de tiempo distintos dentro del período 1960-2008. Por lo tanto, se exploró en que años se encuentra la mayor cantidad de estaciones. El gráfico 3.1 muestra el promedio del número de estaciones por año. A partir de este se observó que en la década del '60 hay menor cantidad de estaciones, luego la misma va en aumento, hasta 1971 donde permanece constante, decreciendo a partir del año 2001, incluso luego del año 2004 se observó una gran caída debido a que a partir de allí no se contó con información acerca de las estaciones de UTE.

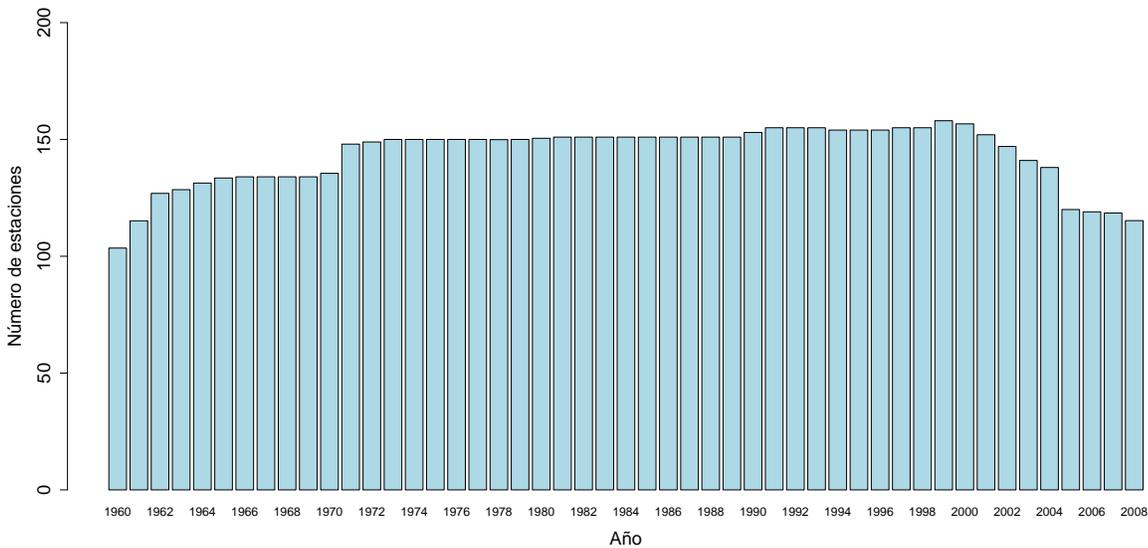


Figura 3.1: Promedio del número de estaciones pluviométricas por año.

El gráfico 3.2 acota las estaciones de estudio solamente a aquellas pertenecientes a la DNM, ya sean de la red pluviométrica o meteorológicas, este gráfico permitió observar que la caída en la cantidad total de estaciones en el 2004 se debió solamente a la finalización de las series pertenecientes a UTE. En este gráfico se observó que desde el '60 al '70 crece en los primeros tres años, luego permanece constante hasta 1970 donde se observó un incremento, luego se estancó hasta el año 1999 que llega a su punto más alto y finalmente decreció hasta el 2006.

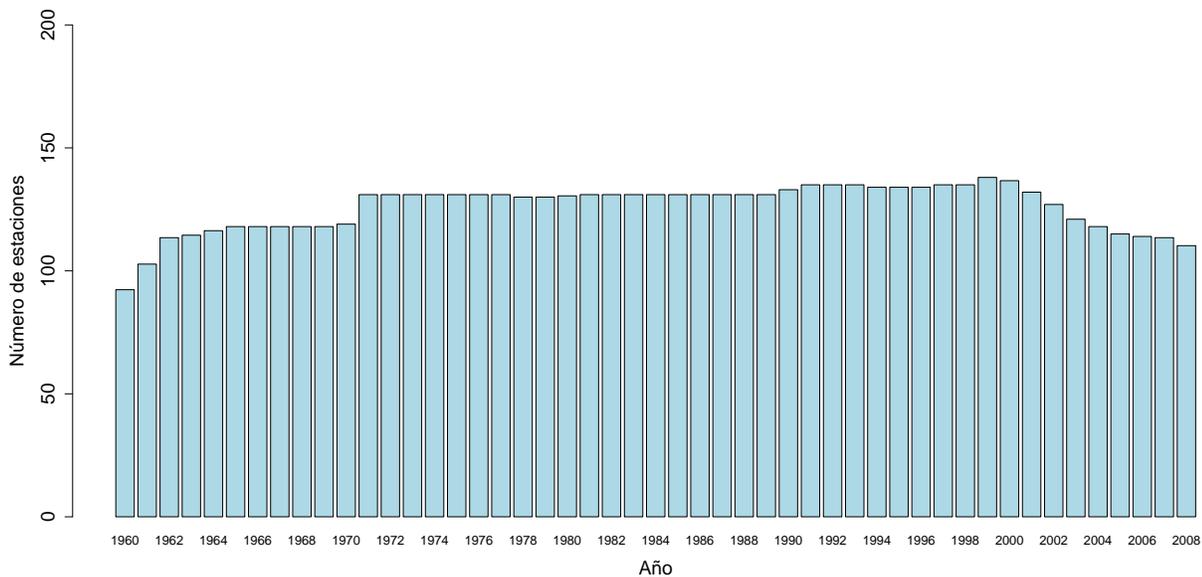


Figura 3.2: Promedio del número de estaciones pluviométricas por año para DNM y Meteorológica.

Al estudiar el porcentaje de datos faltantes por año (ver gráfico 3.3); se observó que a medida que pasa el tiempo el porcentaje de datos faltantes disminuyó progresivamente alcanzando un período desde el año 1996 hasta el 2003 con un bajo porcentaje de datos faltantes (en ese intervalo de tiempo la máxima cantidad de datos faltantes en un año está en el entorno del 3%).

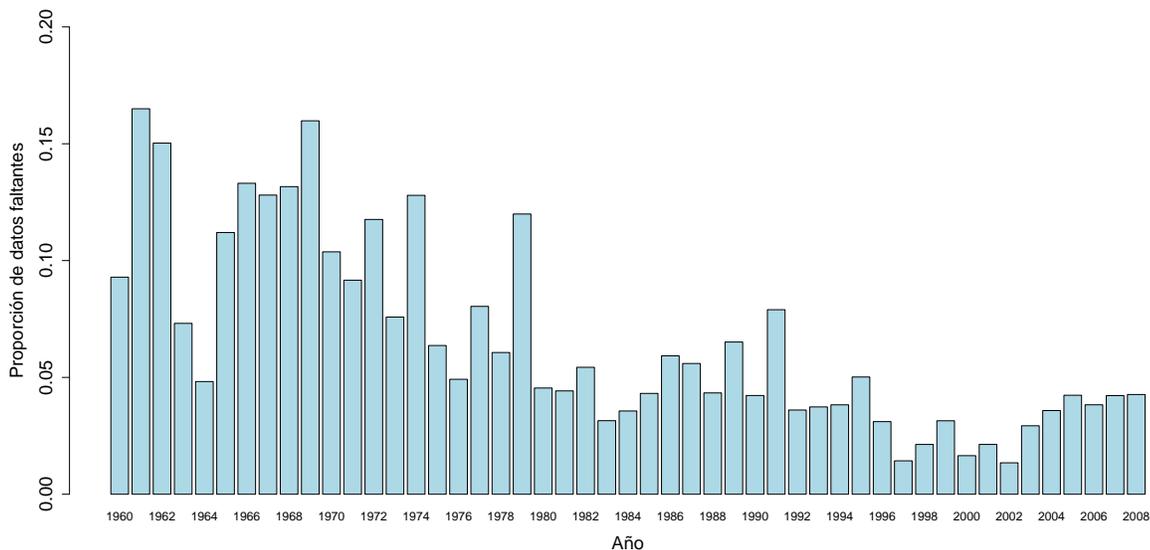


Figura 3.3: Proporción de datos faltantes por año.

Se construyó el mismo gráfico pero solo considerando las estaciones pertenecientes a la DNM (ver gráfico 3.4), se observó que hay solo dos años (1961 y 1979) donde el porcentaje

de datos faltantes supera el umbral de 10 %. Al examinar el gráfico, desde el año 1961 hasta el año 1972 hubo un marcado descenso en el porcentaje de datos faltantes, luego en el año 1973 hubo un gran incremento para nuevamente comenzar a descender hasta el año 1983 (a excepción del año 1979, donde hay un alto porcentaje de datos faltantes en relación a los demás años). El porcentaje de datos faltantes luego del año 1979 se situó en torno al 4 % o 5 % a excepción de los años desde el 1997 hasta 2002 donde este ronda el 2 % anual.

Se detectó que antes del año 1978 el porcentaje de datos faltantes es mucho menor para las estaciones de la DNM que para el total de las estaciones, esto se debe a que en las estaciones de UTE el porcentaje de datos faltantes es mucho mayor entre esos años. Luego el comportamiento de las dos gráficas es bastante similar, dándose un porcentaje de datos faltantes muy parecido.

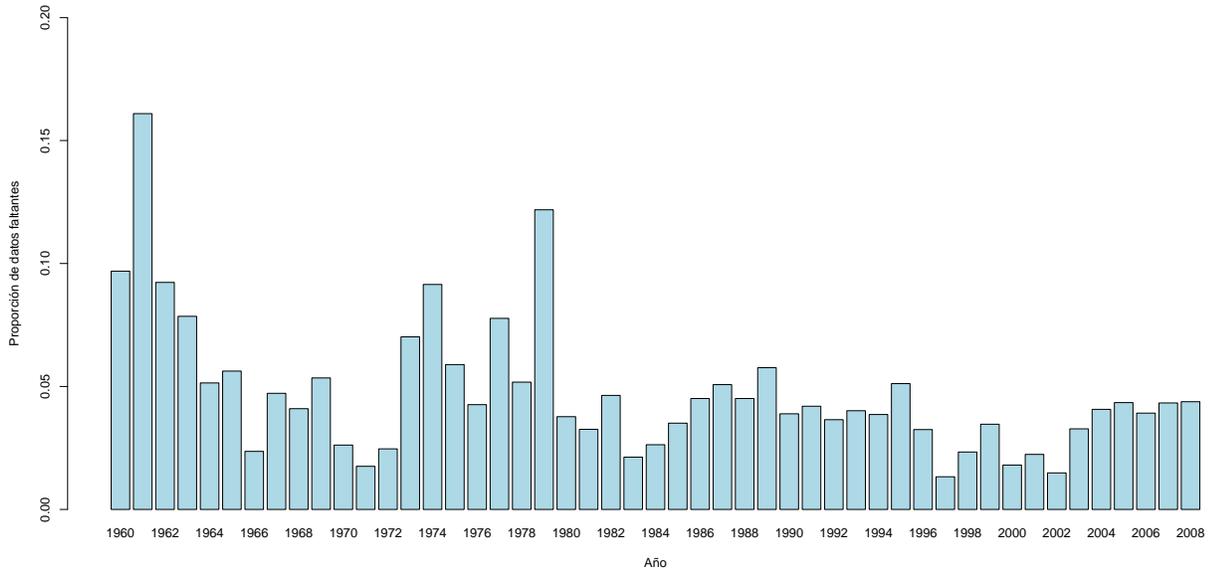


Figura 3.4: Proporción de datos faltantes por año para DNM y las meteorológicas.

También se realizó el análisis de datos faltantes por mes del año considerando todas las estaciones del período 1960-2008 (ver gráfico 3.5). A partir del gráfico se concluyó que no hay ningún mes que posea significativamente un porcentaje de datos faltantes más alto que los demás, en todos los meses el porcentaje promedio de datos faltantes está en el entorno del 6 % y 7 %.

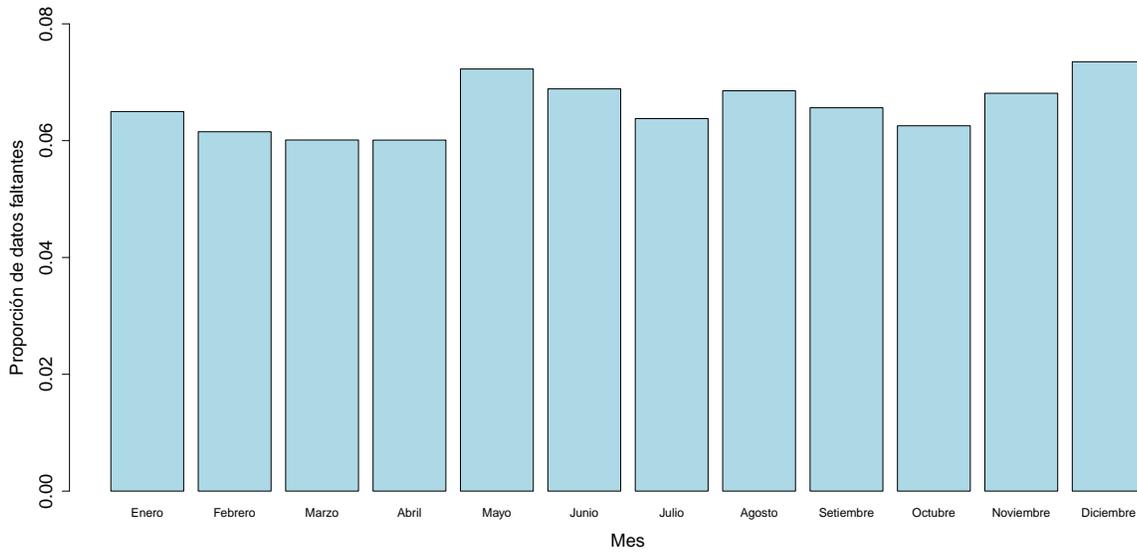


Figura 3.5: Proporción de datos faltantes por mes.

La evolución de la distribución mensual de los datos faltantes por década (ver gráfico 3.6) muestra que para la década del '60, en diciembre y mayo hay un porcentaje de datos faltantes bastante más alto que el resto de los meses, aproximadamente un 2% más, el cual podría estar acompañado de la presencia de vacaciones, sin embargo el mes de enero es un mes que está un poco por debajo de la media, lo cual nos llevó a concluir que no hay influencia de las vacaciones a través de los años en la presencia de datos faltantes.

A medida que pasan los años no sólo disminuye el porcentaje de datos faltantes lo cual concuerda con el gráfico de la figura 3.3, sino que también cambia un poco la estructura, de todos modos se podría asumir que el porcentaje de datos faltantes es uniforme en los meses.

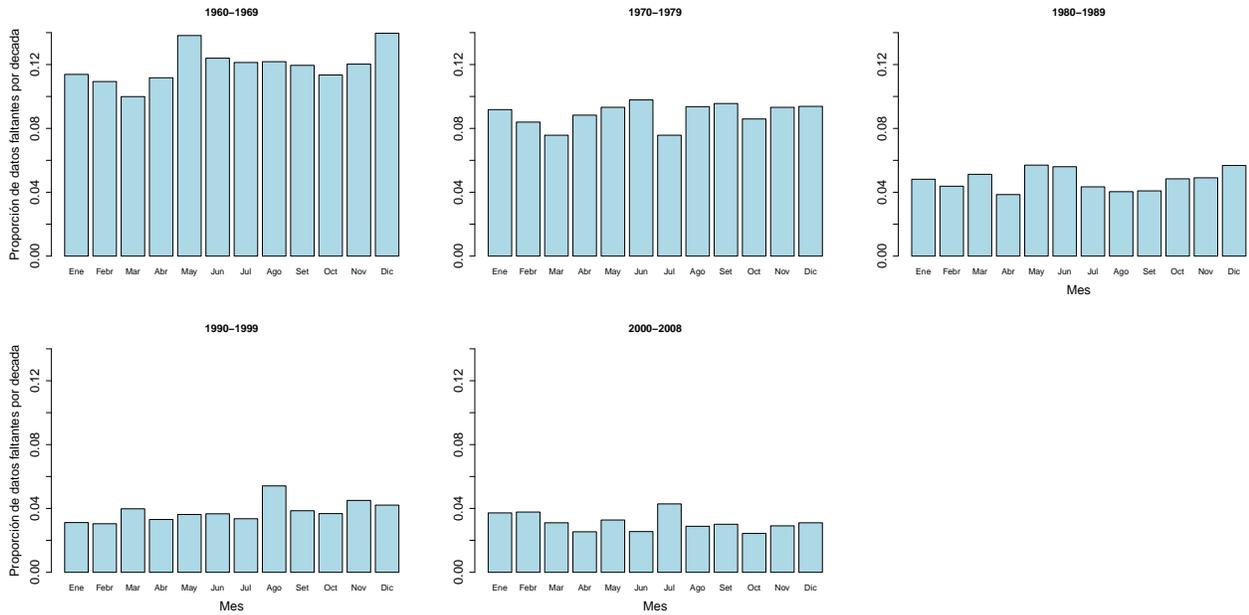


Figura 3.6: Evolución de la distribución de la proporción de los datos faltantes por mes en distintas décadas.

Se analizó la distribución del promedio de los datos faltantes según fuente de información (ver gráfico 3.7), y se detectó que las estaciones de UTE tienen una gran influencia sobre los datos faltantes, ya que el porcentaje de datos faltantes en estas estaciones se situó entre un 20% y un 25% por mes. Luego en las estaciones de la red pluviométrica de la DNM el porcentaje de datos faltantes mensuales fue aproximadamente de un 5% de los datos. En tercer lugar las meteorológicas presentaron datos faltantes que se sitúan entre un 2% y un 3%. Finalmente en las estaciones de INIA casi no hubo presencia de datos faltantes.

A priori se pensó que UTE es una mala fuente de información. Sin embargo, en estas estaciones tienen los registros de datos faltantes fueron diarios, en contraposición, en las estaciones red pluviométrica de la DNM y en las estaciones Meteorológicas, los registros de datos faltantes fueron mensuales y no diarios, por lo se planteó la hipótesis de que hay datos diarios que faltan en la DNM o en las Meteorológicas, pero que a priori fueron considerados como días sin lluvia.

Para las estaciones meteorológicas y de la red pluviométrica hubo una presencia de datos faltantes en diciembre y mayo sensiblemente superior. Sin embargo esto no ocurrió para las estaciones de las demás fuentes, ya que en las estaciones del INIA casi no se presenciaron datos faltantes, y en las de estaciones de UTE hubo una muy alta presencia de datos faltantes, pero no hubo ningún mes en el que se presente un cambio significativo respecto a los otros.

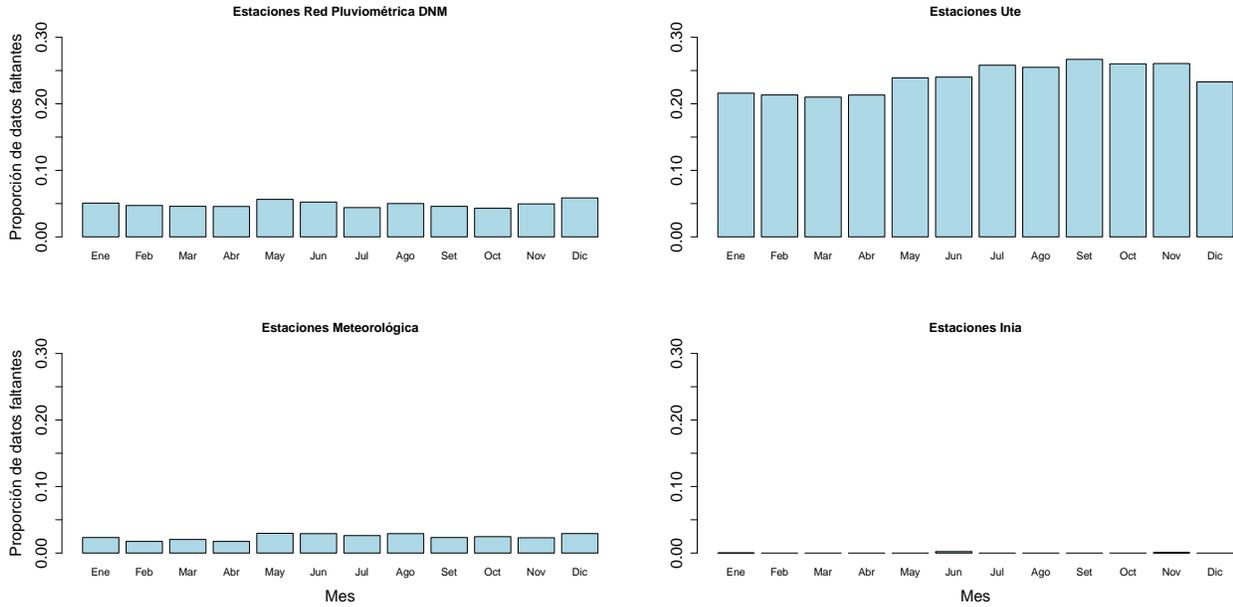


Figura 3.7: Proporción de datos faltantes por mes y fuente de información.

### 3.6. Conclusiones del capítulo.

En este capítulo se realizó una depuración de los datos. A partir de este análisis se pueden formular recomendaciones a la DNM acerca de como proceder al procesamiento de datos. Una primer consideración es que se debe utilizar un archivo secuencial con las fechas ya prefijadas para evitar errores en la generación del archivo calendario. Otro aspecto a considerar es que se debería lograr la uniformidad en la nomenclatura utilizada ya sea para considerar un dato como faltante, como también para considerar una traza.

También se realizó un análisis de datos faltantes, a partir del cual se concluyó en un primer lugar que las estaciones de UTE están muy comprometidas en relación a los datos faltantes que poseen, por otro lado las fuentes con menos datos faltantes son INIA y las estaciones meteorológicas de la DNM.

## Capítulo 4

# Reducción de la base de datos y distancias entre las estaciones.

En el capítulo dos se representaron las estaciones pluviométricas en el mapa, y en el capítulo tres se realizó un control de calidad de la base de datos y un análisis primario de los datos faltantes. En este capítulo se realizó una selección primaria de estaciones con aquellas estaciones que cumplieran con ciertos criterios. Esto fue de suma importancia ya que permitió diferenciar cuales estaciones son las de “mejor calidad” de las que son de “peor calidad” entre ellas cuales son prácticamente inutilizables. Finalmente se calcularán las distancias entre estaciones pluviométricas lo cual fue muy útil a la hora de detectar datos faltantes.

### 4.1. Selección de las estaciones pluviométricas.

#### 4.1.1. Por años de registro.

En el gráfico 4.1 se muestra la distribución geográfica de las estaciones pluviométricas según años de registro. En el trabajo realizado por Vicente-Serrano, et al.[12], se consideró que un período de registro menor de 15 años es muy corto para realizar un análisis climático, por lo tanto nosotros vamos a seguir el mismo criterio. De todos modos los períodos con dicha característica pueden ayudar a la detección datos faltantes en otras series.

Se observó que Lavallega tiene una densidad de pluviómetros aceptable en comparación a los otros departamentos, sin embargo los períodos de las series son muy cortos lo que dificulta cualquier posible análisis. Además se observó que las series más cortas se encuentran al oeste del departamento, región contigua a Florida que tiene solamente dos estaciones. Una parte de la información de Lavallega se pensaba usar para mejorar la calidad de los registros pluviométricos de las estaciones del departamento de Florida.

La mayoría de las estaciones cubren el período entre 31 y 50 años (color rojo en el gráfico 4.1), esto es esperable ya que la mayor parte de las estaciones son pertenecientes a la DNM y a este organismo le pedimos el período 1960-2008 (49 años). Mientras que las estaciones de más de 50 años (color verde en el gráfico 4.1) son muy pocas, estas son todas pertenecientes a UTE y se encuentran concentradas en la región nordeste del país.

Finalmente se decidió acotar el período de las estaciones de UTE al período 1960-2008 a modo de poder compararlas con las demás estaciones.

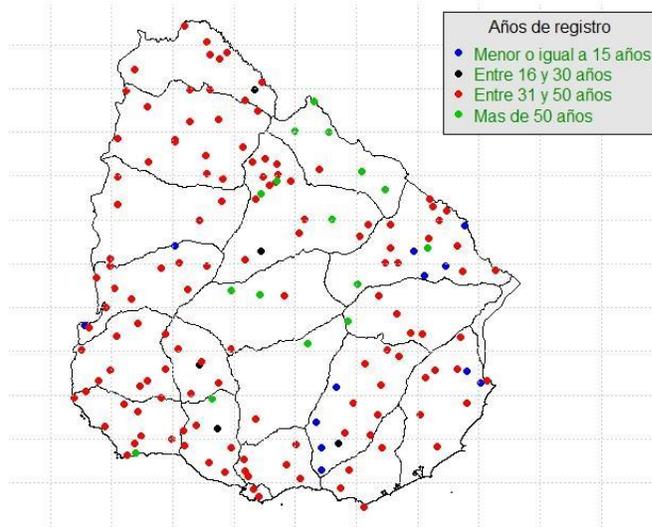


Figura 4.1: Distribución geográfica de los puntos en el mapa de Uruguay según años de registro.

#### 4.1.2. Por datos faltantes.

En esta sección se estudió la distribución geográfica de los puntos según porcentaje de datos faltantes (ver figura 4.2) donde se seleccionó aquellas estaciones que no sobrepasan un cierto umbral.

Para este estudio se seleccionó un umbral de 10%. El mismo representa casi cinco años de datos faltantes en 49 años de registro (nuestro período es 1960-2008). Este umbral fue elegido así por un lado porque más de un 10% de datos faltantes es un porcentaje muy alto como para utilizar dicha estación pluviométrica. Por otro lado si se considera un umbral de un 5% se pierden a priori estaciones pertenecientes a estaciones meteorológicas de la DNM, la cual se consideró como una buena fuente de información.

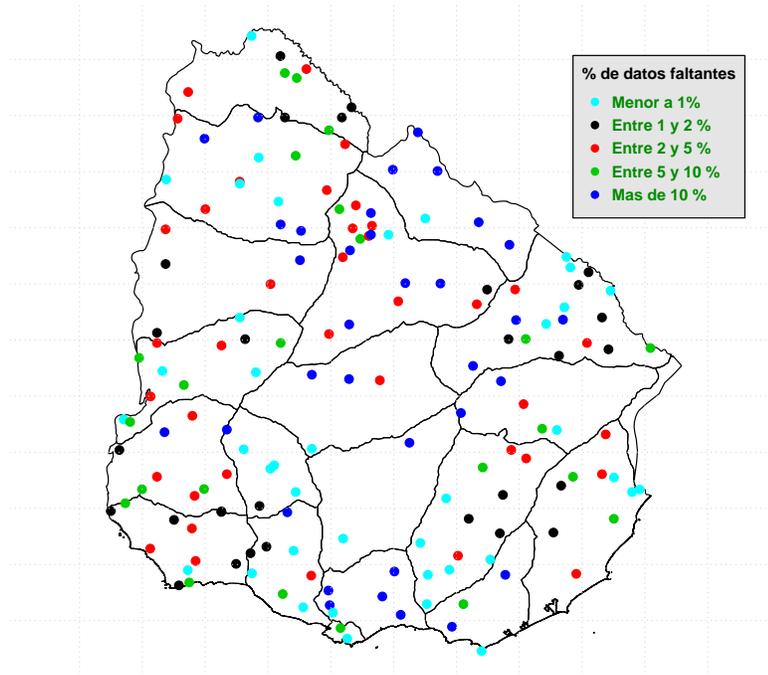


Figura 4.2: Distribución geográfica de los puntos según porcentaje de datos faltantes.

En la tabla 4.1 se muestran los porcentajes de datos faltantes según fuente de información. A partir de ella se observó que todas las estaciones de UTE poseen más de un 10% de datos faltantes, por otro lado tanto las estaciones INIA como Meteorológicas de DNM presentaron puntos con menor porcentaje de datos faltantes que las pluviométricas de DNM. Otro aspecto que se obtiene de la tabla es que hay 34 puntos con más de un 10% de datos faltantes, lo que representa el 21% del total de las estaciones.

	Menor a 1%	Entre 1 y 2%	Entre 2 y 5%	Entre 5 y 10%	Mas de 10%
Pluviométricas DNM	29	28	38	21	19
INIA	5	0	0	0	0
Meteorológicas DNM	5	1	0	2	0
UTE	0	0	0	0	15

Cuadro 4.1: Clasificación de las estaciones por datos faltantes según fuente de información.

El gráfico 4.3 presenta los puntos donde el porcentaje de dato faltante es menor a un 10%. Se observó que los departamentos de Rivera, Canelones, Paysandú, Florida, Durazno, Treinta y Tres y Maldonado presentan serios problemas en cuanto a la cantidad de estaciones pluviométricas con menos de un 10% de datos faltantes. Mientras que Tacuarembó se encuentra comprometida respecto a la distribución espacial de las estaciones pluviométricas.

Finalmente un aspecto a destacar es que hubo regiones como los departamentos de Rivera

y Durazno que tenían varias estaciones con un largo de la serie mayor a 50 años, un largo más que suficiente para realizar un análisis de precipitación, sin embargo su porcentaje de datos faltantes fue superior al 10% lo cual las convirtió en estaciones prácticamente inutilizables.

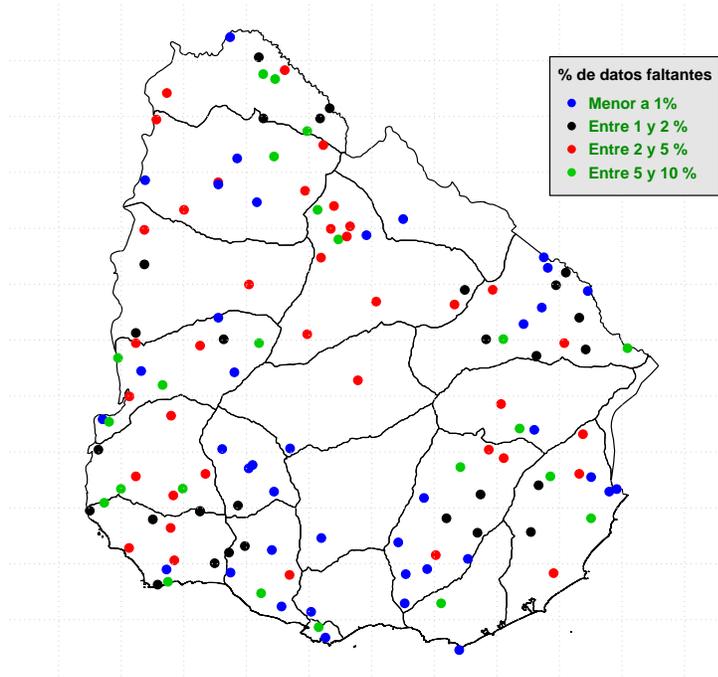


Figura 4.3: Distribución geográfica de los puntos según porcentaje de datos faltantes quitando aquellas estaciones con más de 10% de datos faltantes.

#### 4.1.3. Por porcentaje de días sin lluvia sobre el total de registros.

Cuando no hubo registro de precipitación a priori se supuso que no llovió. Sin embargo si una estación “*x*” tiene un alto porcentaje de días sin lluvia, esto se puede deber a que en realidad cierta parte de los datos faltantes podría estar enmascarado en los días sin lluvia.

El cálculo del porcentaje de días sin lluvia es sobre el total de días con registro, es decir que no se tomaron en cuenta datos faltantes. En la tabla 4.2 se muestran los porcentajes de días sin lluvia según la fuente de información.

	Menor a 75 %	Entre 75 y 80 %	Entre 80 y 85 %	Más de 85 %
Pluviométricas DNM	0	11	117	7
INIA	2	3	0	0
Meteorológicas DNM	7	0	1	0
UTE	0	3	10	2

Cuadro 4.2: Clasificación de las estaciones por porcentaje de días sin lluvia en la serie según fuente de información.

A partir de la tabla, se consideró demasiado alto un porcentaje de más del 85 % de días sin lluvia (umbral) ya que existen estaciones donde el registro de días sin lluvia es un 15 % menos (ver sección 3.4). Por lo tanto solo se consideró para detección de datos faltantes estaciones que no sobrepasen este umbral. La mayoría de estas pertenecen a las estaciones pluviométricas de DNM.

Observando la tabla tenemos que los porcentajes de días sin lluvia para los pluviómetros del INIA y para los de los de estaciones meteorológicas de la DNM son en su gran mayoría inferiores al 75 %, lo que representa menos de 23 días lluviosos por mes. Por otro lado los porcentajes de días sin lluvia para los pluviómetros de la red pluviométrica de la DNM y los de UTE en su gran mayoría se encuentran entre 80 % y 85 % lo que representa entre 24 y 26 días sin lluvia promedio por mes. Esto muestra una clara presencia de datos faltantes ocultos entre los días sin lluvia tanto para estaciones de la red pluviométrica de la DNM como para los de UTE.

En el gráfico 4.4 se muestra la distribución geográfica del porcentaje de días sin lluvia, donde se evidencia que la región del norte de Rocha posee estaciones que tienen un buen largo de serie, un porcentaje de datos faltantes aceptable, pero un porcentaje de días sin lluvia superior al 85 %, lo que hace a estas estaciones altamente dudosas.

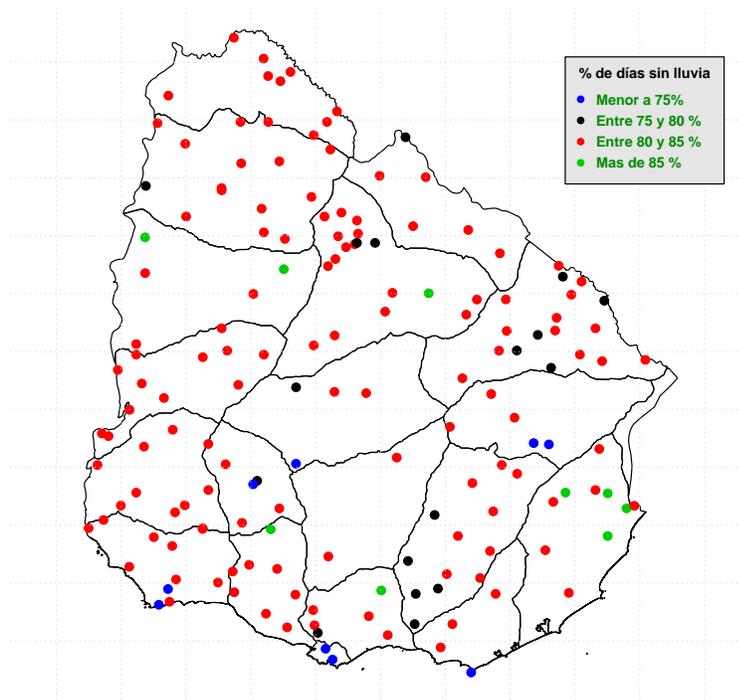


Figura 4.4: Distribución geográfica de los puntos según porcentaje de días sin lluvia.

#### 4.1.4. Selección final de estaciones pluviométricas.

Considerando los tres tipos de selección como válidos, las estaciones seleccionadas se redujeron de 163 a 114 estaciones lo que constituye un 70 % de las estaciones pluviométricas. La representación final de los puntos se muestra en el gráfico 4.5.

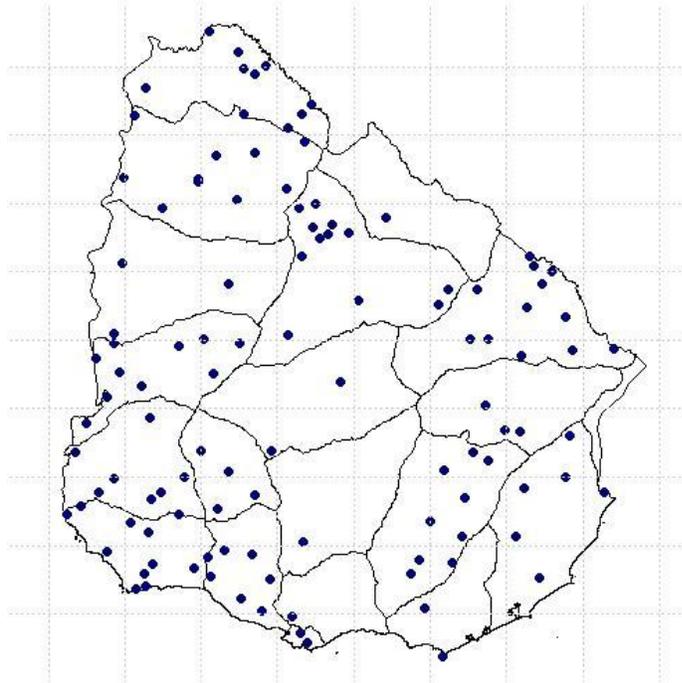


Figura 4.5: Distribución geográfica de los puntos luego del depurado.

En la tabla 4.3 se muestra de qué fuente de información se quitaron las series. En la selección de las series fueron escogidas todas las estaciones del INIA y las meteorológicas de la DNM. Por otro lado se quitó el 25 % de las estaciones de DNM y todas las estaciones de UTE.

	Pluviométricas DNM	INIA	Meteorológicas DNM	UTE
Antes de la selección	135	5	8	15
Después de la selección	101	5	8	0

Cuadro 4.3: Clasificación de las estaciones por fuente de información, antes y después de la selección de las estaciones.

## 4.2. Distancias entre estaciones pluviométricas.

Con la selección final de estaciones se calculó la distancia entre ellas, en esta sección se detalló la metodología utilizada a partir de las coordenadas geográficas. Este cálculo fue muy útil ya que para la detección de datos faltantes en las series pluviométricas se suele mirar series cercanas.

Luego de un correcto tratamiento de las coordenadas en el capítulo dos, se procedió a calcular las distancias entre los puntos.

No se tomaron en cuenta diferencias de altura entre estaciones ya que las mismas no son significativas en Uruguay, debido a que su relieve no presenta grandes cambios.

Se consideró que la superficie de la tierra no es un plano recto por lo que en todos los cálculos de distancia se tuvo en cuenta el arco que describe el recorrido entre dos puntos.

Además este tipo de cálculo omitió el hecho de que la Tierra no es perfectamente esférica lo que haría aún más complejo el mismo. De todos modos, esta suposición no afectó el resultado final, ya que estudios anteriores estimaron que el error es de aproximadamente un kilómetro cada 12000 kilómetros o sea un poco más de media cuadra en 650 Km, que es la distancia máxima que se podría considerar en Uruguay.

Por lo tanto utilizando conocimientos de trigonometría y considerando por un lado el radio de la tierra igual a 6367km, por otro el ángulo en radianes formado por la ubicación de 2 puntos respecto al centro de la tierra, tenemos que la distancia entre 2 puntos se calcula como:

$$d(a, b) = \text{ángulo}(a, b) \times \text{radio terrestre}$$

Al aplicar esta sencilla formula concluimos que el promedio de distancias a la estación más cercana (vecino más cercano) es de 22.69 km., aunque la mediana se sitúa en 21.34 km., con un desvío estándar de 11.96 km. Hay cuatro estaciones cuya distancia a la estación más cercana supera los 50 km., lo cual indica que estas serán difíciles de estudiar, ya que no tendrán estaciones cercanas de comparación. Estas son la 1650 de Tacuarembó cuya distancia a la estación más cercana es de 61.93 km., la 1920 de Durazno cuya distancia a la estación más cercana es de 59.91 km., y luego hay dos estaciones pertenecientes al departamento de Paysandú, estas son la 1502 y la 1553 cuyas distancias a la estación más cercana es de 59.22 km. y 52.09 km. respectivamente

Se construyó una matriz cuadrada que contiene en las filas y en las columnas las estaciones pluviométricas. El elemento en la posición  $(i, j)$  de la matriz representa la distancia entre la estación que corresponde a la fila  $i$ -ésima y la que corresponde a la columna  $j$ -ésima. El uso de esta matriz reduce el tiempo máquina para algunos cálculos posteriores. Está matriz se llama matriz de distancias. En la tabla 5.1 de la sección 5.2 se encuentra una submatriz de distancias igual a la construida aquí pero solo para el departamento de Colonia.

## Capítulo 5

# Caso de estudio: departamento de Colonia.

Si bien se contó con una gran cantidad de puntos, como se observó en el capítulo cuatro, no existe una buena densidad de pluviómetros en algunas zonas del país (la densidad de pluviómetros por departamento medida como la cantidad de pluviómetros por mil  $km^2$ ). Además no todas las series seleccionadas presentan la misma calidad (donde la calidad de pluviómetros es un resultado derivado de observar la cantidad de datos faltantes respecto del total de años de registro de la estación, etc.). Teniendo en cuenta estos aspectos, se ha seleccionado un departamento para realizar el análisis de los datos, que posteriormente podría ser extendido a todo el Uruguay.

### 5.1. ¿Por qué se eligió Colonia?

Se realizó una tabla donde se especifica para cada departamento la cantidad de pluviómetros que posee el mismo así como también la cantidad de pluviómetros que registraron datos en cada una de las décadas desde 1960 al 1999 y del 2000 al 2008 (ver tabla 5.1).

Departamento	Superficie ( $km^2$ )	% territorial	Estaciones	1960-1969	1970-1979	1980-1989	1990-1999	2000-2008
Artigas	11928	6.82	10	10	9	9	9	9
Canelones	4536	2.59	1	0	0	1	1	1
Cerro Largo	13648	7.80	12	12	12	12	12	12
Colonia	6106	3.49	11	10	11	11	11	11
Durazno	11643	6.65	2	2	2	2	2	2
Flores	5144	2.94	4	3	3	4	4	4
Florida	10417	5.95	1	1	1	1	1	1
Lavalleja	10016	5.72	8	5	5	8	8	5
Maldonado	4793	2.74	2	1	2	2	2	2
Montevideo	530	0.30	2	2	2	2	2	2
Paysandú	13922	7.95	5	2	5	5	5	2
Río Negro	9282	5.30	9	6	7	9	9	7
Rivera	9370	5.35	1	0	0	1	1	0
Rocha	10551	6.03	6	5	5	6	6	4
Salto	14163	8.09	10	7	8	10	10	8
San José	4992	2.85	6	4	4	6	6	5
Soriano	9008	5.15	9	4	9	9	9	9
Tacuarembó	15438	8.82	11	8	10	11	11	10
Treinta y Tres	9529	5.44	4	0	3	4	4	4

Cuadro 5.1: Número de estaciones en cada departamento por década.

Esta tabla arrojó el resultado evidente que Montevideo es el departamento con mayor densidad de pluviómetros, ya que en la capital del país hay dos estaciones meteorológicas, y la superficie del mismo solo contiene el 0.30% de la superficie total del territorio. Sin embargo, debido a su pequeña superficie territorial, no se incluyó solo a Montevideo para hacer un análisis preliminar de Uruguay. Alternativamente se optó por unir Montevideo con Canelones y realizando esto se obtuvo que Colonia es el mejor departamento en los dos sentidos anteriormente expuestos (ver figura 5.1), por lo tanto se eligió este departamento para un análisis inicial.

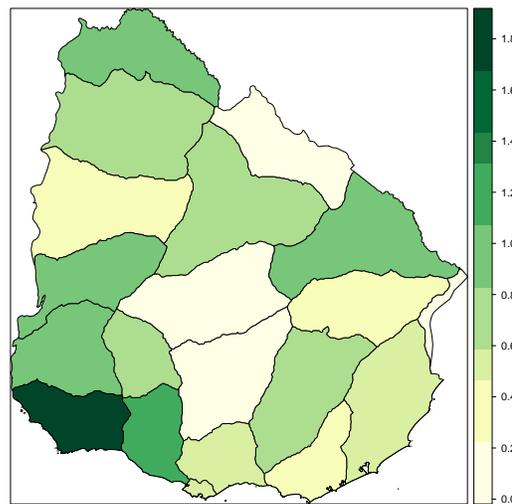


Figura 5.1: Densidad de pluviómetros en cada departamento, observar que Montevideo y Canelones están unidos.

## 5.2. Descripción de las estaciones pluviométricas.

Una vez elegido el departamento de Colonia se construyó una sub base de datos con todas las estaciones pluviométricas del mismo en formato .txt que es más amigable para el R desde el punto de vista de la lectura del mismo.

Se redujo la matriz de distancias calculada en la sección 4.2 a una matriz de distancias entre las estaciones de Colonia (ver tabla 5.2).

Estación	2471	2476	2480C	2570	2611	2657A	2662A	2702	2737	Met. 2774	La Estanz.
2471	0.00	54.83	94.87	71.23	47.01	84.64	125.20	116.77	91.54	87.12	98.27
2476	54.83	0.00	41.13	17.05	32.30	40.28	71.63	65.53	56.12	57.27	46.99
2480C	94.87	41.13	0.00	29.03	68.79	48.11	43.84	47.07	67.38	73.62	36.16
2570	71.23	17.05	29.03	0.00	39.79	28.51	54.59	48.82	47.17	50.87	30.53
2611	47.01	32.30	68.79	39.79	0.00	40.25	85.86	74.62	44.55	40.37	57.41
2657A	84.64	40.28	48.11	28.51	40.25	0.00	47.44	34.71	19.55	25.62	19.67
2662A	125.20	71.63	43.84	54.59	85.86	47.44	0.00	15.45	58.23	67.27	28.49
2702	116.77	65.53	47.07	48.82	74.62	34.71	15.45	0.00	43.16	52.28	18.67
2737	91.54	56.12	67.38	47.17	44.55	19.55	58.23	43.16	0.00	9.16	34.97
Met. 2774	87.12	57.27	73.62	50.87	40.37	25.62	67.27	52.28	9.16	0.00	43.09
La Estanz.	98.27	46.99	36.16	30.53	57.41	19.67	28.49	18.67	34.97	43.09	0.00

Cuadro 5.2: Matriz de distancias entre estaciones del departamento de Colonia.

Para este departamento se tienen once estaciones pluviométricas de las cuales una es estación de INIA (llamada La Estanzuela), una estación meteorológica de la DNM (a partir de aquí estación meteorológica) y las demás son pertenecientes a la red pluviométrica de la DNM (a partir de aquí estaciones pluviométricas, ya que las otras son ambas estaciones meteorológicas). A excepción de la estación de La Estanzuela que cubre el período 1965-2008 las demás estaciones cubren el período 1960-2008, por lo que se utilizarán todos los años en cuestión en lugar de reducir el largo de las series a aquellos años donde la densidad de estaciones es mayor en el país.

En la figura 5.2 se muestra la distribución espacial de las estaciones en el departamento de Colonia, donde se muestra que las estaciones están bien distribuidas en el departamento, más allá de alguna región con concentración de estaciones como el suroeste del departamento, y por otro lado regiones como el nordeste o noroeste donde solamente hay una estación.

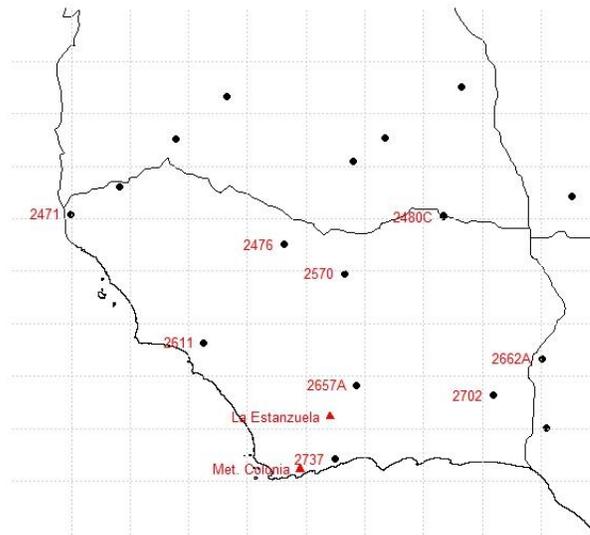


Figura 5.2: Distribución geográfica de las estaciones pluviométricas en el departamento de Colonia.

En la tabla 5.3 se muestran las características observadas en la tabla descrita en la sección 3.4, pero solamente considerando las estaciones de Colonia. A partir de ella se extrajo las siguientes conclusiones.

En primer lugar las estaciones analizadas tienen menos de un 2.5% de datos faltantes, exceptuando las estaciones pluviométricas 2737 y la 2657A que tienen más.

Luego el porcentaje de días sin lluvia presentó diferencias entre las estaciones de La Estanzuela y meteorológica donde el porcentaje de días sin lluvia ronda el 73%, respecto a las estaciones pluviométricas donde el porcentaje de días sin lluvia es aproximadamente un 80%.

En tercer lugar se observó que el día 31 de mayo de 1985 fue un día especial ya que cinco de las once series registraron su máximo de precipitación en esa fecha, el valor registrado fue diferente para las distintas estaciones, este valor fue desde entre 140mm registrados en la estación pluviométrica 2662A, hasta 220.10mm en la estación meteorológica, además todas estas estaciones están en el centro y sur de Colonia.

Finalmente la máxima cantidad de precipitación registrada en cinco días consecutivos varió entre 256mm registrados en la estación pluviométrica 2570 y 359mm en la estación pluviométrica 2611, estos máximos se registraron básicamente en los años 1966, 1988 y 2007.

Estación	Período	Nro. de D.F.	Nro. de S.L.	% de D.F.	% de días SL	Máximo	Fecha	Máximo 5 días	Año
2471	1960-2008	305	14727	1.70	82.28	160.00	24/02/1973	338.00	1966
2476	1960-2008	306	14597	1.71	81.56	150.00	12/05/1981	294.00	1966
2480C	1960-2008	213	14554	1.19	81.32	235.00	10/02/2003	349.00	2003
2570	1960-2008	395	14337	2.21	80.10	213.00	13/01/1976	256.00	2007
2611	1960-2008	398	14622	2.22	81.70	192.00	31/05/1985	359.00	2007
2657A	1960-2008	857	13868	4.79	77.48	140.00	18/12/2002	328.00	1988
2662A	1960-2008	273	14464	1.53	80.81	140.00	31/05/1985	270.00	2007
2702	1960-2008	235	14400	1.32	80.87	162.00	31/05/1985	250.00	1966
2737	1960-2008	1060	14205	5.80	77.78	195.00	31/05/1985	240.00	2001
Met. 2774	1960-2008	229	13070	1.28	73.02	220.10	31/05/1985	228.70	1988
La Estanz.	1965-2008	0	11554	0.00	72.71	173.00	17/03/1966	260.20	1966

Cuadro 5.3: Características de las estaciones del departamento de Colonia, para esta tabla D.F.=Datos Faltantes y S.L.=Sin Lluvia., Máximo y Fecha= Cantidad de precipitación máxima registrada en 1 día y la fecha de ocurrencia, Máximo 5 días y Año= Máxima precipitación acumulada registrada en 5 días y el año de ocurrencia.

### 5.3. Datos faltantes en Colonia.

En esta sección se realizó un análisis de datos faltantes similar al de la sección 3.5 pero no tan exhaustivo, ya que en este caso la finalidad fue comparar como se comporta Colonia en este aspecto respecto a los demás departamentos. Además se estudiará la evolución temporal del porcentaje de datos faltantes.

En una primer instancia se graficó el porcentaje de datos faltantes por año en promedio (ver gráfico 5.3). La comparación de este gráfico con el gráfico 3.4, que refiere a la totalidad de las estaciones del país pertenecientes a la DNM (mas allá que en Colonia haya una estación perteneciente al INIA), muestra que el porcentaje de datos faltantes es significativamente menor en las estaciones de Colonia respecto la totalidad de las de DNM, a excepción de los años 1960, 1976 y 1979 donde el porcentaje de datos faltantes es mayor en Colonia. Además en los periodos de 1970-1973, 1980-1986 y 1999-2005 no hay datos faltantes para el departamento de Colonia.

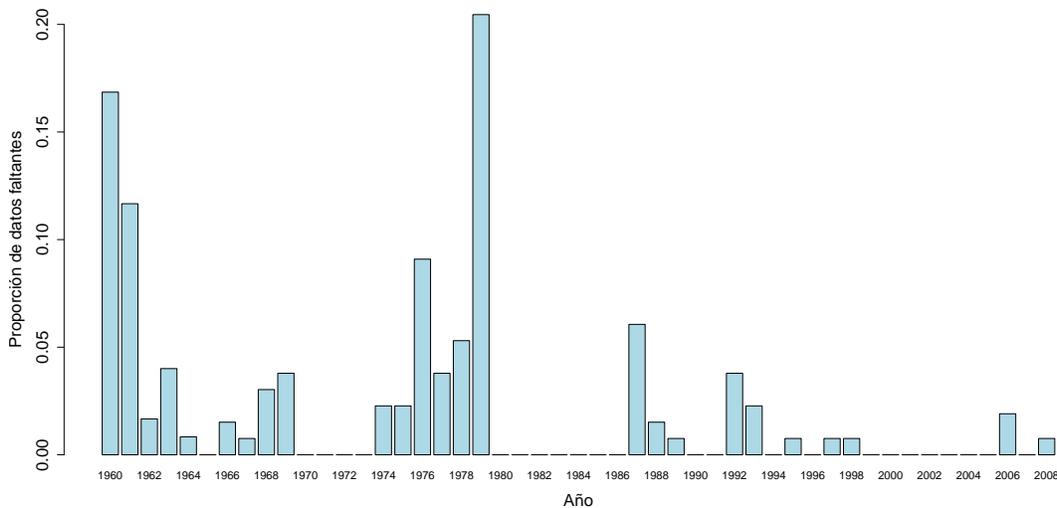


Figura 5.3: Proporción anual de porcentaje de datos faltantes para Colonia en el periodo 1960-2008.

Luego se construyó el gráfico de porcentaje de datos faltantes pero de forma mensual

(ver gráfico 5.4). El promedio de porcentaje de datos faltantes va en aumento desde el mes de diciembre hasta el mes de mayo donde alcanzó el máximo de 5 % de datos faltantes, luego entre junio y noviembre tiene un comportamiento bastante irregular. Comparando con el mismo tipo de gráfico pero referido a todas las estaciones del país pertenecientes a la DNM (ver gráfico 3.7) se visualiza que para todos los meses el porcentaje de datos faltantes es entre un 1 % y un 3 % menor en Colonia.

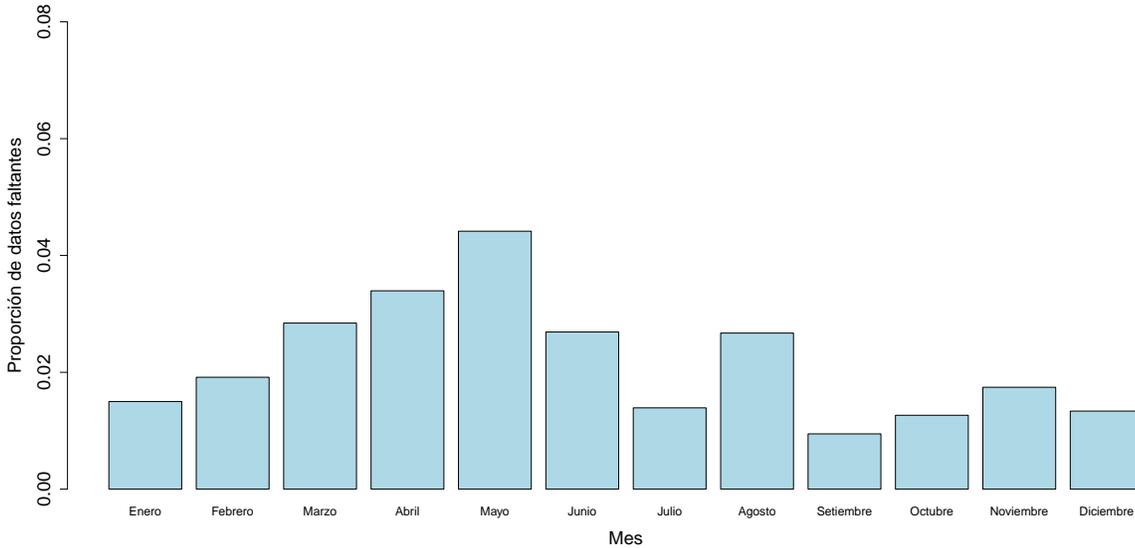


Figura 5.4: Proporción mensual de porcentaje de datos faltantes para Colonia en el periodo 1960-2008.

Finalmente se graficó la evolución de los datos faltantes a través de las décadas (ver figura 5.5). A partir de este gráfico se observó que la distribución de los datos faltantes presentó niveles similares en las décadas del 60 y 70, sin embargo cambió el lugar de acumulación de masa, en la década del 60 se acumuló en los meses desde abril a agosto (entorno al 7% a excepción de julio que ronda el 5%), y en la década del 70 se acumuló desde marzo a mayo aunque con niveles mayores (entorno al 9%), posteriormente descendió en la década del 80, y permaneció más o menos similar en las décadas restantes.

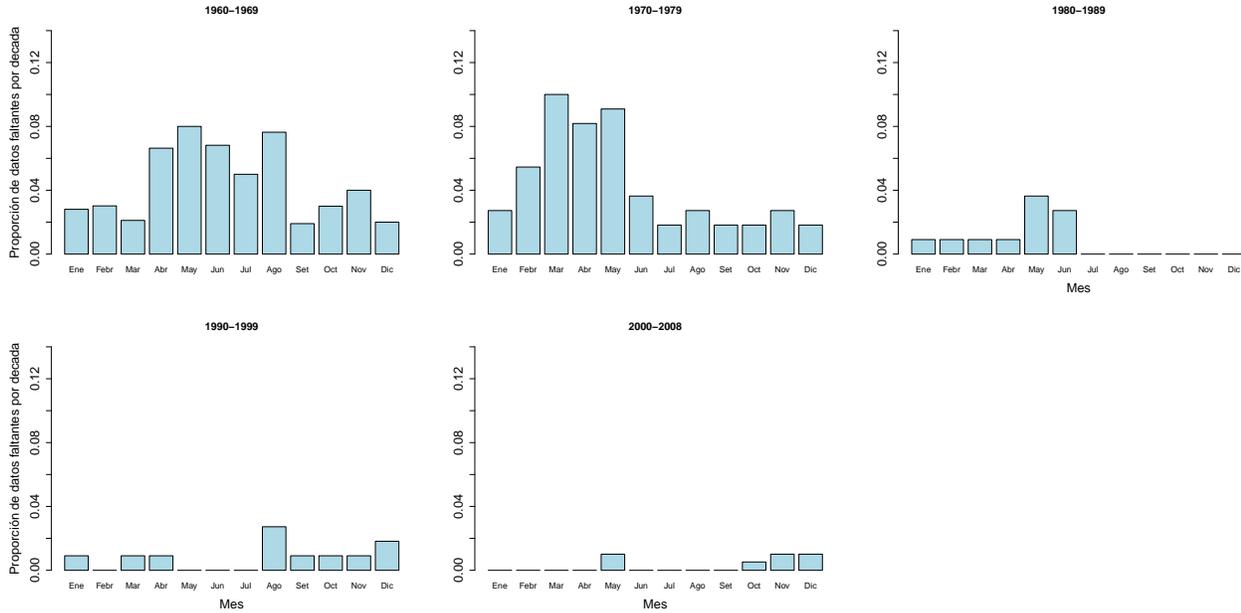


Figura 5.5: Evolución de la distribución de la proporción de los datos faltantes por mes en distintas décadas, para Colonia en el periodo 1960-2008.

Gracias a este análisis de datos faltantes se concluyó<sup>321</sup> que fue adecuado elegir Colonia como caso de estudio preliminar.

## 5.4. Estudio de correlaciones y relación con las distancias entre estaciones.

### 5.4.1. Marco teórico.

En esta sección se desarrolló el concepto de correlación, especialmente la correlación muestral y entre ellas se detalló la correlación de Pearson y Spearman.

Dadas dos variables aleatorias reales  $X$  e  $Y$  el coeficiente poblacional de correlación de Pearson viene dado por la siguiente ecuación:

$$\rho_{XY} = \frac{E((X - \mu_X) \cdot (Y - \mu_Y))}{\sigma_X \cdot \sigma_Y}$$

donde  $\sigma_X = \sqrt{\text{var}(X)}$ . El coeficiente de correlación ( $\rho_{XY}$ ) toma valores en el rango  $[-1,1]$ .

Dicho coeficiente es independiente de la escala de medida mide la tendencia lineal entre  $X$  e  $Y$  para covariar. Cuanto más próximo a 1 (en valor absoluto) es este coeficiente, la relación lineal entre  $X$  e  $Y$  es mayor. Note que cuando  $\rho_{XY} = 0$  no necesariamente las variables son independientes, esto si ocurre bajo distribución normal de  $X$  e  $Y$ , por lo tanto fijando la hipótesis adicional de que  $X$  e  $Y$  tengan distribución conjunta normal bivariada se puede concluir la independencia.

Al hacer inferencia sobre la correlación, cuando la muestra  $\{(X_i, Y_i)^T : i = 1, \dots, n\}$  consiste en variables aleatorias independientes e idénticamente distribuidas, un estimador del coeficiente de correlación es:

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^n ((X_i - \bar{X}) \cdot (Y_i - \bar{Y}))}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Este es el estimador máximo verosímil para la correlación cuando el vector (X,Y) tiene distribución normal bivariada. Además, este estimador tiene una distribución asintóticamente normal con media  $\rho_{XY}$  y varianza  $\frac{(1-\rho_{XY})^2}{n}$  donde n es la cantidad de datos. De todas formas, debido a que  $\hat{\rho}_{XY}$  converge de forma lenta a la distribución asintótica, este resultado no es usado generalmente para hacer inferencia acerca de  $\rho_{XY}$ . En lugar de esto, se utiliza la inferencia basada en la transformación z de Fisher :

$$z = \frac{1}{2} \ln \left( \frac{1 + \hat{\rho}_{XY}}{1 - \hat{\rho}_{XY}} \right).$$

La misma converge rápidamente a una distribución  $\mathcal{N}(\frac{1}{2} \ln \left( \frac{1 + \rho_{XY}}{1 - \rho_{XY}} \right), \frac{1}{n-3})$  cuando  $\rho_{XY}$  es distinto de cero. De esta manera se encuentra fácilmente un intervalo de confianza (IC) para la transformación z de Fisher. Deshaciendo la misma se encuentra el IC para  $\rho_{XY}$ .

Una prueba aproximada para la hipótesis  $H_0) \rho_{XY} = 0$  se realiza calculando:

$$T = |\hat{\rho}_{XY}| \sqrt{\frac{n-2}{1 - \hat{\rho}_{XY}^2}} \sim t_{n-2}$$

donde  $t_{n-2}$  representa una distribución t student con  $n-2$  grados de libertad. El intervalo de confianza y la prueba anterior requieren el supuesto de normalidad.

Para hacer inferencia sobre  $\rho_{XY}$  se puede emplear alternativamente un enfoque no paramétrico basado en rangos de las observaciones. La muestra  $\{(X_i, Y_i)^T : i = 1, \dots, n\}$  es reemplazada por la correspondiente muestra de rangos  $\{(R_{X_i}, R_{Y_i})^T : i = 1, \dots, n\}$  donde  $R_{X_i}$  es el rango de  $X_i$ , y  $R_{Y_i}$  es el rango de  $Y_i$ . La dependencia entre X e Y es estimada con el *coeficiente de correlación de Spearman*  $\rho_{XY}^S$ :

$$\hat{\rho}_{XY}^S = \frac{\sum_{i=1}^n R_{Y_i} \cdot R_{X_i} - N}{\sqrt{(\sum_{i=1}^n R_{Y_i}^2 - N) \cdot (\sum_{i=1}^n R_{X_i}^2 - N)}}$$

donde:  $N = \left(\frac{n+1}{2}\right)^2$

Este coeficiente de correlación muestral está basado en rangos y también tiene la ventaja de recorrer valores entre [-1,1].

En nuestro caso las muestras son grandes ( $n \geq 30$ ), por lo que para poner a prueba

$$H_0) \rho_{XY} = 0$$

vs

$$H_1) \rho_{XY} \neq 0$$

con un valor de significación de  $(1 - \alpha) \cdot 100\%$  (o a nivel  $\alpha$ ) tenemos el siguiente IC:  $\pm Z_{1+\frac{\alpha}{2}} \sqrt{n-1}$  donde  $Z_{1-\frac{\alpha}{2}}$  es el cuantil  $(1 - \frac{\alpha}{2})$  de la distribución normal estándar. Lógicamente valores críticos para pruebas de una sola dirección son análogos.

En nuestro caso calcular el coeficiente de correlación de Spearman no es útil ya que tenemos una gran cantidad de empates (los ceros de no lluvia son cerca del 75% de los datos de la serie) lo cual lleva a no poder calcular el rango de forma efectiva, Al calcular la correlación entre estaciones elegimos el coeficiente de correlación de Pearson, aunque los datos tengan que ser independientes para que el estimador sea de máxima verosimilitud.

### 5.4.2. Distancias entre estaciones.

Para la detección de los datos faltantes de una estación objetivo, usar información de estaciones cercanas puede ser de suma utilidad. Por eso fue importante estudiar las distancias entre las estaciones de Colonia.

Se realizaron boxplot de distancias según estación en Colonia (ver gráfico 5.6). Se detectó que la estación en la cual se tendrán mayores dificultades para detectar datos faltantes desde el punto de vista de las distancias es la estación 2471 situada al noroeste de Colonia, ya que las distancias de la misma a las demás estaciones son mayores que tomando como referencia cualquier otra estación. En contraposición en las estaciones 2657A, 2570 y La Estanzuela, las distancias a las demás estaciones son menores en comparación al resto.

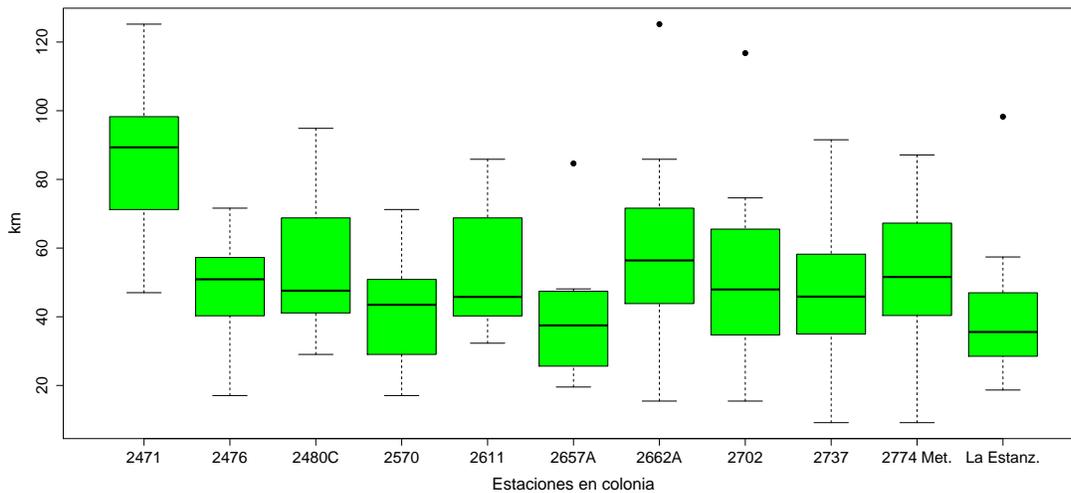


Figura 5.6: Boxplot de distancias en Km. según estación en Colonia.

### 5.4.3. Estudio de las correlaciones.

Se estudió la correlación de precipitaciones entre estaciones y se relacionó con la distancia, para eso se utilizaron tres medidas de interés:

- Correlación de Pearson mensual, se realizó a partir de lo definido en la sección 5.4.1, tomando las series como mensuales.
- Correlación de Pearson diaria, se realizó a partir de lo definido en la sección 5.4.1, tomando las series como diarias.
- Concordancia diaria entre estaciones. Se realizó a partir de lo definido en la sección 3.3.

Al igual que en la sección 3.3, es necesario mencionar que la cantidad de datos faltantes no afecta la estructura de correlación, ya que se consideró que el dato faltante es independiente a la presencia de lluvia en ese día.

Al graficar las medidas ya definidas respecto de la distancia entre estaciones pluviométricas, se espera observar un decaimiento en las medidas cuando la distancia aumenta. Al graficar la correlación mensual respecto de la distancia se visualizó la relación esperada entre

las variables. No obstante en los gráficos a escala diaria existió una serie de puntos con una correlación bastante baja entre sí, independientemente de la distancia (ver figura 5.7).

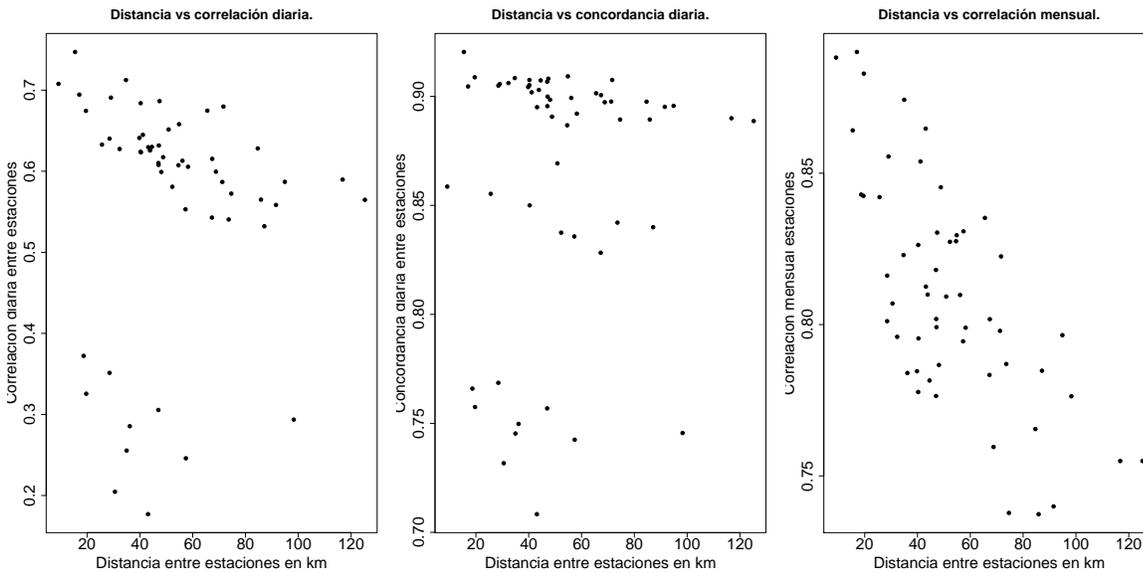


Figura 5.7: Gráfico de dispersión relacionando distancia con distintos valores. El primero compara distancia vs correlación diaria, el segundo compara distancia vs concordancia diaria y el tercero compara distancia vs correlación mensual.

Ante esta problemática se identificaron dichas estaciones y se encontró que la serie de puntos en cuestión corresponden a medidas de la estación de La Estanzuela. Al comparar los registros de las series se descubrió que hay un desfase de un día entre los registros en La Estanzuela y los demás registros (lo que en las estaciones de la DNM se registró en un día en La Estanzuela se registró al día siguiente).

Por lo tanto se modificó la estación de La Estanzuela para poder compararla con las demás estaciones (llevar la estación de La Estanzuela un día hacia atrás). En los próximos gráficos se comparan las distintas medidas de relación entre estaciones antes y después de modificar la estación de La Estanzuela:

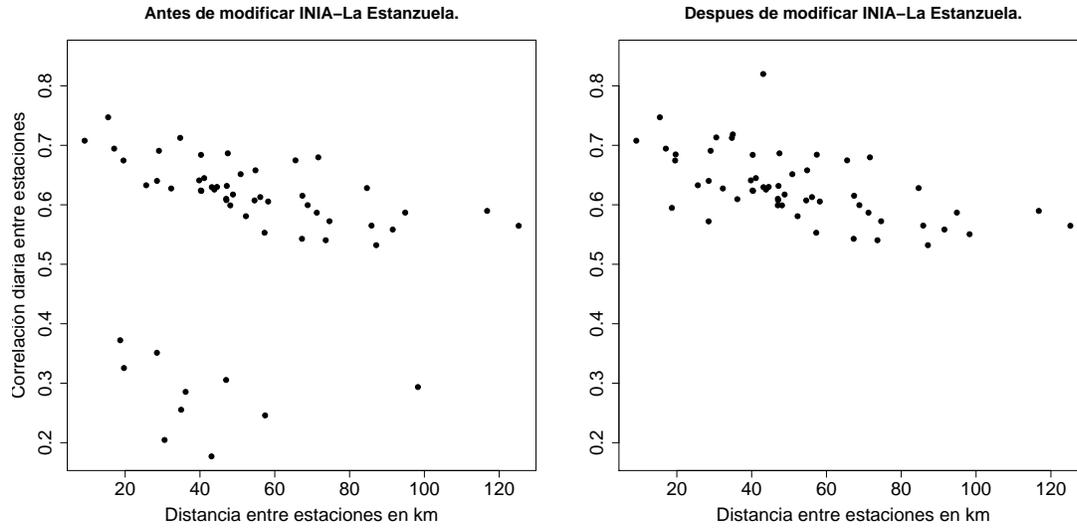


Figura 5.8: Comparación entre correlación diaria antes y después de modificar la serie de La Estanzuela.

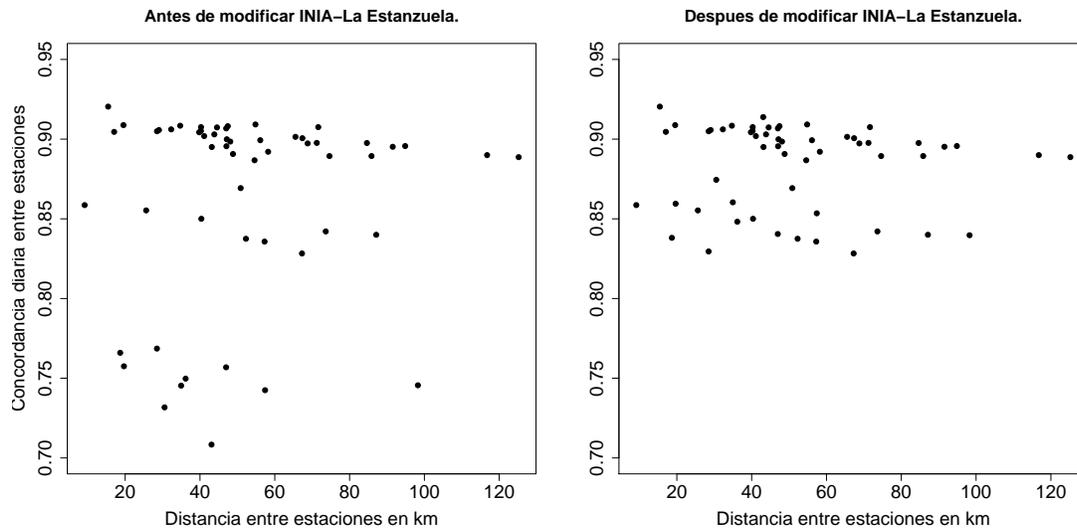


Figura 5.9: Comparación entre concordancia diaria antes y después de modificar la serie de La Estanzuela.

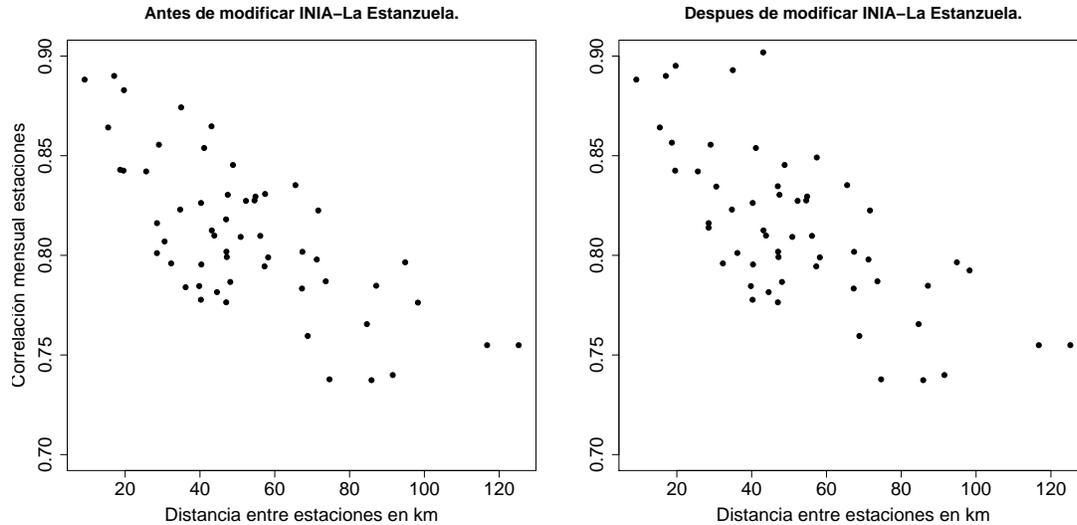


Figura 5.10: Comparación entre correlación mensual antes y después de modificar la serie de La Estanzuela.

En el Anexo A.2 se encuentran las matrices de correlación diaria, concordancia diaria y correlación mensual para las estaciones de Colonia luego de modificar la serie de La Estanzuela.

Al modificar la serie de La Estanzuela, la correlación entre distancia y correlación diaria pasó de  $-0.015$  a  $-0.579$ , con esta modificación se mostró como a medida que aumenta la distancia disminuye la correlación. La correlación entre distancia y concordancia diaria es muy baja en los dos casos, esto se debió a que en las estaciones de La Estanzuela y la meteorológica hay aproximadamente un 7 % menos de días lluviosos respecto a las estaciones pluviométricas, o cual impacta en la concordancia (ver nube de puntos inferiores en el gráfico 5.8).

La correlación entre distancia y correlación mensual pasó de  $-0.702$  a  $-0.690$  al modificar la serie de la estación de La Estanzuela, esta disminución no es significativa y la correlación es buena.

Al dejar solo las estaciones pluviométricas (sin La Estanzuela y sin la meteorológica), la correlación entre distancia y correlación diaria pasó de  $-0.579$  a  $-0.674$  (que es una diferencia significativa). La correlación entre distancia y correlación mensual no cambió en forma significativa, y la correlación entre distancia y concordancia diaria pasa de  $-0.035$  a  $-0.6919$  que es una diferencia muy significativa la cual resulta en una correlación bastante alta. Dada esta diferencia en la concordancia diaria se concluyó que las estaciones de La Estanzuela y la Meteorológica tienen un comportamiento distinto a las estaciones pluviométricas.

Se comparó la concordancia entre las estaciones con la concordancia que nos hubiera dado si las mismas fueran independientes. Teóricamente, si el evento llovió en la estación “a”, es independiente al evento llovió en la estación “b”, por la definición de independencia la probabilidad de que en las dos estaciones se de el mismo evento debería ser de:

$$P(\text{evento est.a} = \text{evento est.b}) = P(\text{lluvia est.a}) \cdot P(\text{lluvia est.b}) + P(\text{no lluvia est.a}) \cdot P(\text{no lluvia est.b}),$$

En el gráfico 5.11 se muestra que los puntos calculados bajo la hipótesis de independencia están bastante por debajo que lo observado en las series, por lo cual se rechazó cualquier hipótesis de independencia.

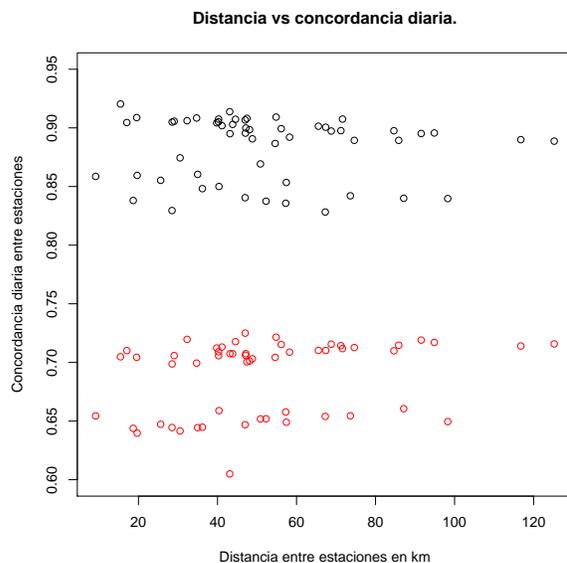


Figura 5.11: Concordancia diaria entre estaciones pluviométricas (negro), concordancia diaria simulando independencia (rojo).

## 5.5. Estudio de días lluviosos en el mes y lluvia mensual.

Luego del análisis de correlaciones se determinaron algunas variables de estudio que son importantes para estudiar la homogeneidad de la serie. Dichas variables fueron útiles para observar el comportamiento de las estaciones. En esta sección se detallaran aspectos relacionados a días lluviosos en el mes y lluvia mensual para cada estación.

Se construyó un boxplot de la cantidad de días lluviosos como un promedio mensual para todas las estaciones (ver figura 5.12). Se observó una mayor variabilidad en los meses de abril y agosto, por otro lado los meses menos variables fueron febrero y julio. Finalmente en octubre se registró la mayor cantidad de días lluviosos.

La mayor cantidad de días lluviosos en promedio se registró en los meses de febrero de 1977 y 1990, marzo de 2002 y 2007, y octubre del 2001, los cuales aparecen como los mayores outliers dentro de los gráficos de boxplot.

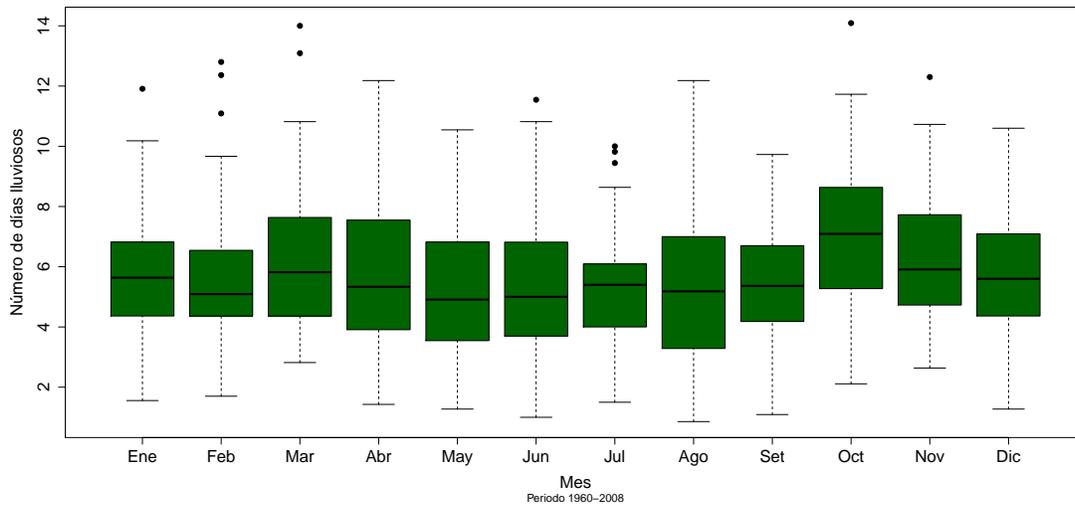


Figura 5.12: Boxplot de cantidad de días lluviosos para el total de estaciones en Colonia.

Por otro lado se realizó un boxplot de la cantidad promedio de lluvia por mes para el total de las estaciones (ver figura 5.13). Aquí se observó que los meses de junio y julio fueron los meses de menor variación, que, junto con los meses de mayo, agosto y setiembre fueron los meses que en mediana menos llovió. Por otra parte octubre es el mes que más llovió junto con marzo que es bastante más variable.

En marzo hubo dos años en los que llovió mucho más que lo normal. Dicho aspecto no se debió a algún dato mal digitado, ya que se trata de un promedio mensual de lo que llovió en todas las estaciones.

El outlier más significativo se observó en marzo del año 2007, que por lo que ya se vió fue el mes en el cual se registraron más días lluviosos en promedio. Otro dato significativo en marzo se observó en el año 1988. El mismo no se corresponde con un outlier en el mes de marzo, sin embargo en promedio llovió 10,8 días, que si bien no llega a ser un outlier, al observar el boxplot de marzo, esta es una cantidad bastante alta.

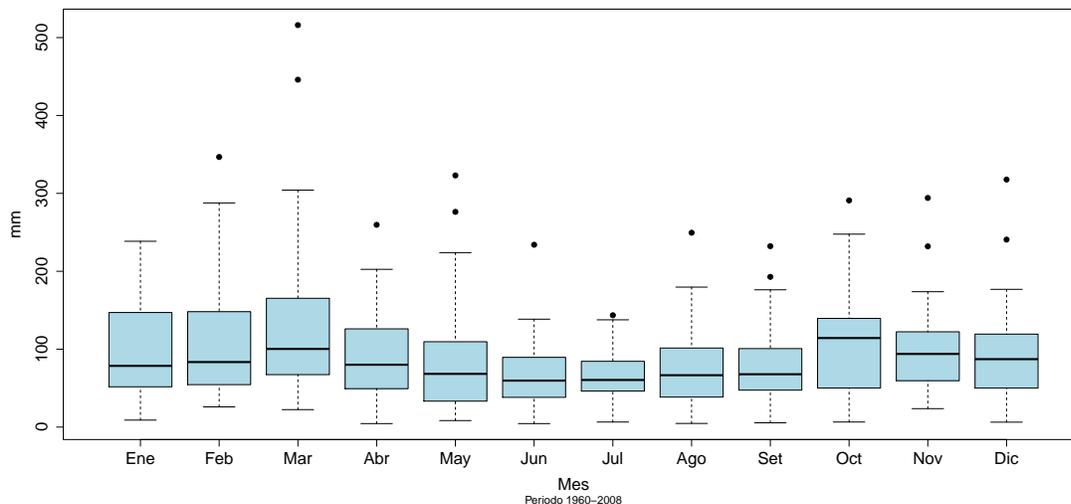


Figura 5.13: Boxplot de acumulado mensual para el total de estaciones en Colonia.

A los efectos de realizar una análisis estacional de las precipitaciones acumuladas se definieron como meses de verano a diciembre, enero y febrero, como meses de otoño a marzo, abril y mayo, como meses de invierno a junio, julio y agosto y como meses de primavera a setiembre octubre y noviembre.

Con las estaciones así definidas se realizó un gráfico Q-Q entre ellas (ver gráfico 5.14), donde se concluyó que en verano es la época que llovió mas, luego en primavera llovió un poco menos, después en orden descendente viene otoño y donde menos llovió es en invierno.

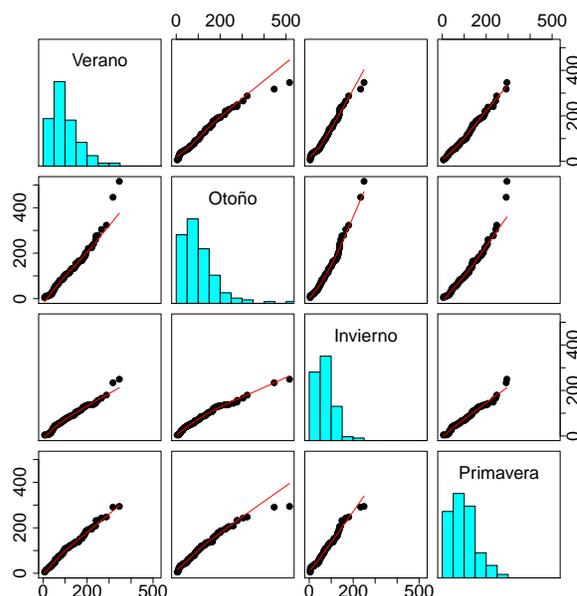


Figura 5.14: Gráfico Q-Q entre lluvia acumulada en mm, en las estaciones del año en Colonia en el periodo 1960-2008.

Observación: Hubo dos puntos que se consideraron como atípicos ya que esos años hubo

demasiadas precipitaciones, especialmente en verano, sin embargo estos datos no fueron considerados como influyentes, ya que la recta “regresora” no se modificó en forma significativa debido a estos puntos (no afectó la relación de precipitación acumulada entre las distintas épocas del año), por lo que la inclusión de estos puntos no trajo ningún tipo de problema.

En las tablas 5.4 y 5.5 se encuentran medidas de interés tanto para la cantidad de lluvia mensual como para la cantidad de días lluviosos en el mes, estas son el promedio mensual, el desvío standard y los respectivos intervalos de confianza de cada promedio. Vemos que en marzo llovió en promedio bastante más que en el resto de los meses, sin embargo, este valor está influenciado por los dos outliers evidenciados en el boxplot del gráfico 5.13.

En octubre llovió en promedio 17.8 mm menos que en marzo, sin embargo en octubre se registró la mayor cantidad de días promedio con lluvia.

	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Set.	Oct.	Nov.	Dic.
A.M.	102.3	108.3	130.0	90.5	81.1	68.3	67.3	73.6	79.4	112.2	96.0	91.4
D.S. lluvia mens	61.6	73.1	98.0	58.9	67.6	43.4	32.2	52.1	47.4	66.2	53.7	58.2
I.C. inferior	25.6	36.8	36.7	15.8	12.8	12.5	23.1	8.1	25.7	29.9	26.4	24.8
I.C. superior	214.7	259.0	293.8	192.7	210.5	132.1	135.3	162.4	171.0	229.1	170.9	175.3

Cuadro 5.4: Lluvia mensual, donde D.S.:Desvío Standard, I.C.:Intervalo de Confianza, A.M:Acumulado Medio.

	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Set.	Oct.	Nov.	Dic.
Media días lluv. mens.	5.7	5.8	6.4	5.9	5.2	5.3	5.3	5.3	5.5	7.0	6.3	5.7
D.S. días lluv. mens.	2.2	2.3	2.7	2.7	2.3	2.4	2.0	2.6	1.9	2.6	2.2	2.0
I.C. inferior	2.4	3.2	2.9	2.1	2.1	2.0	2.1	1.7	2.9	3.0	3.3	2.4
I.C. superior	9.6	10.5	10.8	11.0	8.9	9.5	9.1	9.0	8.8	10.6	9.5	8.9

Cuadro 5.5: Días lluviosos en el mes, donde DS: Desvío Standard, IC:Intervalo de Confianza.

Las variables días lluviosos y cantidad de lluvia fueron estudiadas por mes y estación, con el objetivo de comparar las dos medidas en las distintas estaciones.

En el gráfico 5.15 se visualizó que en la estación de La Estanzuela y en la meteorológica se registró en promedio entre dos y tres días lluviosos más que por mes que las estaciones pluviométricas. Sin embargo la forma de la distribución de los días lluviosos por mes es similar.

En términos generales se observó un crecimiento de enero a marzo, una disminución hasta mayo, un nivel constante hasta setiembre, un salto abrupto en octubre y una disminución hasta diciembre llegando a un nivel similar al de enero.

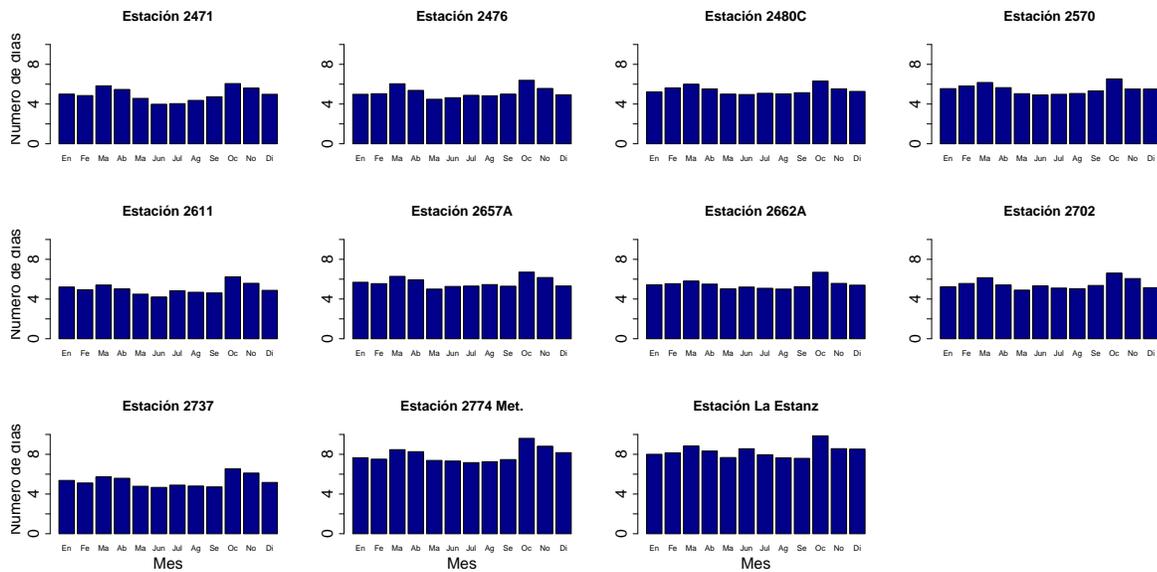


Figura 5.15: Gráfico de barras de días lluviosos por mes según estación en Colonia.

En el gráfico 5.16 se ha representado la cantidad de lluvia mensual por estación. Se observó a partir de este, que la cantidad de lluvia mensual en promedio es prácticamente la misma (a excepción de los eneros de la estación 2471 y 2570, diciembre en la 2480C y octubre en la 2737). En estas estaciones se dió un crecimiento en el total de lluvias desde enero a marzo, luego decreció el mismo hasta julio, luego creció nuevamente hasta octubre, y finalmente decreció un poco en noviembre y diciembre, donde la lluvia mensual en diciembre es un poco inferior a la de enero.

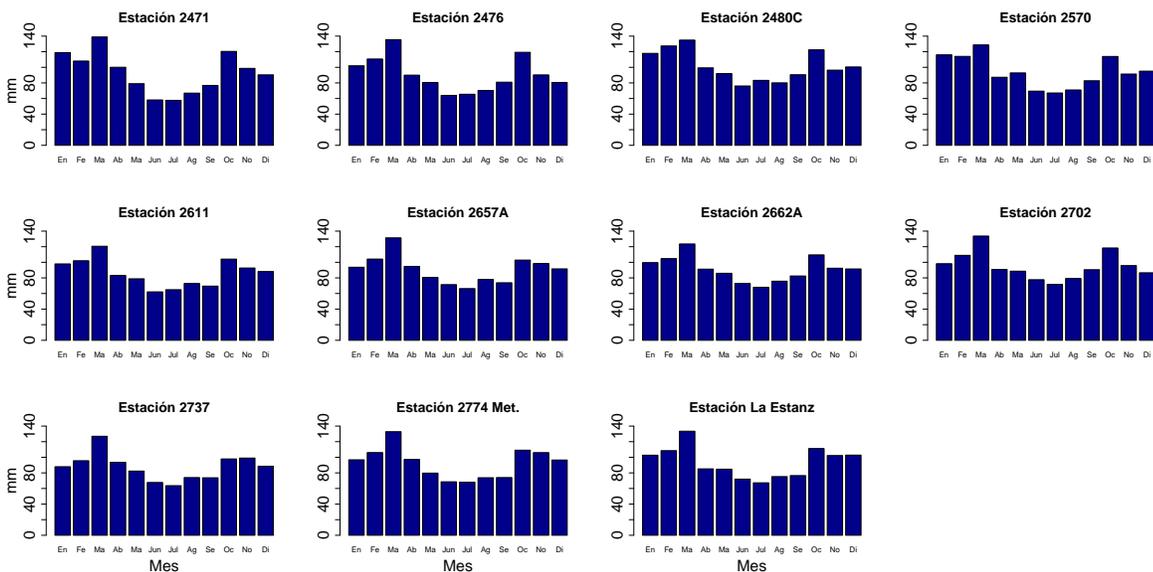


Figura 5.16: Gráfico de barras de lluvia mensual por mes según estación en Colonia.

Para estudiar si hay más días lluviosos por mes en las estaciones de La Estanzuela y la meteorológica respecto de las estaciones pluviométricas se construyó un gráfico Q-Q

entre días lluviosos en cada estación pluviométrica con respecto a la meteorológica y La Estanzuela (ver gráficos 5.17 y 5.18).

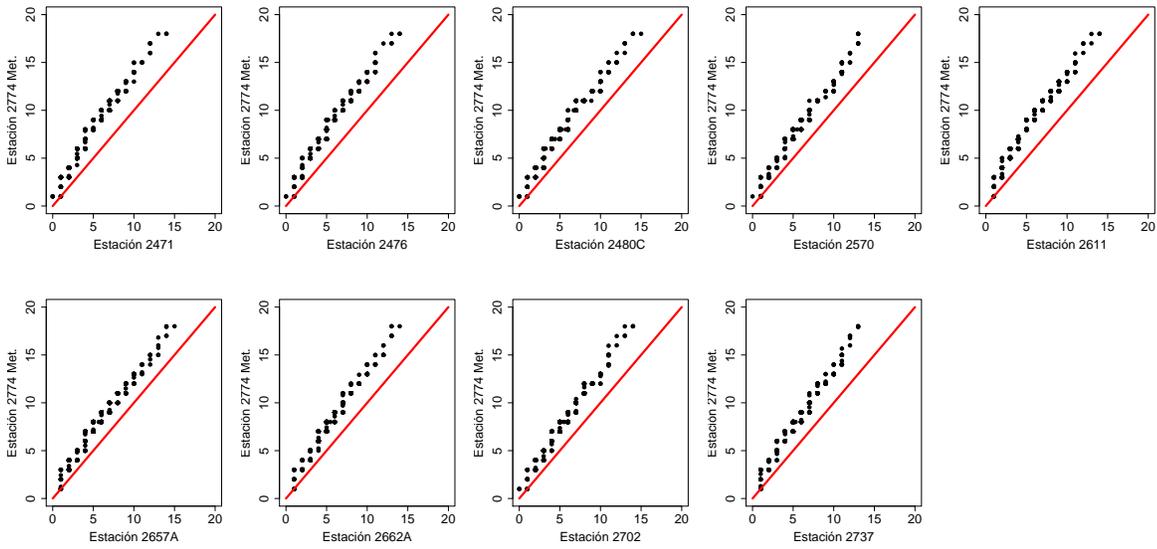


Figura 5.17: gráfico Q-Q entre días lluviosos en cada estación pluviométrica con respecto a la meteorológica en Colonia.

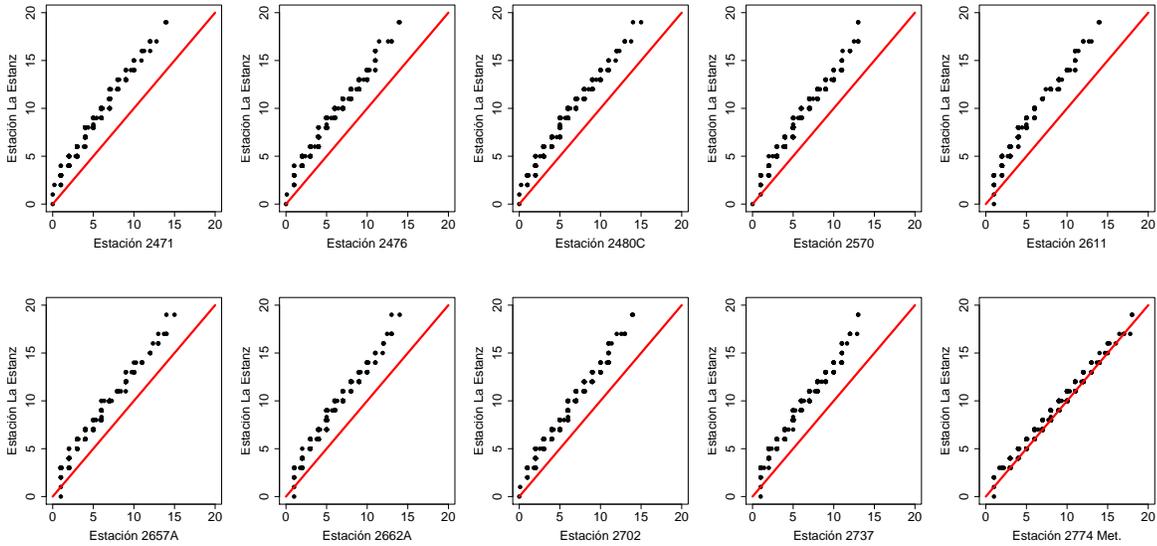


Figura 5.18: gráfico Q-Q entre días lluviosos en cada estación perteneciente a la DNM(incluyendo a la meteorológica) con respecto a La Estanzuela en Colonia.

Tanto en el gráfico 5.17 como en el 5.18 se verificó que si bien no hay una diferencia entre cantidad de precipitación mensual, efectivamente se registraron más días lluviosos en La Estanzuela y en la meteorológica respecto a las estaciones pluviométricas.

Luego se estudió la correlación de Pearson entre lluvia mensual y días lluviosos por estación, obteniéndose un valor de 0.70 aproximadamente para las estaciones pluviométricas.

cas, mientras que para La Estanzuela y la meteorológica se obtuvo un valor del orden de 0.60. Esto indica que efectivamente hay correlación entre las dos variables en cuestión. Es importante observar que esto no necesariamente tiene que ocurrir siempre.

También se estudió correlación entre lluvia mensual y días lluviosos por estación discriminando según el mes, obteniéndose que en los meses de julio y noviembre hay una correlación menor, destacándose los casos de la estación La Estanzuela, 0.33 en julio y 0.38 en noviembre. De todas formas se concluyó que las dos variables de interés están débilmente correlacionadas ya que el intervalo de confianza no fue lo suficientemente amplio como para contener al cero.

## 5.6. Densidad estimada de precipitación diaria.

Los datos de precipitación diaria tienen la peculiaridad de tener una altísima cantidad de empates, estos ocurren cuando no hay precipitación. La solución planteada a este problema fue calcular la densidad condicionada a que llovió, es decir se calculó  $\hat{P}(x/x > 0)$  donde  $x$  representa la variable aleatoria de precipitación diaria.

Dado que ninguna distribución conocida se ajustó a la precipitación diaria, se procedió a una estimación no paramétrica de dicha distribución, para ello se consideraron las dos propiedades básicas de cualquier densidad, la primera es que integra uno en el recorrido de la variable aleatoria y además que la función de densidad es siempre mayor o igual a cero. Luego se definió el siguiente estimador:

$$\hat{f}(y) = \frac{1}{nh} \cdot \sum_{i=1}^n K\left(\frac{y - y_i}{h}\right),$$

donde  $K(x)$  representa una función que integra a uno en su recorrido. Para simplificar las cosas se eligió como núcleo  $K$  a la distribución normal estandarizada. Por otro lado  $h$  es la ventana que se eligió para estimar la densidad ( $h$  puede ser una función de la cantidad de datos  $n$ , y se debe elegir de modo que tienda a cero pero que la convergencia a cero sea de orden inferior que la convergencia de  $n$  a infinito, ya que de lo contrario quedan algunas ventanas sin suficientes observaciones para realizar la estimación). Este método se denomina estimación por kernels utilizando el kernel normal.

A partir del gráfico 5.19 se detectó que la densidad estimada en las estaciones pluviométricas son muy distintas a las de La Estanzuela y la meteorológica, siendo estas últimas dos muy similares entre sí. Aunque todas presentaron una asimetría con cola hacia la derecha (lo cual es lógico ya que desde el punto de vista de la intensidad, la probabilidad de lluvias de alta intensidad es mucho menor a la probabilidad de que las lluvias sean de baja intensidad). Las estaciones de La Estanzuela y la Meteorológica presentaron un grado de apuntamiento (kurtosis) mucho mayor que las estaciones pluviométricas.

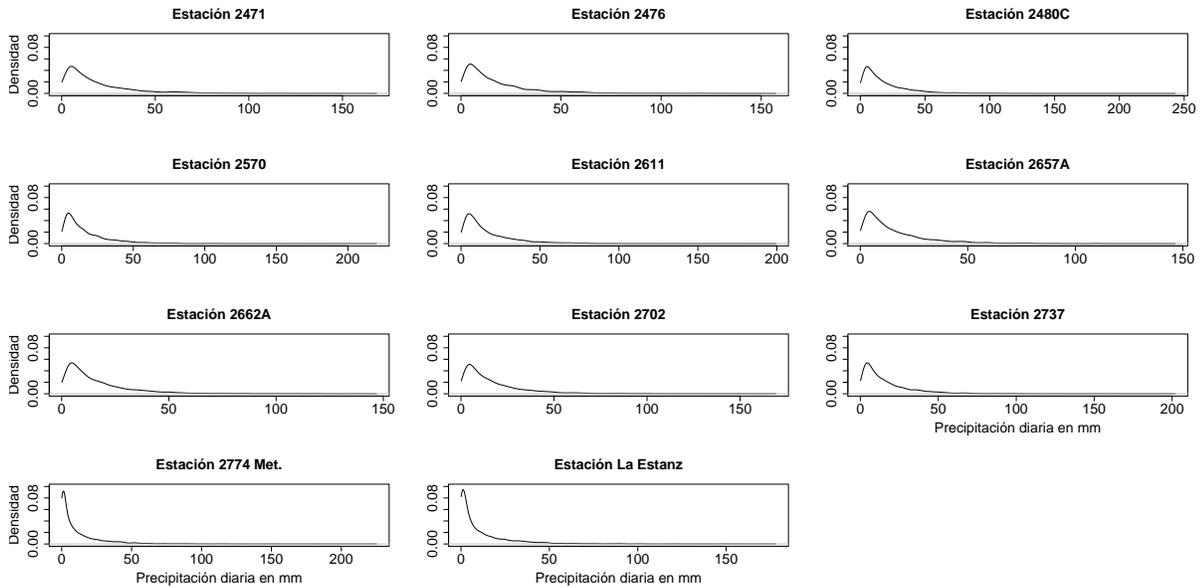


Figura 5.19: Densidad estimada de cantidad de lluvia dado que llovió en cada estación de Colonia.

Una segunda densidad calculada fue la densidad de precipitación dado que llovió más de 5 mm. (ver gráfico 5.20), esta se estimó de manera análoga a la anterior. Al realizar esta nueva restricción las densidades son bastante parecidas, al menos en su apuntamiento y comportamiento asimétrico.

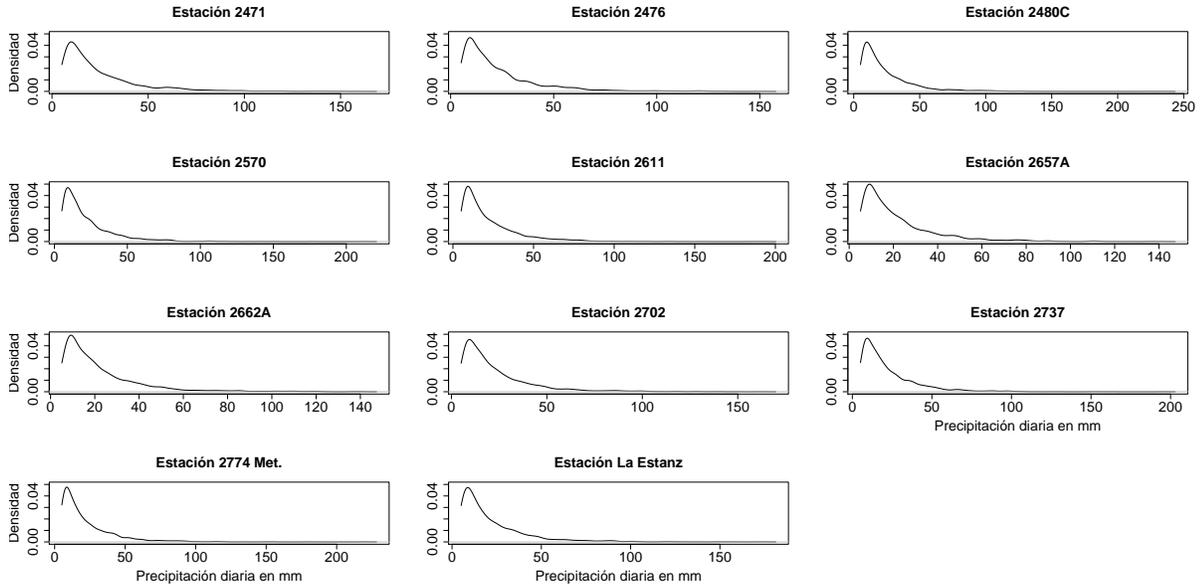


Figura 5.20: Densidad estimada de cantidad de lluvia dado que llovió más de 5 mm. en cada estación de Colonia.

En el Gráfico 5.21 se hallan todas las densidades de precipitaciones en un solo gráfico, condicionado a precipitaciones mayores a 5 mm y a 10mm. Se observó que todas las estaciones experimentan la misma caída, lo que si se cambia para cada estación, y específicamente al condicionarlas a precipitaciones mayores a 5 mm, es el grado de apuntamiento.

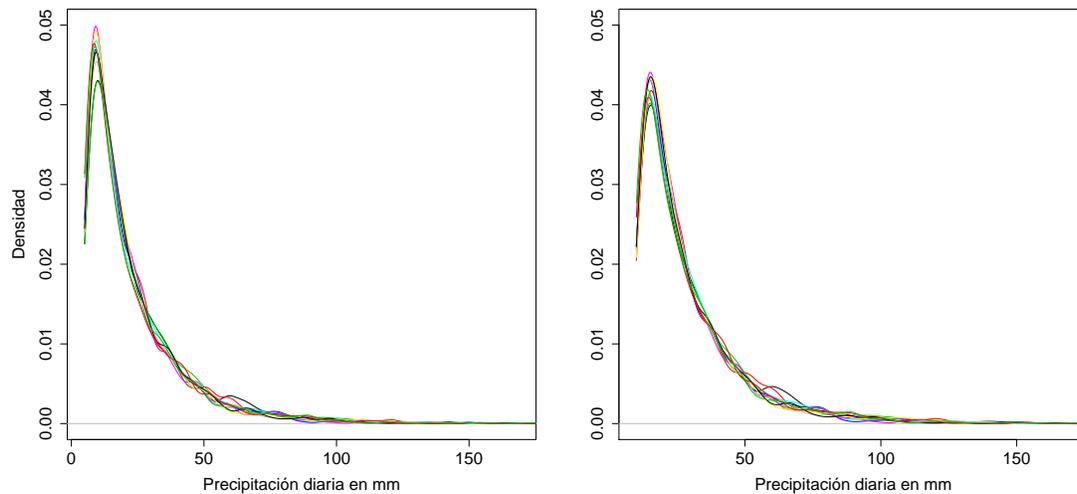


Figura 5.21: Densidad estimada dado que llovió más de 5mm en el gráfico a la izquierda y más de 10 mm en el gráfico a la derecha, para cada estación por separado.

## 5.7. Análisis de precipitación.

Antes de analizar las formas en las que se pueden detectar los datos faltantes, fue necesario observar ciertos aspectos, uno de ellos está relacionado a las diferentes cantidades de lluvia observadas en las precipitaciones.

En este análisis fue necesario tener en cuenta que en la estación de La Estanzuela se comenzó con el registro de datos a partir del año 1965, por lo que en dicha estación no se registró la mayor cantidad de lluvia, debido a que la misma tiene cinco años menos de registro que el resto de las estaciones (cuyo período de registro es 1960-2008).

En la tabla 5.6 se muestra la cantidad de días lluviosos y la diferencia entre cada estación y la estación en la que llovió más días.

Se observó una amplia diferencia entre la cantidad de lluvia registrada en las estaciones pluviométricas y la registrada en la estación meteorológica y en La Estanzuela. La cantidad de días lluviosos fue similar entre las propias estaciones pluviométricas, así como también entre las estaciones meteorológica y La Estanzuela.

En promedio se registraron 1546 días lluviosos más en la estación meteorológica con respecto a las estaciones pluviométricas (cuatro años y tres meses aproximadamente). Lo que marcó una aproximación a la cantidad de datos (faltantes) que se deberían encontrar en el proceso de detección de datos faltantes.

Estación	Días lluviosos	Diferencia
Estación 2471	2866	1733
Estación 2476	2995	1604
Estación 2480C	3131	1468
Estación 2570	3166	1433
Estación 2611	2878	1721
Estación 2657A	3173	1426
Estación 2662A	3161	1438
Estación 2702	3172	1427
Estación 2737	2931	1668
Estación 2774 Met.	4599	0
Estación La Estanz.	4336	263

Cuadro 5.6: Cantidad de días lluviosos y su diferencia con la estación de mayor precipitación. observación: dif = diferencia entre días lluviosos en la estación y en la estación donde llueve mas.

Otro aspecto de interés refiere a la detección de períodos de tiempo caracterizados por un mejor registro de las precipitaciones. A modo de ejemplo, es posible que el reemplazo de un observador por otro más responsable, o un cambio en la forma de medir los datos, produzcan cambios en la calidad de los registros.

En la tabla 5.7 se muestra la diferencia entre días lluviosos en cada estación respecto a la estación donde llueve más pero por década.

La máxima cantidad de días lluviosos por década se registró en las estaciones de La Estanzuela o en la meteorológica. Antes de construir la tabla 5.7 se pensó que a medida que pasa el tiempo los registros son mejores y por lo tanto que la diferencia entre días lluviosos en las estaciones es cada vez menor, sin embargo no se observó ningún patrón a la baja entre la diferencia de días lluviosos entre ellas.

Además al sumar las diferencias por estación, para las estaciones pluviométricas se obtuvo una diferencia total promedio de 1674.4 días, este promedio es 128 días mayor que el promedio hallado sin diferenciar por décadas, lo cual repercutió en aumentar la aproximación a la cantidad esperada de datos faltantes a detectar.

Estación/Años	[1960,1970)	[1970,1980)	[1980,1990)	[1990,2000)	[2000,2008]	Total
Estación 2471	301	480	384	387	309	1861
Estación 2476	263	395	343	458	273	1732
Estación 2480C	229	383	330	385	269	1596
Estación 2570	212	346	350	356	297	1561
Estación 2611	317	398	381	435	318	1849
Estación 2657A	280	382	296	326	270	1554
Estación 2662A	218	373	321	374	280	1566
Estación 2702	180	412	303	365	295	1555
Estación 2737	312	465	350	399	270	1796
Estación 2774 Met.	0	102	6	0	20	128
Estación La Estanz.	339	0	0	52	0	391

Cuadro 5.7: Diferencia entre días lluviosos en la estación y en la estación donde llueve más por década.

Finalmente se clasificó a la cantidad de días lluviosos pero según intensidad de lluvia registrada (ver tabla 5.8). Esto se hizo de forma similar a la clasificación realizada por los autores que trabajaron en la región de Calabria, Italia [3] en el estudio esbozado en la sección 1.3. En dicho trabajo se clasificó a la intensidad de lluvia en:

- Precipitación suave de 0.1 a 5 milímetros registrados en el día.
- Precipitación moderada de 5.1 a 15 milímetros registrados en el día.
- Precipitación moderada a fuerte de 15.1 a 50 milímetros registrados en el día.
- Precipitación fuerte de 50.1 a 80 milímetros registrados en el día.
- Precipitación severa si esta es mayor a 80 milímetros registrados en el día.

	Suave	Moderada	Moderada a fuerte	Fuerte	Severa	Faltante
Estación 2471	752	964	947	149	54	305
Estación 2476	829	1031	945	143	47	306
Estación 2480C	785	1071	1073	139	63	213
Estación 2570	887	1117	991	136	35	395
Estación 2611	786	1048	883	129	32	399
Estación 2657A	958	1109	968	108	30	857
Estación 2662A	906	1108	997	103	47	273
Estación 2702	908	1072	1023	112	57	326
Estación 2737	879	1004	906	104	38	1030
Estación 2774 Met.	2344	1124	940	135	56	231
Estación La Estanz.	2280	1009	896	103	48	0

Cuadro 5.8: Cantidad de días lluviosos por clasificación de intensidad de lluvia diaria según estación.

La gran diferencia respecto a los registros se observó en lluvias suaves, lo que denotó una carencia por parte de las estaciones pluviométricas a la hora de registrar este tipo de lluvias. Debido a estas diferencias se realizaron los gráficos 5.22 y 5.23.

Tanto en el Gráfico 5.22 como en el 5.23 representan estimaciones para las estaciones de Colonia de densidades condicionadas a que la precipitación fue clasificada como suave en un caso y como severa en el otro, al realizar esto estudiamos ambos extremos de la densidad de lluvia.

En el gráfico 5.22 se observó que las densidades estimadas para las estaciones pluviométricas son muy distintas a las densidades estimadas para las estaciones de La Estanzuela y la meteorológica.

En las estaciones pluviométricas detectamos que la densidad tiene un comportamiento sinusoidal con picos en 1mm, 2mm, 3mm, 4mm y 5mm. Desde el pico 1mm hasta el pico 3mm el valor de la densidad crece (el pico en 1 es menor que el pico en 2mm que, a su vez, es menor que el pico en 3mm), sin embargo este comportamiento es debido a que en las estaciones pluviométricas no hay precisión a la hora de medir la precipitación.

Otro aspecto observado es que el pico en 4mm es inferior a los demás. Esto se debe a que se cuando el valor está entre 4mm y 5mm el observador eligió 5 mm ya que este es un

número más atractivo. Los patrones recién descritos no se observaron en las estaciones de La Estanzuela y la meteorológica, las cuales manifestaron un crecimiento en su densidad hasta 0.5 mm para luego descender.

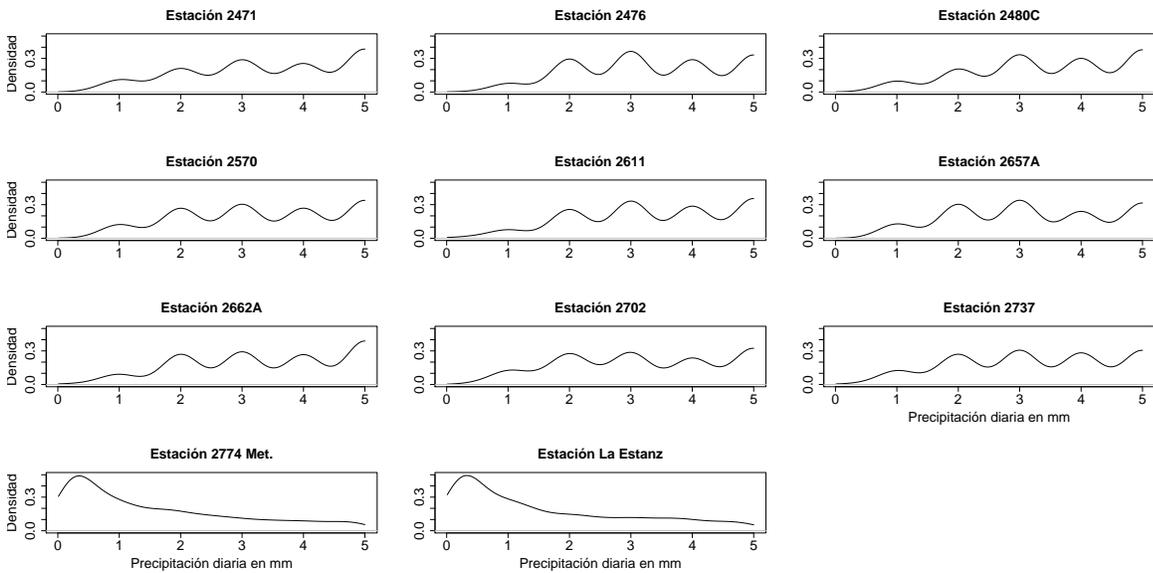


Figura 5.22: Densidad estimada de cantidad de lluvia dado que la precipitación diaria fue suave.

En el gráfico 5.23 a diferencia del gráfico 5.22 las densidades son bastante similares para todas las estaciones observadas, aunque se siguen observando algunas diferencias, especialmente entre las estaciones de La Estanzuela y la meteorológica respecto a las estaciones pluviométricas.

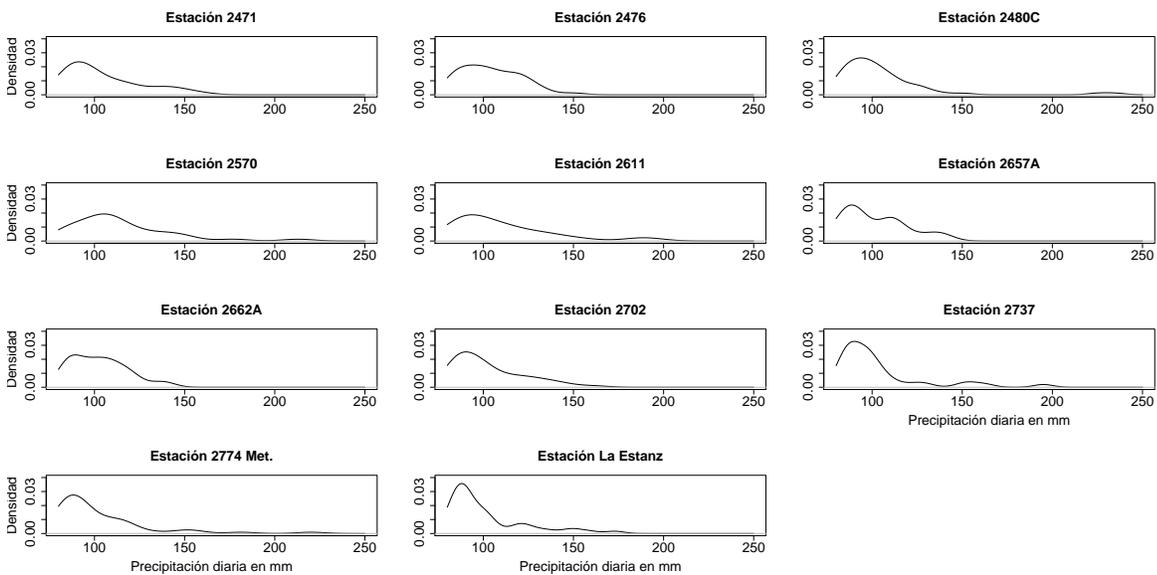


Figura 5.23: Densidad estimada de cantidad de lluvia dado que la precipitación diaria fue severa.

## 5.8. Conclusiones del capítulo.

Por un lado no se hallaron diferencias entre las estaciones de Colonia respecto a la intensidad de lluvia registrada, por otro sí hubo diferencias respecto a la cantidad de días lluviosos en el mes. Se registraron más días lluviosos en las estaciones de La Estanzuela y la meteorológica que en las estaciones pluviométricas (en promedio son 1674 días lluviosos de diferencia).

Se detectó que la diferencia en la cantidad de registros de precipitación se encuentra en lluvias de intensidad suave (entre 0.1mm y 5.0mm). Además, la medición en las estaciones pluviométricas es mucho menos precisa que en las estaciones de La Estanzuela y la meteorológica, notándose que para las estaciones pluviométricas la mayoría de los registros de intensidad de precipitación se encuentran en números enteros, es decir que no se registró la cifra decimal.

Debido a estos aspectos se concluyó que hay una gran diferencia entre la calidad de las estaciones pluviométricas con respecto a La Estanzuela y la meteorológica. Tanta es la misma que dejó prácticamente inutilizables a las estaciones pluviométricas, ya se estimó que por lo menos esas series tienen un 10 % más de datos faltantes que los que se preveía al inicio del capítulo.

Finalmente este capítulo por un lado nos permitió asegurar cierta confianza en los registros de las estaciones de La Estanzuela y la meteorológica. Por otro se desechó la intención de cumplir nuestro objetivo principal que era el de obtener una base de datos confiables de precipitación para el total del país, debido a que hay un alto porcentaje de datos faltantes en las estaciones pluviométricas, las cuales representan el 89 % del total de estaciones que en el final del capítulo cuatro se consideraron como correctas.

## Capítulo 6

# Modelo para la estimación de ocurrencia de lluvia.

Debido a la alta cantidad de ceros registrados principalmente en estaciones pluviométricas, que es altamente probable que una importante cantidad sean datos faltantes, especialmente en estaciones pluviométricas, nos vemos ante la imposibilidad de detectar un dato faltante a partir de estaciones cercanas o de la propia serie. Por otro lado no es posible reconstruir la información de precipitación a escala diaria para las series.

Ante este inconveniente, se optó por realizar un modelo para la estimación de ocurrencia de lluvia en estaciones de Colonia, dado que a partir de este modelo podremos describir características de la precipitación las cuales tienen uso directo en diversos campos de estudio.

### 6.1. Selección de la metodología.

#### 6.1.1. Primer intento para el modelado de la ocurrencia de lluvia.

Sea  $X$  la variable aleatoria que mide la ocurrencia de lluvia, la misma vale 1 cuando se registra lluvia y 0 en caso contrario. Un modelo muy simple para modelar esta variable es el esquema de Bernoulli, basado en una sucesión de variables aleatorias independientes e idénticamente distribuidas,  $X_i \sim Ber(p)$  con  $i = 1, \dots, n$ , donde  $p$  representa la probabilidad de lluvia en un día cualquiera, y  $n$  el largo de la serie. La verosimilitud del modelo es:

$$L(p) \propto p^{\text{Número de días lluviosos}} \cdot (1-p)^{\text{Número de días secos}},$$

donde la estimación de  $p$  que maximiza la verosimilitud del modelo es  $\hat{p} = \sum_i \frac{X_i}{n}$

Primero se deberá probar si el supuesto de independencia es viable, para ello se utilizó la estación Meteorológica, ya que la misma fue clasificada como confiable y por lo tanto útil para ilustrar la metodología, luego se repitió el procedimiento para el resto de las estaciones.

Si el modelo de independencia fuese correcto, se esperaría observar un total de  $m \cdot (1-\hat{p}) \cdot \hat{p}$  días secos seguidos de días lluviosos, donde  $m$  es la cantidad de días donde es posible observar dicho cambio. En el caso de presencia de datos faltantes, por ejemplo:

$$(1, 0, 1, 1, 1, -99,9, -99,9, 1, 0, 0, 0, 0, 1)$$

no se considera el cambio de 1 a  $-99,9$  ni de  $-99,9$  a 1, por lo tanto  $m$  sería igual a nueve.

Para la estación meteorológica los datos del modelo son:  $m = 17660$ , y  $\hat{p} = 0,260$ , por lo tanto  $m \cdot \hat{p} \cdot (1 - \hat{p}) = 3400,46$ . La tabla 6.1 contiene el conteo total para la estación

meteorológica, con el conteo esperado bajo el supuesto de independencia mostrado entre paréntesis.

	Hoy Seco	Hoy Llueve	Total
Ayer Seco	10459 (9662.36)	2603 (3400.46)	13062
Ayer Llovió	2603 (3400.46)	1995 (1196.72)	4598
Total	13062	4598	17660

Cuadro 6.1: Relación en la ocurrencia de lluvia entre ayer y hoy.

Para poner a prueba la independencia se utilizó la prueba ji cuadrado planteado en la sección 3.5 a través de la siguiente expresión:

$$\chi^2 = \sum_{j=1}^2 \sum_{i=1}^2 \frac{(x(i, j)_{\text{Empírico}} - x(i, j)_{\text{Teórico}})^2}{x(i, j)_{\text{Teórico}}},$$

donde  $x(i, j)$  representa el elemento  $i, j$  de la tabla 6.1, para el caso teórico el mismo es bajo el supuesto de independencia y el empírico es el observado en la serie.

Este estadístico bajo  $H_0$  tiene una distribución asintótica  $\chi^2$  con  $(Nro. de filas - 1) \cdot (Nro. de columnas - 1) = 1$  grados de libertad, por lo tanto para rechazar la hipótesis nula con un nivel de significación  $\alpha = 0,05$ , el valor de estadístico se debe encontrar en la región crítica determinada por el intervalo  $(3,84, +\infty)$ . El valor del estadístico para esta estación es 972.21 por ende se rechazó la hipótesis de independencia para la estación meteorológica, debido a que el valor del estadístico pertenece a la región crítica.

Para las demás estaciones de Colonia también se rechazó la hipótesis de independencia. Ya que no se cumple la independencia, se descarta el modelo inicial (Bernoulli), y se analizan las probabilidades condicionales de ocurrencia de lluvia:

$$P(X_{i+1} = l | X_0 = k_0, \dots, X_i = k_i) \text{ donde } l, k_0, \dots, k_i \in \{0, 1\}$$

Una manera simple de especificar las probabilidades anteriores es a través de las condiciones que caracterizan una cadena de Markov, con espacio de estados  $\mathcal{E} = \{t, \infty\}$ .

### 6.1.2. Marco teórico.

Decimos que  $(X_n)_{n \geq 0}$  es una cadena de Markov a valores en el espacio de estados  $\mathcal{E}$ , de primer orden con distribución inicial  $\lambda$ , y matriz de transición P, si:

1.  $X_0$  tiene distribución  $\lambda$ .
2. Las probabilidades condicionales  $P(X_{n+1} = i_{n+1} | X_0 = i_0 \dots X_n = i_n) = P(X_{n+1} = i_{n+1} | X_n = i_n)$  para todo  $i_0, i_1, \dots, i_n, i_{n+1} \in E$ . Esta condición indica que la distribución de  $X_{n+1}$  (asociado al futuro) dado  $X_n$  (asociado al presente) es independiente de  $X_0, X_1 \dots X_{n-1}$  (asociado al pasado).

Para definir una cadena de Markov de orden mayor a uno, es necesario modificar la segunda condición. Por ejemplo si se quiere definir una cadena de Markov de segundo orden la segunda condición sería:  $P(X_{n+1} = i_{n+1} | X_0 = i_0 \dots X_n = i_n) = P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1})$  para todo  $i_0, i_1, \dots, i_n, i_{n+1} \in E$ .

A partir de los datos se debe estimar la matriz de transición P. Bajo el supuesto de que la cadena de Markov es homogénea, es decir los elementos de la matriz P permanecen constantes a través del tiempo:

$$p_{i,j} = P(X_n = i/X_{n-1} = j) = P(X_n = i, X_{n-1} = j)/P(X_{n-1} = j)$$

y dada una serie de largo N:

$$P(X_n = i/X_{n-1} = j) \approx \frac{N(X_n = i, X_{n-1} = j)}{N(X_{n-1} = j)}$$

donde  $N(X_n = i, X_{n-1} = j)$  es el número de veces que la cadena pasa del estado j al estado i en el período de tiempo de largo N, y  $N(X_{n-1} = j)$  es el número de veces que la cadena ocupa el estado j,  $X_{n-1} = j$  con  $n = 1, \dots, N$ . Esta fórmula (estimación máximo verosímil de los elementos de P) fue aplicada a nuestros datos.

### Propiedad fuerte de Markov.

Antes de enunciar la propiedad fuerte de markov, decimos que la variable aleatoria  $T : \Omega \rightarrow \{0, 1, \dots\} \cup \{\infty\}$  es llamada “tiempo de parada” si el evento  $\{T=n\}$  depende solo de  $X_0, X_1 \dots X_n$  para  $n = 0, 1, \dots$ . Por ejemplo  $T_j = \inf\{n \geq 1 : X_n = j\}$  es un tiempo de parada porque  $\{T_j = n\} = \{X_1 \neq j, X_2 \neq j, \dots, X_{n-1} \neq j, X_n = j\}$ .

La propiedad fuerte de Markov se enuncia de la siguiente manera: sea  $(X_n)_{n \geq 0}$  una cadena de Markov homogénea de primer orden  $(\lambda, P)$  y sea T un tiempo de parada asociado a  $(X_n)_{n \geq 0}$ . Entonces dado que  $T < \infty$  y  $X_T = i$ ,  $(X_{T+n})_{n \geq 0}$  es Markov de primer orden  $(\delta_i, P)$  y es independiente de  $X_0, X_1 \dots X_T$ .

Una aplicación de la propiedad fuerte de Markov es la cadena de Markov homogénea  $(X_n)_{n \geq 0}$  observada solo en ciertos tiempos. En una primera instancia suponemos que J es un subconjunto del espacio de estados E y que se observó la cadena solo cuando la misma toma valores en J. El proceso resultante  $(Y_m)_{m \geq 0}$  puede ser obtenido formalmente tomando  $Y_m = X_{T_m}$ , donde:

$$T_0 = \inf\{n \geq 0 : X_n \in J\}$$

y para  $m = 0, 1, 2, \dots$

$$T_{m+1} = \inf\{n \geq T_m : X_n \in J\}.$$

Luego se asume que  $P(T_m < \infty) = 1 \forall m$ . Para cada m se puede verificar fácilmente que  $T_m$ , el tiempo de la m-esima visita a J, es un tiempo de parada. Entonces aplicando la propiedad fuerte de Markov se muestra para  $i_0, \dots, i_{m+1} \in J$  que:

$$\begin{aligned} & P(Y_{m+1} = i_{m+1} | Y_0 = i_0, \dots, Y_m = i_m) \\ &= P(X_{T_{m+1}} = i_{m+1} | X_{T_0} = i_0, \dots, X_{T_m} = i_m) \\ &= P_{i_m}(X_{T_1} = i_{m+1}) = \bar{p}_{i_m, i_{m+1}}. \end{aligned}$$

En las series de ocurrencia de lluvia aquí estudiadas se tienen tres estados: 1 si llueve, 0 si no llueve y -99.9 cuando es dato faltante. En esta aplicación queda visto que se puede trabajar solamente con los estados 0 y 1 (que serían los estados observados) que nos interesa utilizar y modelar.

### 6.1.3. Segundo intento para el modelado de la ocurrencia de lluvia.

En la sección 6.1.1 se asumió que el proceso de ocurrencia de lluvia puede modelarse mediante cadenas de Markov cuyo marco teórico se presentó en la sección anterior. Allí se explicó que las cadenas de Markov se caracterizan por su matriz de transición y es fundamental poner a prueba si la misma es invariante a través del tiempo, o sus elementos dependen del tiempo. Se tomó a la estación meteorológica como estación de referencia, para la cual ya vimos que el esquema de Bernoulli no era adecuado para el modelado de ocurrencia de lluvia y se consideró una cadena de Markov homogénea de primer orden. Se estimaron las probabilidades de transición mediante el método de máxima verosimilitud y para poner a prueba si dichas probabilidades varían respecto del tiempo en forma anual, se consideró una ventana de seis días, donde suponemos que dentro de cada ventana las probabilidades de transición se mantienen invariantes.

Para estudiar la estacionariedad de las probabilidad de transición de las cadenas de Markov de orden 1 propuesta para el modelado de ocurrencia de lluvia, se hizo referencia al trabajo realizado por Anderson y Goodman (1957)[1], donde se estudiaron distintos aspectos estadísticos de las cadenas de Markov, entre ellos se realizó el desarrollo de una prueba de hipótesis para probar si la matriz de transición es invariante a través del tiempo. Se pone a prueba la hipótesis

$$H_0) p_{ij}(t) = p_{ij}(t = 1, \dots, T)$$

vs

$$H_1) \text{ no } H_0,$$

utilizando el test del cociente de máximas verosimilitudes. La máxima verosimilitud bajo la hipótesis nula:

$$\prod_{t=1}^T \prod_{i,j} \hat{p}_{ij}^{n_{ij}(t)},$$

y bajo la hipótesis alternativa la máxima verosimilitud es:

$$\prod_{t=1}^T \prod_{i,j} \hat{p}_{ij}(t)^{n_{ij}(t)}.$$

Por lo tanto el cociente de máximas verosimilitudes tiene la siguiente forma:

$$\lambda = \prod_{t=1}^T \prod_{i,j} \left[ \frac{\hat{p}_{ij}}{\hat{p}_{ij}(t)} \right]^{n_{ij}(t)}.$$

En nuestro estudio se tomaron intervalos de seis días, de manera que la cantidad de intervalos de tiempos es  $T = 366/6 = 61$ .  $m$  representa a la cantidad de estados de la cadena, por lo tanto aquí  $m = 2$ .

Bajo  $H_0$  el estadístico  $-2\log(\lambda)$  se distribuye asintóticamente  $\chi^2$  con  $(T-1) \cdot [m(m-1)] = 120$  grados de libertad, y por lo tanto para rechazar la hipótesis nula con un nivel de significación  $\alpha = 0,05$ , el valor de estadístico se debe encontrar en la región crítica determinada por el intervalo  $(146,57, +\infty)$ . Al realizar esta prueba en la estación meteorológica, obtuvimos que  $-2\log(\lambda) = 185,56$  por ende se rechazó la hipótesis de estacionariedad debido a que el valor del estadístico pertenece a la región crítica.

Análogamente se realizó la misma prueba para las demás estaciones en el departamento de Colonia llegando a idéntico resultado. Por lo tanto se concluyó que para el ajuste de datos de ocurrencia de lluvia se deben aplicar cadenas de Markov no estacionarias.

## 6.2. Desarrollo utilizando cadenas de Markov no estacionarias (Markov NE).

### 6.2.1. Estimación del modelo para las probabilidades de lluvia.

Stern y Coe (1984) [11] aplicaron modelos de tipo Markoviano no estacionario a datos de precipitación diaria, y propusieron técnicas para la selección de modelos.

Aquí se implementó dicha metodología a las estaciones pluviométricas de La Estanzuela, la meteorológica (las dos confiables) y a una estación pluviométrica con el propósito de comparar estimaciones provenientes de datos con distinto grado de confiabilidad.

La estación pluviométrica fue elegida de tal forma que no tuviera demasiados datos faltantes, y que fuera cercana a las estaciones de La Estanzuela y a la meteorológica a fin de poder comparar los resultados de la aplicación de los modelos propuestos. En la tabla 5.1 de la sección 5.1 se encuentra la matriz de distancia entre las estaciones de Colonia, y en la tabla 5.3 se encuentran los porcentajes de datos faltantes para dichas estaciones. Considerando ambas tablas se seleccionó la estación pluviométrica 2611.

A continuación se mostrará como aplicar la metodología de Stern y Coe con los datos de la estación meteorológica, luego se replicará la metodología a las estaciones de La Estanzuela y a la pluviométrica 2611.

A modo de poder comparar los resultados en las tres estaciones analizadas se considerará el período 1966-2008.

En la sección 6.1.2 se definió el concepto de cadenas de Markov y también se detalló como estimar la matriz de transición.

Siguiendo el estudio de Stern y Coe, en nuestro análisis se contempló la no estacionariedad para distintos períodos del año. Luego se ajustaron curvas continuas a las probabilidades de transición estimadas. La cantidad de curvas a ajustar depende del orden de la cadena de Markov. En el caso de una cadena de segundo orden, deben ser estimados 1464 valores ( $4 \times 366 = 1464$ ) dados los 366 días en el año (como máximo) y las cuatro probabilidades de transición. Las probabilidades de transición de una cadena de Markov NE de segundo orden están definidas por:

$$p_{hi,1}(t) = P[J(t) = 1/J(t-1) = i, J(t-2) = h] \text{ con } h, i = 0, 1; t = 1, \dots, 366$$

En el caso de una cadena de primer orden, deben ser estimados 732 valores ( $2 \times 366 = 732$ ) dados los 366 días en el año y las dos probabilidades de transición. Las probabilidades de transición de una cadena de Markov NE de primer orden están definidas por:

$$p_{i,1}(t) = P[J(t) = 1/J(t-1) = i], \text{ con } i = 0, 1; t = 1, \dots, 366$$

Dados los 366 días en el año, la estimación de las  $p_{hi}(t)$  se obtuvo mediante:

$$r_{hi,1}(t) = \frac{n_{hi1}(t)}{n_{hi+}(t)} \text{ con } h, i = 0, 1; t = 1, \dots, 366$$

Para contemplar la no estacionariedad de la variable, se utilizaron períodos de estimación cortos. En la tabla 5.5 de la sección 5.5 (estación meteorológica) se observa que en promedio llueven entre 5.2 y 7.0 días por mes, lo que se traduce en un día de lluvia cada 4.5–6 días. Por esta razón en nuestro estudio se optó por períodos de seis días, aspecto que facilitó la agrupación de los días. Gracias a esta agrupación la variable  $t$  en vez de variar de 1 a 366 lo hace de 1 a 61 ( $366/6$ ).

Para evaluar el orden de la cadena de Markov NE, se aplicó la prueba de máxima verosimilitud propuesta en el artículo de Anderson y Goodman [1] (página 99). Dado el modelo multinomial detrás de las cadenas de Markov NE, la log-verosimilitud del modelo para una cadena de Markov NE de segundo orden se calculó con la siguiente fórmula:

$$l = \sum_{t=1}^{61} \sum_{h,i} [n_{hi1}(t) \cdot \log(p_{hi,1}(t)) + n_{hi0}(t) \cdot \log(1 - p_{hi,1}(t))]$$

De la misma forma para cadenas de Markov NE de primer orden, la log-verosimilitud se calculó de la siguiente forma:

$$l = \sum_{t=1}^{61} \sum_i [n_{i1}(t) \cdot \log(p_{i,1}(t)) + n_{i0}(t) \cdot \log(1 - p_{i,1}(t))]$$

Antes de pasar a los resultados debemos mencionar que cadenas de tercer orden no fueron consideradas debido a problemas de estimación, ya que algunos parámetros serían evaluados con muy pocos datos.

Para la estación meteorológica se obtuvo una log-verosimilitud máxima de -8420.20 con cadenas de Markov NE de segundo orden y una log-verosimilitud máxima de -8329.07 con cadenas de Markov NE de primer orden. Para poner a prueba la hipótesis

$H_0$ ) alcanza con utilizar una cadena de Markov de primer orden

vs

$H_1$ ) es necesario utilizar una cadena de Markov de segundo orden,

se utilizó la prueba del cociente máxima de verosimilitudes, donde el estadístico de prueba  $G$  se definió de la siguiente manera:

$$G = -2 \cdot l(\text{Cadena de Markov de orden 1}) + 2 \cdot l(\text{Cadena de Markov de orden 2}).$$

Bajo  $H_0$ ) el estadístico de prueba  $G$  se distribuye  $\chi^2$  con  $m^2(m-1) - m(m-1) = 2$  grados de libertad. El valor de  $G$  para la estación meteorológica es 182.27 que cae dentro de la región crítica dada por el intervalo (5,99, + inf) a un nivel de significación del 5%. Por lo tanto se rechazó la hipótesis de que el modelo no mejora al aumentar el orden de la cadena de Markov NE.

Por lo tanto se optó en elegir cadenas de segundo orden y calcular los cuatro modelos que a ella le corresponden, estos son  $\hat{p}_{001}(t)$ ,  $\hat{p}_{011}(t)$ ,  $\hat{p}_{101}(t)$  y  $\hat{p}_{111}(t)$ .

Las probabilidades de transición  $\hat{p}_{000}(t)$ ,  $\hat{p}_{010}(t)$ ,  $\hat{p}_{100}(t)$  y  $\hat{p}_{110}(t)$  se calculan respectivamente a partir de los cuatro modelos anteriores, ya que  $\hat{p}_{000}(t) = 1 - \hat{p}_{001}(t)$ ,  $\hat{p}_{010}(t) = 1 - \hat{p}_{011}(t)$ ,  $\hat{p}_{100}(t) = 1 - \hat{p}_{101}(t)$  y  $\hat{p}_{110}(t) = 1 - \hat{p}_{111}(t)$ .

A los efectos de modelar la variable de respuesta dicotómica  $Y$  cuyas modalidades son lluvia (codificada con el uno) y no lluvia (codificada con el cero), se empleó el modelo de regresión logística. Dicho modelo se basa en asumir que:

$$P(Y = 1/X_i = x_i) = \frac{1}{1 + e^{-g(x_i)}} = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}},$$

o de manera complementaria:

$$P(Y = 0/X_i = x_i) = \frac{1}{1 + e^{g(x_i)}},$$

donde  $Y$  es la variable binaria a modelar,  $X_i$  representa al conjunto de variables explicativas, y  $g$  es una función de los datos. En nuestro caso el modelo logístico con el que se trabajó es el siguiente:

$$p_{hi,1}(t) = \frac{e^{g_{hi}(t)}}{1 + e^{g_{hi}(t)}}. \quad (1)$$

A partir de la fórmula (1), se puede escribir  $g_{hi,1}(t)$  como:

$$g_{hi,1}(t) = \ln\left(\frac{p_{hi,1}(t)}{1 - p_{hi,1}(t)}\right).$$

La función  $g$  puede tomar distintas formas, aquí se utilizaron series de Fourier:

$$g_{hi,1}(t) = a_{hi0} + \sum_{k=1}^m [a_{hik} \cdot \sin(kt') + b_{hik} \cdot \cos(kt')]$$

donde  $t' = \frac{2,6 \cdot \pi \cdot t - 3}{366}$  y  $t = 1, \dots, 61$ .

El número de armónicos,  $m$ , fue decidido con técnicas de regresión múltiple donde las variables explicativas ingresaron al modelo en un orden preestablecido. Modelos con mayores valores de  $m$  fueron estimados sucesivamente hasta encontrar un modelo que mejorara el ajuste.

En la tabla 6.2 se muestran los valores obtenidos de la log-verosimilitud para las cuatro probabilidades de transición variando  $m$ , en esta tabla 001 representa el modelo de ocurrencia de lluvia dado que no llovió en los dos días previos, 101 representa el modelo de ocurrencia de lluvia dado que no llovió el día anterior y llovió dos días antes, 011 representa el modelo de ocurrencia de lluvia dado que llovió el día anterior y no llovió dos días antes y 111 representa el modelo de ocurrencia de lluvia dado que llovió el día anterior y llovió en los dos días previos.

$m$	001	101	011	111
0	-4560.04	-1178.62	-1595.04	-1261.59
1	-4535.12	-1176.08	-1587.85	-1255.30
2	-4528.12	-1174.82	-1585.64	-1254.26
3	-4524.92	-1173.16	-1584.83	-1252.45
4	-4523.07	-1172.97	-1584.02	-1250.66
5	-4522.68	-1172.55	-1583.80	-1250.13
6	-4521.91	-1171.96	-1583.70	-1250.13
7	-4520.95	-1170.72	-1581.86	-1249.31

Cuadro 6.2: Log-verosimilitud para los distintas transiciones de una cadena de Markov NE de segundo orden variando  $m$ .

Para elegir el modelo con el  $m$  óptimo se utilizó el Criterio de Información de Akaike (AIC), el cual está basado en la penalización de la función log-verosimilitud mediante la cantidad de parámetros estimados:

$$AIC = -2l + \alpha(2m + 1),$$

donde  $2m + 1$  es la cantidad de parámetros utilizados,  $\alpha$  es el factor de ponderación que debe ser un número real mayor a cero (en nuestro caso  $\alpha = 1$ ) y  $l$  es la log-verosimilitud del modelo estimado.

En la tabla 6.3 se muestran los valores obtenidos de AIC para las cuatro probabilidades de transición variando  $m$ .

$m$	001	101	011	111
0	9121.08	2358.24	3191.08	2524.18
1	9073.25	2355.15	3178.69	2513.60
2	9061.24	2354.64	3176.27	2513.52
3	9056.85	2353.31	3176.65	2511.89
4	9055.14	2354.94	3177.04	2510.33
5	9056.36	2356.09	3178.60	2511.26
6	9056.82	2356.92	3180.40	2513.26
7	9056.90	2356.44	3178.72	2513.63
$m$ óptimo	4	3	2	4

Cuadro 6.3: AIC para las distintas transiciones de una cadena de Markov NE de segundo orden variando  $m$ .

las cuatro funciones de probabilidad de transición estimadas  $\hat{p}_{hi}(t)$  con sus respectivos intervalos de confianza al 95% (basados en el error standard de  $\hat{g}(t)$ ) se muestran en la siguiente gráfica, donde el tamaño de cada punto es proporcional a la cantidad de datos utilizado para la medición de la probabilidad de transición en el instante de tiempo en cuestión:

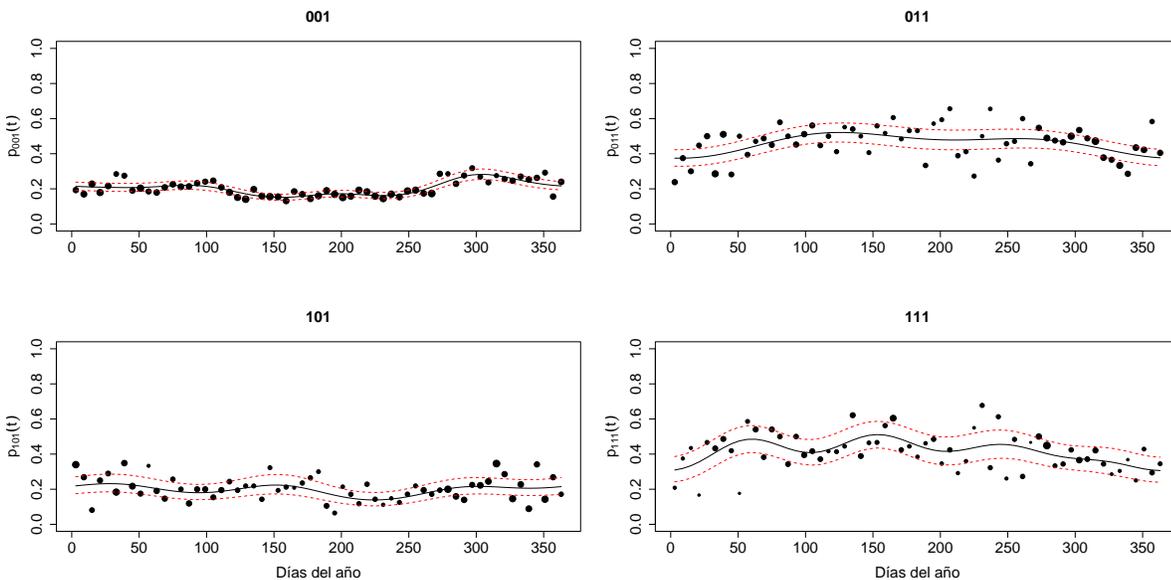


Figura 6.1: Funciones de probabilidades de transición estimadas: intervalos de confianza al 95% (punteado rojo) para la estación meteorológica.

### 6.2.2. Simulación del modelo estimado y comparación con la serie observada.

Una vez estimado el modelo para las probabilidades de precipitación, se simuló el mismo con el fin de compararlo con los datos observados.

La simulación consiste en replicar el modelo estimado anteriormente de forma de generar observaciones nuevas. Al realizar un número considerable de simulaciones (siempre y cuando

el modelo esté bien especificado) se obtiene la distribución de cualquier variable de interés que se desprenda a partir de los datos simulados. En nuestro caso la variable de interés es el número de días esperados para que llueva.

Para calcular la distribución de la variable objetivo, en primer lugar se simuló una cantidad de datos equivalente a 500 series pluviométricas. Luego se creó a partir de las simulaciones una tabla donde para cada día del año se encuentran las frecuencias simuladas de los días esperados para que llueva nuevamente.

Una vez construida la tabla con los datos simulados se compararon sin discriminar por momento del año, la distribución de los días esperados para que llueva con el modelo simulado (Distribución Teórica) y la serie de ocurrencia de lluvia en la estación meteorológica (Distribución con los datos observados).

En el gráfico 6.2 se muestra que las dos distribuciones son bastante similares entre si, sin embargo en los primeros tres días la probabilidad empírica es menor que en la teórica, lo que indica una falta en la cantidad de registros de días con lluvia. A continuación se analizó la cola de la distribución con los datos observados de forma individual ya que puede estar influenciado por el momento del año.

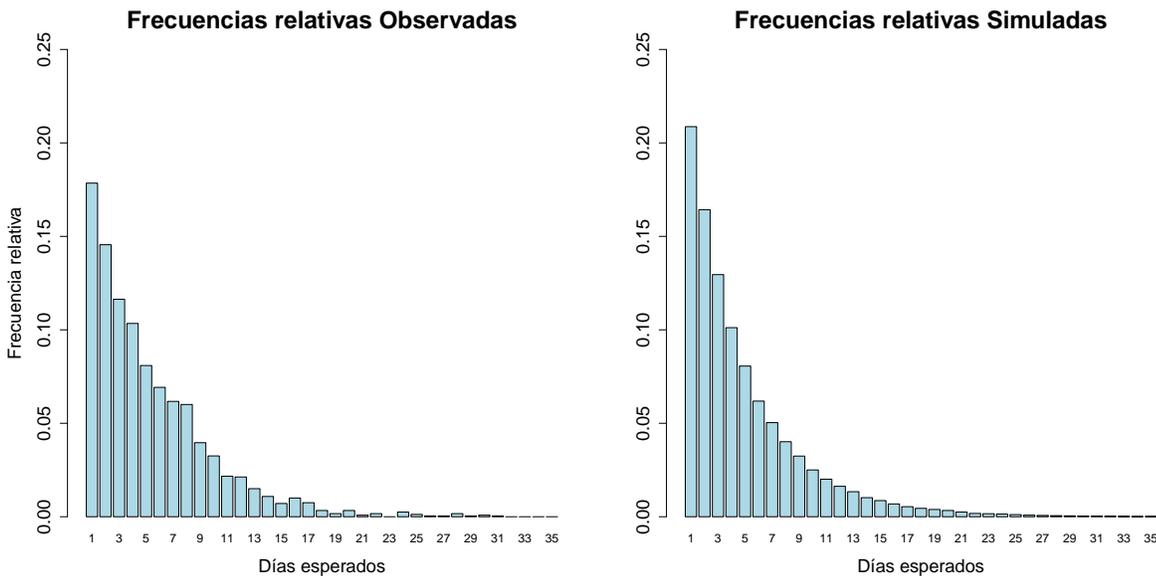


Figura 6.2: Distribución de los días esperados para que llueva con los datos observados en la estación meteorológica (izquierda) y con datos simulados (derecha).

Se analizó la cola de la distribución observada de los días esperados para que llueva si la presencia de la racha seca fue debida a una omisión en los registros o si la misma es probable.

Se define una racha como seca si la probabilidad de observar una racha igual o mayor de días secos es menor a un cierto umbral. Para este estudio se definió como umbral al 0,01 (1%) y además se calcularon los días con rachas secas mayores a 0,005 y 0,001 a modo más bien informativo.

La tabla 6.4 muestra las fechas donde se registraron rachas secas. Se observaron que cinco de las 18 rachas secas (27.8% aproximadamente) se encuentran en los primeros cuatro años de registro, lo cual puede indicar un marcado faltante de registros en esos años.

Por otro lado el modelo detectó rachas secas en períodos que de antemano se había

caracterizados por períodos secos. Por ejemplo entre el año 1988 y 1989 hubo un período seco y a partir del modelo se detectaron tres rachas secas en dicho período.

Año inicial	Mes inicial	Día inicial	Días esperados para lluvia	<i>Prob. &lt; 0,5 %</i>	<i>Prob. &lt; 0,1 %</i>
1966	7	30	31	Si	No
1966	11	11	18	No	No
1967	11	23	17	No	No
1968	3	30	30	Si	Si
1969	11	28	19	No	No
1977	4	9	21	No	No
1978	12	31	22	No	No
1988	5	17	31	No	No
1988	12	20	26	Si	No
1989	6	22	28	No	No
1994	11	9	16	No	No
1996	7	10	25	No	No
1997	7	14	32	Si	No
2000	2	28	21	No	No
2005	10	4	18	Si	No
2007	5	16	29	No	No
2008	10	27	18	Si	No
2008	12	2	19	No	No

Cuadro 6.4: Rachas secas donde la probabilidad de las mismas es menor al 1 % para la estación meteorológica.

Se estudió la sensibilidad del modelo en la modificación de la ventana utilizada para detección de rachas secas. Para ello se utilizaron además de la ventana de 6 días, ventanas de 5 y 7 días, en la tabla 6.5 se muestra la cantidad de rachas secas observadas utilizando una ventana de 5, 6 y 7 días, donde se observó que no hay prácticamente diferencias en la detección de rachas secas. Dicha tabla muestra la racha que empieza el día 8/5/1987 fue detectada por la ventana de 5 y 7 días pero no por la ventana de 6.

A partir de esto se deduce que el modelo no presenta grandes cambios al modificar el tamaño de la ventana, sugiriendo una cierta estabilidad en el modelo.

Año inicial	Mes inicial	Día inicial	Días esperados para lluvia	Ventana 5 días	Ventana 6 días	Ventana 7 días
1966	7	30	31	Si	Si	Si
1966	11	11	18	Si	Si	Si
1967	11	23	17	Si	Si	Si
1968	3	30	30	Si	Si	Si
1969	11	28	19	Si	Si	Si
1977	4	9	21	Si	Si	Si
1978	12	31	22	Si	Si	Si
1987	5	8	26	Si	No	Si
1988	5	17	31	Si	Si	Si
1988	12	20	26	Si	Si	Si
1989	6	22	28	Si	Si	Si
1994	11	9	16	Si	Si	Si
1996	7	10	25	No	Si	Si
1997	7	14	32	Si	Si	Si
2000	2	28	21	Si	Si	Si
2005	10	4	18	Si	Si	Si
2007	5	16	29	Si	Si	Si
2008	10	27	18	Si	Si	Si
2008	12	2	19	Si	Si	Si

Cuadro 6.5: Rachas secas donde la probabilidad de las mismas es menor al 1 % para la estación meteorológica.

### 6.3. Aplicación del modelo a la estación de La Estanzuela.

#### 6.3.1. Estimación del modelo para las probabilidades de lluvia.

En esta sección se aplicó el mismo procedimiento descrito anteriormente a la estación La Estanzuela, exponiendo sus resultados.

Se obtuvo una log-verosimilitud máxima de -8622.49 con cadenas de Markov NE de segundo orden y una log-verosimilitud máxima de -8684.41 con cadenas de Markov NE de primer orden. Para poner a prueba la hipótesis

$H_0$ ) alcanza con utilizar una cadena de Markov NE de primer orden

vs

$H_1$ ) es necesario utilizar una cadena de Markov NE de segundo orden,

se utilizó el estadístico de prueba G definido anteriormente, el mismo es igual a 123.83. Dado que bajo  $H_0$ ) el estadístico de prueba G se distribuye  $\chi^2$  con 2 grados de libertad, el valor resultante cae dentro de la región crítica establecida por el intervalo  $(5,99, +\infty)$  a un nivel de significación del 5 %, lo que llevó al rechazo de la hipótesis de que el modelo no mejora al aumentar el orden de la cadena de Markov NE. Por lo tanto se utilizaron cadenas de Markov NE de segundo orden.

En la tabla 6.6 se muestran los valores obtenidos de la log-verosimilitud para las cuatro probabilidades de transición variando el número de armónicos  $m$ .

$m$	001	101	011	111
0	-4633.05	-1202.16	-1628.30	-1313.50
1	-4614.15	-1200.68	-1623.81	-1300.40
2	-4609.23	-1199.93	-1621.94	-1296.33
3	-4605.65	-1197.45	-1621.82	-1294.79
4	-4604.55	-1197.32	-1620.67	-1292.52
5	-4604.45	-1197.30	-1619.43	-1292.02

Cuadro 6.6: Log-verosimilitud para los distintos estados de una cadena de Markov NE de segundo orden variando  $m$  para la estación La Estanzuela.

En la tabla 6.7 se muestran los valores obtenidos de AIC para las cuatro probabilidades de transición variando  $m$ , Se observa que los  $m$  óptimos para La Estanzuela son iguales a los de la estación meteorológica (ver tabla 6.3).

$m$	001	101	011	111
0	9267.10	2405.32	3257.60	2628.00
1	9231.29	2404.36	3250.61	2603.80
2	9223.47	2404.86	3248.88	2597.66
3	9218.31	2401.91	3250.65	2596.59
4	9218.10	2403.63	3250.35	2594.05
5	9219.91	2405.61	3249.85	2595.04
$m$ óptimo	4	3	2	4

Cuadro 6.7: AIC para los distintos estados de una cadena de Markov NE de segundo orden variando  $m$  para la estación La Estanzuela.

Las cuatro funciones de probabilidad de transición estimadas  $\hat{p}_{hi}(t)$  con sus respectivos intervalos de confianza al 95 % (basados en el error standard de  $\hat{g}(t)$ ) se muestran en la siguiente gráfica, donde el tamaño de cada punto es proporcional a la cantidad de datos utilizado para la medición de la probabilidad de transición en el instante de tiempo  $t$  en cuestión:

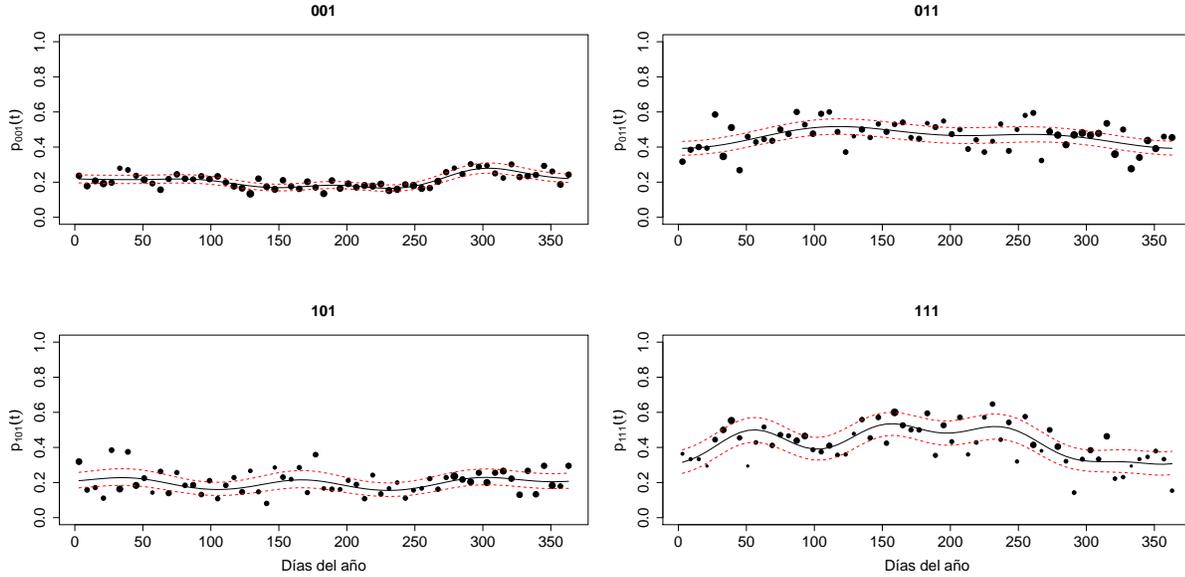


Figura 6.3: Funciones de probabilidades de transición estimadas: intervalos de confianza al 95 % (punteado rojo) para la estación de La Estanzuela.

### 6.3.2. Simulación del modelo estimado y comparación con la serie observada.

Un vez estimado el modelo para las probabilidades de precipitación, se simuló el mismo con el fin de compararlo con los datos observados.

Para obtener la distribución del número de días esperado para que llueva, en primer lugar se simuló una cantidad de datos equivalente a 500 series pluviométricas. Luego se creó a partir de las simulaciones una tabla donde para cada día del año se encuentran las frecuencias simuladas de los días esperados para que llueva nuevamente.

Una vez construida la tabla con los datos simulados se compararon sin discriminar por momento del año, la distribución de los días esperados para que llueva con el modelo simulado (Distribución Teórica) y la serie de ocurrencia de lluvia en la estación de La Estanzuela (Distribución con los datos observados).

En el gráfico 6.4 se muestra que las dos distribuciones son bastante similares entre si, sin embargo en los primeros tres días la probabilidad empírica es menor que en la teórica, lo que indica una falta en la cantidad de registros de días con lluvia.

Al comparar la distribución teórica del estimador en la estación de La Estanzuela y la misma en la estación meteorológica, se observó que las mismas son muy similares.

A continuación se analizó la cola de la distribución observada de forma individual ya que puede estar influenciada por el momento del año.

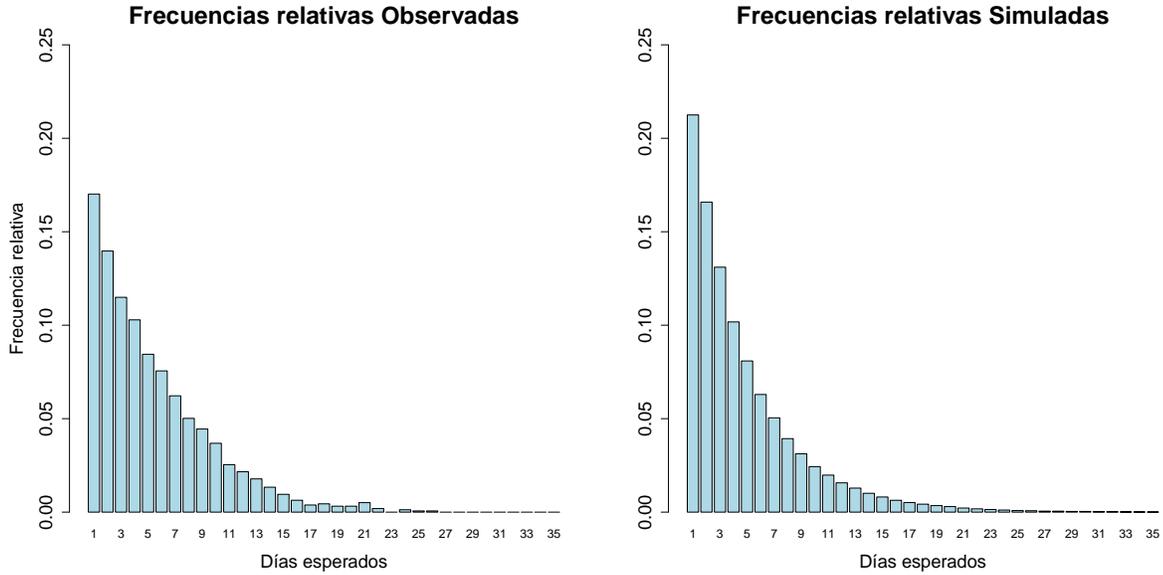


Figura 6.4: Distribución de los días esperados para que llueva con los datos observados en la estación La Estanzuela (izquierda) y con los datos simulados (derecha).

Finalmente para analizar la cola de la distribución observada, se estudiaron aquellos períodos donde la racha de días secos es larga, si la presencia de la misma fue debida a una omisión en los registros o si es probable.

En la tabla 6.8 se muestran los períodos donde se presentaron rachas secas. A diferencia del análisis realizado con la estación pluviométrica, aquí no hay ningún período que se destaque por una mayor cantidad de rachas secas.

Se destaca el caso de 4/5/1987 donde la serie muestra 58 días consecutivos con 0, lo que es muy dudoso, no siendo un problema del modelo, sino un problema de la serie.

Año inicial	Mes inicial	Día inicial	Días esperados para lluvia	Prob. < 0,5 %	Prob. < 0,1 %
1966	11	9	20	Si	No
1967	11	23	17	No	No
1973	11	1	15	No	No
1977	4	9	21	No	No
1978	12	31	22	No	No
1987	5	4	58	Si	Si
1988	12	20	26	Si	No
1989	6	22	27	No	No
1996	7	10	25	No	No
1997	2	19	19	No	No
2000	2	28	21	No	No
2001	9	8	21	No	No
2005	10	3	19	Si	No
2008	4	2	25	Si	No
2008	10	29	16	No	No

Cuadro 6.8: Rachas secas donde la probabilidad de las mismas es menor al 1% para la estación La Estanzuela.

## 6.4. Aplicación del modelo a la estación pluviométrica 2611.

### 6.4.1. Estimación del modelo para las probabilidades de lluvia.

En esta sección se aplicó el mismo procedimiento descrito anteriormente a la estación pluviométrica 2611, los resultados fueron los siguientes.

Se obtuvo una log-verosimilitud máxima de -6582.72 con cadenas de Markov NE de segundo orden y una log-verosimilitud máxima de -6680.27 con cadenas de Markov NE de primer orden. Para poner a prueba la hipótesis

$H_0$ ) alcanza con utilizar una cadena de Markov NE de primer orden

vs

$H_1$ ) es necesario utilizar una cadena de Markov NE de segundo orden,

se utilizó el estadístico de prueba G definido anteriormente, el mismo es igual a 195.11. Dado que bajo  $H_0$ ) el estadístico de prueba G se distribuye  $\chi^2$  con 2 grados de libertad, el valor resultante cae dentro de la región crítica establecida por el intervalo (5,99, + ínf) a un nivel de significación del 5 %, lo que llevó al rechazo de la hipótesis de que el modelo no mejora al aumentar el orden de la cadena de Markov NE. Por ende se utilizaron cadenas de Markov de segundo orden.

En la tabla 6.9 se muestran los valores obtenidos de la log-verosimilitud para las cuatro probabilidades de transición variando el número de armónicos  $m$ .

$m$	001	101	011	111
0	-4335.33	-743.39	-1131.75	-528.27
1	-4316.34	-741.97	-1129.23	-525.01
2	-4313.56	-742.30	-1128.73	-523.02
3	-4311.15	-741.45	-1127.93	-523.42
4	-4310.03	-739.45	-1127.79	-524.17
5	-4308.75	-739.07	-1127.47	-525.08

Cuadro 6.9: Log-verosimilitud para los distintos estados de una cadena de Markov NE de segundo orden variando  $m$  para la estación pluviométrica 2611.

En la tabla 6.10 se muestra los valores obtenidos de AIC para las cuatro probabilidades de transición variando  $m$ , esta muestra claros cambios en la estimación del  $m$  óptimo para la estación pluviométrica 2611 respecto al obtenido para las estación meteorológica y en La Estanzuela (ver tablas 6.3 y 6.7).

$m$	001	101	011	111
0	8671.65	1487.78	2264.50	1057.53
1	8635.67	1486.94	2261.46	1053.03
2	8632.12	1489.60	2262.46	1051.03
3	8629.29	1489.89	2262.86	1053.85
4	8629.06	1487.91	2264.58	1057.34
5	8628.51	1489.14	2265.95	1061.16
$m$ óptimo	5	1	1	2

Cuadro 6.10: AIC para los distintos estados de una cadena de Markov NE de segundo orden variando  $m$  para la estación pluviométrica 2611.

Las cuatro funciones de probabilidad de transición estimadas  $\hat{p}_{hi}(t)$  con sus respectivos intervalos de confianza al 95 % (basados en el error standard de  $\hat{g}(t)$ ) se muestran en la siguiente gráfica, donde el tamaño de cada punto es proporcional a la cantidad de datos

utilizado para la medición de la probabilidad de transición en el instante de tiempo  $t$  en cuestión:

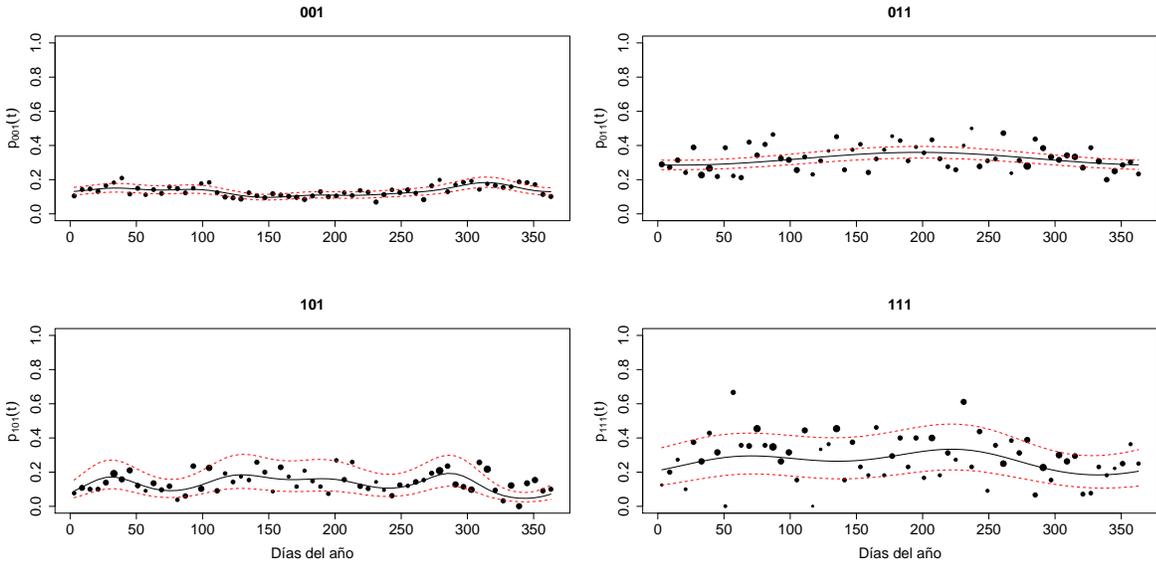


Figura 6.5: Funciones de probabilidades de transición estimadas: intervalos de confianza al 95 % (punteado rojo) para la estación pluviométrica 2611.

#### 6.4.2. Simulación del modelo estimado y comparación con la serie observada.

Un vez estimado el modelo para las probabilidades de precipitación, se simuló el mismo a modo de compararlo con los datos observados.

Para obtener la distribución del número de días esperados para que llueva, en primer lugar se simuló una cantidad de datos equivalente a 500 series pluviométricas. Luego se creó a partir de las simulaciones una tabla donde para cada día del año se encuentran las frecuencias simuladas de los días esperados para que llueva nuevamente.

Una vez construida la tabla con los datos simulados se compararon sin discriminar por momento del año, la distribución de los días esperados para que llueva con el modelo simulado (Distribución Teórica) y la serie de ocurrencia de lluvia en la estación pluviométrica 2611 (Distribución con los datos observados).

En el gráfico 6.6 se muestra que las dos distribuciones son bastante similares entre si, sin embargo en los primeros tres días la probabilidad empírica es menor que en la teórica, lo que indica una falta en la cantidad de registros de días con lluvia.

Al comparar la distribución teórica de la variable en la estación pluviométrica 2611 y la misma en la estación meteorológica, se observó que el recorrido de la distribución en la estación pluviométrica 2611 es mayor, ya que en estación meteorológica la serie tiende a cero a partir del día 29, no obstante en la estación pluviométrica 2611 la misma tiende a cero a partir del día 39.

A continuación se analizó la cola de la distribución observada de forma individual ya que puede estar influenciada por el momento del año.

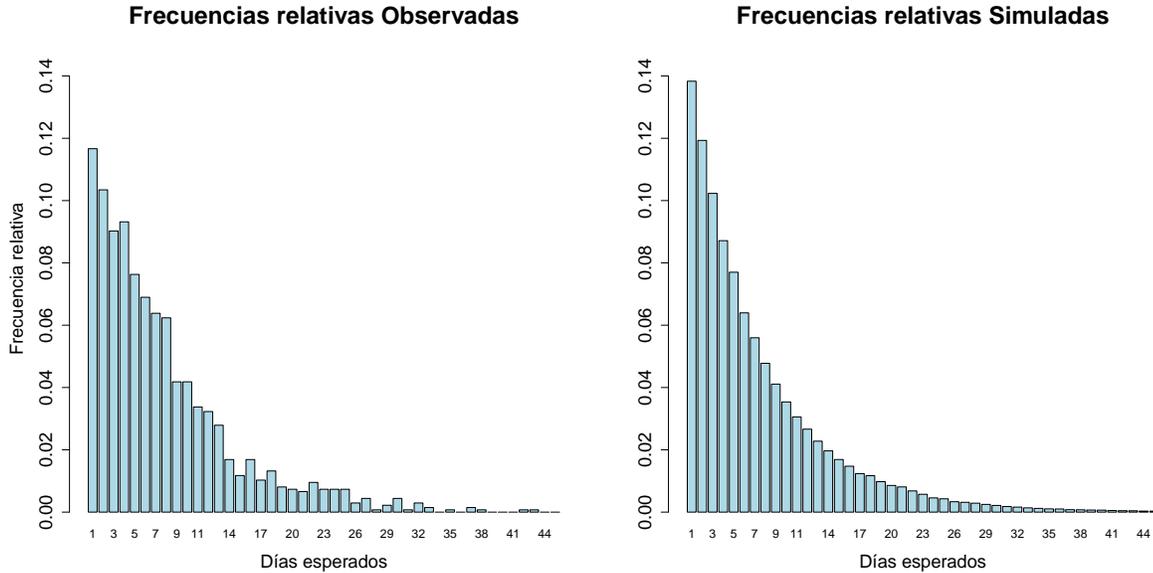


Figura 6.6: Distribución de los días esperados para que llueva con los datos observados en la estación pluviométrica 2611 (izquierda) y con los datos simulados (derecha).

Finalmente para analizar las colas de distribución, se estudió en aquellos períodos donde la racha de días secos es larga, si la presencia de las mismas fue debida a una omisión en los registros o si es probable la misma.

En la tabla 6.11 se muestran los días donde se presentaron rachas secas. A diferencia del análisis realizado con la estación meteorológica, aquí no hay ningún período donde predomine la cantidad de rachas secas.

El modelo en este caso no detectó una alta cantidad de rachas secas. Sin embargo, en el capítulo anterior se probó que esta estación tenía muchos más días sin precipitación que en las estaciones como la estación meteorológica, que tiene una mayor cantidad de rachas secas y además es cercana a la estación pluviométrica 2611.

Aquí el problema radica en que el modelo estimado para la estación pluviométrica 2611 se obtuvo a partir de los datos correspondientes y, como se vió en el gráfico anterior, el recorrido de la variable días esperados para que llueva en este modelo es mayor que en el modelo de la estación meteorológica, lo que implica que el modelo necesite una mayor cantidad de días secos para detectar una racha seca.

El promedio de la cantidad de días esperados para que llueva, dado que el período es considerado como una racha seca, es 35.75 días en la estación pluviométrica 2611, y es 23.39 días en la estación meteorológica, lo cual marca una diferencia de 12.36 días.

Año inicial	Mes inicial	Día inicial	Días esperados para lluvia	<i>Prob. &lt; 0,5 %</i>	<i>Prob. &lt; 0,1 %</i>
1967	11	12	43	Si	Si
1971	10	23	25	No	No
1972	9	17	32	No	No
1976	6	7	44	Si	No
1978	8	8	36	No	No
1981	8	8	36	No	No
1988	8	15	38	No	No
1993	6	9	50	Si	No
1997	2	10	38	Si	No
1999	11	7	26	No	No
2004	3	7	30	No	No
2008	10	26	31	No	No

Cuadro 6.11: Rachas secas donde la probabilidad de las mismas es menor al 1% en la estación pluviométrica 2611.

Una opción alternativa para la detección de rachas secas en la estación pluviométrica 2611 fue utilizar los parámetros del modelo hallado a partir de la estación meteorológica en la estación pluviométrica 2611, de la misma fuente de información y cercana a la estación en cuestión.

Al aplicar dicha opción se detectaron 85 rachas secas (ver tabla 6.12), lo cual demuestra que gran parte de la cantidad de rachas secas se deben pura y exclusivamente a una omisión en el registro del dato.

Año inicial	Mes inicial	Día inicial	Días esperados para lluvia	<i>Prob. &lt; 0,5 %</i>	<i>Prob. &lt; 0,1 %</i>
1966	6	23	28	No	No
1966	7	31	31	Si	No
1966	9	8	24	No	No
1966	11	12	17	No	No
1967	4	20	27	No	No
1967	11	12	43	Si	Si
1968	1	3	23	Si	No
1969	3	25	22	Si	No
1970	6	19	28	No	No
1970	11	21	17	No	No
1971	7	24	27	No	No
1971	10	23	25	Si	Si
1972	1	14	24	Si	No
1972	2	9	23	Si	No
1972	9	17	32	Si	Si
1973	3	15	28	Si	No
1973	7	27	28	No	No
1973	11	1	21	Si	No
1974	7	7	26	No	No
1974	12	20	21	No	No
1975	10	9	18	Si	No
1975	11	16	22	Si	No
1976	6	7	44	Si	Si
1977	4	9	25	Si	No
1978	5	15	30	No	No
1978	8	8	36	Si	No
1978	12	5	25	Si	No
1979	1	1	21	No	No
1979	2	12	20	No	No
1979	8	28	26	No	No
1979	12	15	21	No	No

Año inicial	Mes inicial	Día inicial	Días esperados para lluvia	<i>Prob. &lt; 0,5%</i>	<i>Prob. &lt; 0,1%</i>
1981	8	8	36	Si	No
1981	10	6	19	Si	No
1982	1	11	20	No	No
1982	2	21	23	Si	No
1982	11	3	24	Si	Si
1982	12	12	25	Si	No
1983	7	8	33	Si	No
1983	11	21	17	No	No
1984	7	11	31	Si	No
1984	8	12	33	Si	No
1984	11	11	20	Si	No
1984	12	12	21	No	No
1985	4	29	26	No	No
1985	12	18	25	Si	No
1986	2	15	23	Si	No
1986	12	20	25	Si	No
1987	5	3	31	Si	No
1988	1	29	24	Si	No
1988	8	15	38	Si	Si
1988	12	19	27	Si	No
1989	6	22	28	No	No
1990	6	5	33	Si	No
1990	7	16	29	Si	No
1990	9	11	22	No	No
1992	1	2	24	Si	No
1992	2	22	20	No	No
1993	6	9	50	Si	Si
1993	9	2	26	No	No
1994	2	17	25	Si	No
1994	5	9	34	Si	No
1994	11	9	19	Si	No
1995	8	1	28	No	No
1996	7	10	26	No	No
1996	11	10	19	Si	No
1997	2	10	38	Si	Si
1997	7	13	33	Si	No
1998	2	7	25	Si	No
1998	9	3	23	No	No
1999	9	6	31	Si	Si
1999	10	10	22	Si	No
1999	11	7	26	Si	Si
1999	12	5	22	Si	No
2000	2	28	21	No	No
2000	7	13	26	No	No
2000	11	27	18	No	No
2004	3	7	30	Si	Si
2004	9	7	31	Si	Si
2005	9	30	24	Si	Si
2005	10	25	20	Si	No
2005	12	9	30	Si	Si
2007	5	6	39	Si	No
2008	4	2	25	Si	No
2008	9	6	22	No	No
2008	10	26	31	Si	Si

Cuadro 6.12: Rachas secas donde la probabilidad de las mismas es menor al 1% en la estación pluviométrica 2611, utilizando el modelo hallado para la estación meteorológica.

## 6.5. Conclusiones del capítulo.

En este capítulo se mostró que el uso de la cadena de Markov NE es adecuado para el modelado de la ocurrencia de lluvia.

Dicho modelo manifestó la existencia de diferencias apreciables entre las estaciones de buena calidad y de mala calidad tanto en lo que refiere al número de armónicos utilizados como a las funciones de probabilidades de transición estimadas (especialmente en el estado 111).

La pequeña diferencia en la estimación de la probabilidad de lluvia en la estación de La Estanzuela y la estación meteorológica, muestra el efecto de la variación espacial del fenómeno estudiado. La diferencia en la estimación de la probabilidad de lluvia en la estación de La Estanzuela o la estación meteorológica, respecto a la estación pluviométrica, es resultado de la diferencia en la calidad de las series correspondientes.

En la etapa de selección de orden de la cadena de Markov NE para las tres estaciones analizadas se optó por utilizar el segundo orden para modelar el fenómeno. Sin embargo, al obtener el armónico óptimo se observaron diferencias entre las estaciones de buena calidad y la de mala calidad.

<i>m</i> óptimo	001	101	011	111
Meteorológica	4	3	2	4
La Estanzuela	4	3	2	4
Pluviométrica 2611	5	1	1	2

Cuadro 6.13: *m* óptimo por estación según estado de la cadena de Markov NE.

Observando las gráficas 6.1, 6.3 y 6.5 referidas a las funciones de probabilidad de transición estimadas, la amplia mayoría de los puntos están comprendidos en los respectivos intervalo de confianza al 95 %, lo cual proporciona evidencia a favor de la aplicación de los modelos propuestos.

Se estudió la sensibilidad en la modificación de la ventana utilizada para la detección de rachas secas. Donde se dedujo que el modelo no presenta grandes cambios al modificar el tamaño de la ventana en  $\pm 1$  día, sugiriendo una cierta estabilidad en el modelo estimado.

Una vez estimados los modelos para las probabilidades de precipitación, se simularon los mismos a modo de comparar las simulaciones con los datos observados. Para dicho objetivo se usaron las frecuencias de rachas secas simuladas y observadas.

Se comparó la distribución de los días esperados para que llueva bajo el modelo simulado y la distribución con los datos observados obteniéndose resultados bastante similares en el caso de las estaciones de buena calidad. Sin embargo la estación de peor calidad muestra problemas en la cola de la distribución, y las simulaciones del modelo indican que el número de días esperados para que llueva es mayor que en las estaciones de buena calidad.

Además a partir de la simulación de los modelos, se estimaron las rachas secas de cada estación. En primer lugar se observó que en la estación pluviométrica, el número esperado de días necesarios para que una racha de días secos fuera considerada como racha seca es en promedio 12.36 días mayor que lo esperado en la estación meteorológica. Luego se aplicó el modelo utilizado para la estación meteorológica a la estación pluviométrica, y se demostró que gran parte de la cantidad de rachas secas en esta última se debe pura y exclusivamente a una omisión en el registro del dato.

En segundo lugar comparando las rachas secas en las dos estaciones de buena calidad, se observó que nueve rachas secas coinciden en mes y año, lo cual representa para la estación meteorológica el 50.0 % de las rachas observadas y para la estación de La Estanzuela el 60.0 % de las rachas observadas.

## Capítulo 7

# Conclusiones.

El presente trabajo ha sido realizado en el marco del proyecto “A Europe-SouthAmerica Network For Climate Change Assessment and Impact Studies in La Plata Basin CLARIS-LPB, financiado por el 7mo programa marco de la Unión Europea del cual participa la Unidad de Ciencias de la Atmósfera de Facultad de Ciencias-UdelaR. Por medio de este proyecto se realizó un convenio con la Dirección Nacional de Meteorología (DNM), para la obtención y digitalización de la base de datos diaria de precipitación en Uruguay.

Es importante destacar que no existía hasta la fecha una base de datos digitalizada que además reúne datos provenientes de diferentes fuentes de información para el país.

Se conformó una base de datos diaria proveniente de las siguientes instituciones nacionales: Instituto Nacional de Investigación Agropecuaria (INIA), Usinas y Transmisiones Eléctricas (UTE) y DNM. El período analizado comprende los años 1960 a 2008.

Luego de la digitalización se comenzó el proceso de control de calidad de las series. A partir de 163 estaciones se realizó el mismo que constó de diferentes etapas. Una primera etapa de homogeneización del formato de los archivos, logrando una misma codificación del dato faltante, codificación de lluvia traza, codificación de mes entero sin lluvia, etc.(sección 3.1). Luego se corrigieron errores indiscutibles (por ejemplo valores negativos de precipitación, días inexistentes).

Actualmente el dato de precipitación diaria es medido de diferentes formas en las distintas instituciones. A modo de ejemplo, en la red pluviométrica de DNM dicho dato se mide como la precipitación acumulada en el período de 24 horas comprendido entre las 07:00 AM hora local de un día y las 07:00 del día siguiente, pero el registro se adjudica a la fecha del día anterior. Sin embargo, en las estaciones meteorológicas de DNM dicho dato es considerado como el acumulado entre las 09:00 AM hora local del día al cual se quiere asignar la precipitación y las 09:00 AM del día siguiente.

En el presente informe se tuvieron en cuenta las diferentes formas de medir el dato de precipitación diaria. La uniformidad en la definición de “día” para precipitación es muy importante sobre todo cuando se quiere realizar estudios de eventos extremos y para la comparación entre las series, por otro lado si hubo un evento extremo al comienzo o fin de un mes, la diferencia en la asignación de dicho evento puede cambiar el acumulado mensual.

El análisis de correlación entre las series permitió detectar la falta de un protocolo nacional para la medición de la precipitación diaria.

Como sugerencia a partir de este trabajo en lo que refiere a base de datos diaria de precipitación se recomienda:

- Comenzar a utilizar un archivo secuencial con las fechas ya prefijadas para evitar

errores en la generación del archivo calendario.

- Lograr uniformidad en la nomenclatura utilizada para los registros, de modo de lograr un protocolo nacional en la medición de la variable.
- Mejorar la forma en la cual se recaba la información en la DNM, ya que hasta el momento es por medio de comunicación telefónica.

Al realizar el estudio de datos faltantes se tomó el criterio de descartar las series que presentaran más de 10% de datos faltantes. Este porcentaje representa un aproximado de 5 años sin registros durante el período completo. Como el objetivo de la generación de esta base de datos es con fines de estudios climáticos y de eventos extremos dicho porcentaje es adecuado.

Por otro lado también se descartaron aquellas series en las que el período de registro era menor a 15 años. La aplicación de los mencionados criterios de depuración de la base de datos redujo el número de estaciones de 163 a 114, un 70% de las estaciones pluviométricas que había inicialmente. Las 114 estaciones comprenden todas las estaciones del INIA y las meteorológicas de la DNM, el 75% de las estaciones pluviométricas de la DNM, mientras que fueron descartadas todas las estaciones de UTE.

Visto la selección anterior y teniendo en cuenta, para cada departamento, tanto la cantidad de pluviómetros por cada mil  $km^2$  como la calidad de los pluviómetros, se seleccionó el departamento de Colonia, ya que representa el mejor departamento para llevar a cabo un análisis exhaustivo de la precipitación diaria y de esta forma poder probar metodologías estadísticas en el mejor escenario.

Colonia presenta datos provenientes de dos estaciones meteorológicas (INIA y DNM) y nueve estaciones pluviométricas. A partir de la comparación por década con la estación que presentaba el máximo de días con lluvia, las cuales fueron alternando siempre entre La Estanzuela y la meteorológica, se observa una diferencia muy grande en cuanto a días con registros sin lluvia. Esto muestra más registros de 0 en las estaciones pluviométricas que en las estaciones de referencia.

Se estudiaron las colas de la distribución de la cantidad de lluvia mensual, o sea las precipitaciones menores a 5 mm y mayores a 80 mm, suponiendo que una mayor frecuencia de eventos con precipitación mayores a 80 mm podrían ser la suma de dos o más eventos de lluvia menores. Si esto fuese así generaría uno o más ceros ficticios y un evento extremo que puede llegar también a ser ficticio. Por otro lado el estudio de la cola inferior nos muestra una acumulación de precipitaciones bajas que no se reportan como precipitación. La distribución de la función de densidad para precipitaciones menores a 5 mm muestra comportamientos totalmente diferentes entre las estaciones pluviométricas y las meteorológicas, donde se detectó una falta de precisión en la medición de la variable para estaciones pluviométricas, mientras que para el otro extremo las diferencias no son tan significativas.

Esto hace a las series pluviométricas difíciles de trabajar con los días secos ya que están subestimando las precipitaciones débiles.

Debido a estos aspectos se concluyó, en primer lugar, que hay una gran diferencia entre la calidad de las estaciones pluviométricas con respecto a La Estanzuela o la meteorológica. En segundo lugar, las estaciones pluviométricas proporcionan dudosa información estadística, debido a que se estimó para los registros de estaciones pluviométricas que por lo menos el 10% de días sin registros (o sin precipitación) son datos faltantes.

El análisis detallado que se realizó en el departamento de Colonia, pone de manifiesto la diferencia en cuanto a calidad de los datos provenientes de la red pluviométrica respecto

a las de las estaciones meteorológicas tanto de INIA como de DNM. Esto representa un problema, ya que la red pluviométrica es más densa que la red de estaciones meteorológicas lo cual genera ciertas limitaciones para poder realizar y concluir con confiabilidad algunos estudios referentes a precipitación en el Uruguay.

Por otro lado la reconstrucción de series diarias de precipitación por medios estadísticos es prácticamente imposible dada la variabilidad tanto espacial como temporal de la misma, el intento podría resultar en una serie aun menos confiable.

No obstante el comenzar a detectar los errores e ir desarrollando herramientas para un buen control de calidad es de suma importancia. Con dicho objetivo se generó un modelo de ocurrencia de lluvia. Se demostró que la serie de precipitación no es una variable estacionaria, si bien es algo que a nivel de las ciencias atmosféricas ya se sabe. Esta propiedad de la serie generó la búsqueda en la bibliografía de herramientas estadísticas adecuadas. Se utilizaron cadenas de Markov no estacionarias. Es importante mencionar el conocimiento que me aportó el aprendizaje de esta herramienta, hasta ahora nueva para mi.

El modelo diseñado para la probabilidad de ocurrencia de lluvia, se realizó tomando como referencia la metodología propuesta por Stern y Coe (1984). En dicho trabajo los autores utilizan cadenas de Markov no estacionarias. El modelo simula cada una de las probabilidades de transición de forma anual, agrupando los días del año en intervalos fijos y de esta forma lograr contar con una mayor cantidad de datos para la estimación de las probabilidades de transición en los distintos momentos del año. Se consideraron cadenas de Markov no estacionarias hasta de segundo orden, lo cuál permite utilizar hasta siete armónicos para modelar cada una de las probabilidades de transición para finalmente lograr la modificación en la ventana utilizada para estimar las mismas.

Una vez calibrado el modelo se llevaron a cabo simulaciones con el propósito de corroborar que el modelo reproduzca características deseadas de la estación considerada, tales como los períodos de rachas secas. Esta aplicación del modelo tiene la ventaja de detectar períodos de rachas secas indicando el momento de comienzo y duración de las mismas. Esta propiedad del modelo representa una herramienta muy útil para la caracterización de las secuencias secas, cuando el mismo se aplica a series confiables.

Para cada estación de interés se mostró como estimar las funciones de probabilidad de transición de la cadena de Markov no estacionaria, y la densidad de los días esperados para que llueva. El modelo propuesto permite discernir la calidad de las series observadas. El desempeño del modelo estimado con estaciones de buena calidad aplicado a estaciones de mala calidad evidenció la importancia estratégica de lograr una buena base de datos diaria de precipitación para el país.

Gracias al correcto modelo de la ocurrencia de lluvia podremos describir características de la precipitación las cuales tienen uso directo en diversos campos, uno de ellos es la planificación agrícola. A modo de ejemplo este modelado permite lograr una aproximación a la distribución de la humedad del suelo y conocer las longitudes de días consecutivos secos. La lluvia excesiva, además de las sequías, puede impedir la maduración y la cosecha. Claramente si se lograra una buena calidad en la base de datos, sería de suma importancia para la economía del país.

# Bibliografía

- [1] Anderson, T.W., Goodman, L.A., 1957: *Statistical Inference about Markov Chains* - The Annals of Mathematical Statistics 28 (1), 89-110.
- [2] Correa, J.C., González, N., 2002: *Gráficos estadísticos con R*. - Medellín: Universidad Nacional-Sede Medellín
- [3] Federico, S. , Pasqualoni, L., Avolio, E., Bellecci, C., 2010: *Calabria daily rainfall from 1970 to 2006*. - Nat. Hazards Earth Syst. Sci., 10, pg. 717-722.
- [4] Fuerza Aérea Argentina, Comando de Regiones Aéreas, 1985: *Pluviometría*.- Servicio Meteorológico Nacional. República Argentina.
- [5] Guttorp, P., 1995: *Stochastic Modelling of Scientific Data*. - Gran Bretania: Chapman & Hall/Crc.
- [6] Mittal, H.V., 2011: *R Graphs Cookbook*. - Birmingham: Packt Publishing Ltd.
- [7] Norris, J.R., 1997: *Markov Chain*. - New York: Cambridge University Press.
- [8] Paulhus, J.L., Kohler, M.A., 1952: *Interpolation of missing precipitation records*. - Washington: Monthly wheather review vol.80 nro.8 pg. 129-133.
- [9] Penalba, O., Llano, M.P., 2010. *A climatic analysis of dry sequences in Argentina*. - International journal of climatology 31, pg. 504-513.
- [10] Rodríguez-Iturbe, I., Mejía, J.M., 1974: *The Design of Rainfall Networks in Time and Space*. - Water Resources Research Vol.10, No. 4, pg.713-728.
- [11] Stern, R.D., Coe, R., 1984: *A Model fitting Analysis of Daily Rainfall Data*. - Journal of the Royal Statistical Society. Series A (General), Vol. 147, No. 1, pg.1-34.
- [12] Vicente-Serrano, S.M., Beguería, S., López-Moreno, J.I., García- Vera, M.A., Stepanek, P., 2010: *A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity*. - International journal of climatology 30, pg. 1146-1163.
- [13] Von Storch, H., Zwiers, F.W., 1999: *Statistical analysis in climate research*. - New York: Cambridge University Press
- [14] Wickham, H., 2011: *Ggplot2: Elegant Graphics for Data Analysis*. Houston: Springer.

# Apéndice A

## Tablas utilizadas.

### A.1. Tabla de la sección 2.1

Numero	Departamento	Longitud	Latitud	Año inicio	Año final
1001	Artigas	-57.0900	-30.1400	1960	2008
1013	Artigas	-57.5933	-30.2777	NA	NA
1016	Artigas	-57.3100	-30.2900	NA	NA
1019	Artigas	-56.9690	-30.2980	NA	NA
1021	Artigas	-56.8220	-30.2980	1960	2008
1035	Artigas	-56.5800	-30.4000	1960	2008
1040	Artigas	-57.4400	-30.4310	NA	NA
1044	Artigas	-57.1300	-30.4800	NA	NA
1047	Artigas	-56.7810	-30.4310	1960	2001
1048	Artigas	-56.6700	-30.4700	1960	2008
1050	Artigas	-56.5119	-30.3981	NA	NA
1054	Artigas	-57.6800	-30.5800	1960	2008
1062	Artigas	-56.9050	-30.6130	NA	NA
1066	Artigas	-56.4200	-30.4600	NA	NA
1066A	Artigas	-56.4010	-30.5040	NA	NA
1077	Artigas	-57.0500	-30.6000	NA	NA
1082	Artigas	-56.6300	-30.6200	NA	NA
1086	Artigas	-56.1600	-30.7000	1960	2008
1092	Artigas	-57.3250	-30.7190	NA	NA
1094	Artigas	-57.0300	-30.7800	1960	2008
1099	Artigas	-56.7700	-30.6300	NA	NA
1105	Salto	-57.7780	-30.7890	1960	2008
1109	Artigas	-57.4100	-30.8500	NA	NA
1115	Artigas	-56.7800	-30.7800	1960	2008
1120	Artigas	-56.2500	-30.7800	1960	1977
1141	Artigas	-56.3700	-30.8800	1960	2008
1147	Rivera	-55.5426	-30.8965	1912	2004
1153A	Salto	-57.5280	-30.9460	1960	1993
1159	Artigas	-56.8760	-31.0000	NA	NA
1162	Salto	-56.6800	-31.0800	1960	2008

## A.2. Tabla de la sección 5.4.3

Estación	2471	2476	2480C	2570	2611	2657A	2662A	2702	2737	Met. 2774	La Estanz.
2471	1.00	0.66	0.59	0.59	0.61	0.63	0.56	0.59	0.56	0.53	0.55
2476	0.66	1.00	0.64	0.69	0.63	0.68	0.68	0.67	0.61	0.55	0.60
2480C	0.59	0.64	1.00	0.69	0.60	0.60	0.63	0.61	0.62	0.54	0.61
2570	0.59	0.69	0.69	1.00	0.64	0.64	0.61	0.62	0.63	0.65	0.71
2611	0.61	0.63	0.60	0.64	1.00	0.62	0.57	0.57	0.63	0.62	0.68
2657A	0.63	0.68	0.60	0.64	0.62	1.00	0.69	0.71	0.67	0.63	0.68
2662A	0.56	0.68	0.63	0.61	0.57	0.69	1.00	0.75	0.61	0.54	0.57
2702	0.59	0.67	0.61	0.62	0.57	0.71	0.75	1.00	0.63	0.58	0.59
2737	0.56	0.61	0.62	0.63	0.63	0.67	0.61	0.63	1.00	0.71	0.72
Met. 2774	0.53	0.55	0.54	0.65	0.62	0.63	0.54	0.58	0.71	1.00	0.82
La Estanz.	0.55	0.60	0.61	0.71	0.68	0.68	0.57	0.59	0.72	0.82	1.00

Cuadro A.1: Matriz de correlación diaria para el departamento de Colonia.

Estación	2471	2476	2480C	2570	2611	2657A	2662A	2702	2737	Met. 2774	La Estanz.
2471	1.00	0.91	0.90	0.90	0.91	0.90	0.89	0.89	0.90	0.84	0.84
2476	0.91	1.00	0.90	0.90	0.91	0.91	0.91	0.90	0.90	0.84	0.84
2480C	0.90	0.90	1.00	0.91	0.90	0.90	0.90	0.90	0.90	0.84	0.85
2570	0.90	0.90	0.91	1.00	0.90	0.90	0.89	0.89	0.90	0.87	0.87
2611	0.91	0.91	0.90	0.90	1.00	0.91	0.89	0.89	0.91	0.85	0.85
2657A	0.90	0.91	0.90	0.90	0.91	1.00	0.91	0.91	0.91	0.86	0.86
2662A	0.89	0.91	0.90	0.89	0.89	0.91	1.00	0.92	0.89	0.83	0.83
2702	0.89	0.90	0.90	0.89	0.89	0.91	0.92	1.00	0.90	0.84	0.84
2737	0.90	0.90	0.90	0.90	0.91	0.91	0.89	0.90	1.00	0.86	0.86
Met. 2774	0.84	0.84	0.84	0.87	0.85	0.86	0.83	0.84	0.86	1.00	0.91
La Estanz.	0.84	0.84	0.85	0.87	0.85	0.86	0.83	0.84	0.86	0.91	1.00

Cuadro A.2: Matriz de correlación mensual diaria para el departamento de Colonia.

Estación	2471	2476	2480C	2570	2611	2657A	2662A	2702	2737	Met. 2774	La Estanz.
2471	1.00	0.83	0.80	0.80	0.78	0.77	0.75	0.75	0.74	0.78	0.79
2476	0.83	1.00	0.85	0.89	0.80	0.83	0.82	0.84	0.81	0.79	0.83
2480C	0.80	0.85	1.00	0.86	0.76	0.79	0.81	0.80	0.80	0.79	0.80
2570	0.80	0.89	0.86	1.00	0.78	0.82	0.83	0.85	0.80	0.81	0.83
2611	0.78	0.80	0.76	0.78	1.00	0.78	0.74	0.74	0.78	0.80	0.85
2657A	0.77	0.83	0.79	0.82	0.78	1.00	0.83	0.82	0.84	0.84	0.90
2662A	0.75	0.82	0.81	0.83	0.74	0.83	1.00	0.86	0.80	0.78	0.81
2702	0.75	0.84	0.80	0.85	0.74	0.82	0.86	1.00	0.81	0.83	0.86
2737	0.74	0.81	0.80	0.80	0.78	0.84	0.80	0.81	1.00	0.89	0.89
Met. 2774	0.78	0.79	0.79	0.81	0.80	0.84	0.78	0.83	0.89	1.00	0.90
La Estanz.	0.79	0.83	0.80	0.83	0.85	0.90	0.81	0.86	0.89	0.90	1.00

Cuadro A.3: Matriz de concordancia diaria para el departamento de Colonia.

# Apéndice B

## Código utilizado.

### B.1. Depurado de base de datos, análisis y confección inicial.

```
#bibliotecas
library(foreign)
library(lattice)

x=read.dbf("Estación A Testear.dbf")           #Se abre la estación deseada.
head(x)                                       #Se observa si esta bien el encabezado
dim(x)                                        #Si lo lee bien como una tabla de datos
colnames(x)=c("ESTACION","A.0","MES","DIA","PRECIPITAC") #Se inserta un encabezado en común
año.inicial=x$A.0[1]
año.final=x$A.0[dim(x)[1]]
#####
#Número de años bisiestos y número de días #
#####

años.bisiestos=0
for (i in año.inicial:año.final)
{if(i/4==floor(i/4))
{años.bisiestos=años.bisiestos+1}
}
nro.dias=(2009-1960)*365+años.bisiestos
#####
#Particularidades que pueden existir en la serie #
#####
#Datos NA
e=x$PRECIPITAC
a=is.na(e)
dato=rep(1,length(e))
for(i in 1:length(e))
{if (a[i]==TRUE){(dato[i]=0)}
}
x=cbind(x,dato)
filt1=x[dato==1,]           #Primer filtrado de los datos, se quitan los datos NA
x[dato==0,]                #Si hay dato==0 entonces hay que cambiar el NA por -99.9 ya que no hay registro

#Precipitacion negativa
s.r=rep(1,dim(filt1)[1])
pr.neg=filt1$PRECIPITAC<0
x=cbind(x,s.r)
x[x$s.r==0,]               #El dato sin registro es el -99.9 no se admitirá otro

filt2=x[x$s.r==1,]        #Solo escogemos los datos con pp>=0

#####
#Ordenamiento de los datos por año #
#####
ord.año=c(sort(filt2$A.0)==filt2$A.0)      #Igualo el ordenado a los datos reales
```

```

indicadora=1
año.cambiado=NULL
i=1
while ( i<=length(ord.año) & indicadora==1)
{if (ord.año[i]==FALSE)
{indicadora=0
año.cambiado=filt2$A.0[i]}
i=i+1}
indicadora          # Si es cero mirar los datos, hay un año que esta mal.
año.cambiado        # Si existe el año que esta mal, aca te especifica cual es

#Si supera el supuesto de que los años están ordenados entonces se vera si están ordenados por mes

#####
#Ordenamiento de los datos por año y mes      #
#####
mes.cambiado=NULL
i=año.inicial
indicadora=1
while ( i <= año.final)
{filt2i=filt2[filt2$A.0==i,]
ord.mes=c(sort(filt2i$MES)==filt2i$MES)
j=1
while ( j<=length(ord.mes) & indicadora==1)
{if (ord.mes[j]==FALSE)
{indicadora=0
mes.cambiado=c(filt2i$A.0[j],filt2i$MES[j])}
j=j+1}
i=i+1
}

indicadora          # Si es cero mirar los datos, hay un mes que esta mal.
mes.cambiado        # Si existe el mes que esta mal, acá te especifica cual es

#Si supera el supuesto de que los años y meses están ordenados entonces se vera si están ordenados por día
#####
#Ordenamiento de los datos por año,mes y día  #
#####
dia.cambiado=NULL

for (i in año.inicial:año.final)
{filt2i=filt2[filt2$A.0==i,]
for (j in 1:12)
{filt2ij=filt2i[filt2i$MES==j,]
nul=0
ord.dia=c(sort(filt2ij$DIA)==filt2ij$DIA)
k=1

while ( k<=dim(filt2ij)[1] & indicadora==1 & nul==0)
{if (ord.dia[k]==FALSE)
{indicadora=0
dia.cambiado=rbind(dia.cambiado,c(filt2ij$A.0[k],filt2ij$MES[k],filt2ij$DIA[k]))}
k=k+1}
}
}
}
dia.cambiado
if (!is.null(dia.cambiado))
{colnames(dia.cambiado)=c("A.0","MES","DIA")}

#Si este resultado dio null seguimos con el análisis, si no hay días que están desordenados.
#####
#Datos raros, seguramente debido a mal registro      #
#####

#Meses fuera de los comunes
x[(x$MES<=0 | x$MES>12) & dato==1,]

```

```

#Dias negativos
x[x$DIA<=0 & dato==1,]          # Si hay algun dia negativo o cero, pero con pp>=0 corregir

#Mes febrero
dr=x[x$DIA>28 & x$MES==2 & dato==1,]
error=NULL
for(i in 1:dim(dr)[1]) {
if( dr$A.0[i]/4 != floor(dr$A.0[i]/4) & dr$DIA[i]==29)
{error=rbind(error,dr[i,])}
if( dr$DIA[i]!=29)
{error=rbind(error,dr[i,])}
}
error

#Meses de 30 días con mas de 30 días
x[x$DIA>30 & (x$MES==4|x$MES==6 |x$MES==9 |x$MES==11) & dato==1,]

#Meses de 31 días con más de 31 días
x[x$DIA>31 & (x$MES==1 |x$MES==3 |x$MES==5 |x$MES==7 |x$MES==8 |x$MES==10 |x$MES==12) & dato==1,]

#####
#Confección de los datos          #
#####
x=x[,1:5]

#Importante destacar estas 2 tablas

meses.sin.registro=x[x$PRECIPITAC<0,]

meses.sin.lluvia=x[x$PRECIPITAC==0,]
#####
#Se agrega ceros en los dias que no llovio,          #
#Esto es para que funcione todo el programa          #
#####

for(i in 1:dim(x)[1]){
if (x$PRECIPITAC[i]<=0){
x$DIA[i]=1
x$PRECIPITAC[i]=0
} }

x.comp=NULL
for (i in 1960:2008)
{xi=x[x$A.0==i,]
for(j in 1:12)
{xi[j]=xi[xi$MES==j,]

if (j==1 |j==3 |j==5 |j==7 |j==8 |j==10 |j==12)
{xi[j].rep=matrix(data=rep(0,31*5), nrow =31, ncol =5, byrow = FALSE,dimnames = NULL)
xi[j].rep=data.frame(xi[j].rep)
colnames(xi[j].rep)=colnames(xi[j])
for(k in 1:dim(xi[j])[1])
{xi[j].rep[xi[j]$DIA[k],]= xi[j][k,]
}
for(k in 1:31)
{
if (xi[j].rep$PRECIPITAC[k]==0){ xi[j].rep[k,]= t(c(xi[j][1,1],xi[j][1,2],xi[j][1,3],k,0))}
}
x.comp=rbind(x.comp,xi[j].rep)
}

if(j==4|j==6 |j==9 |j==11){
xi[j].rep=matrix(data=rep(0,30*5), nrow =30, ncol =5, byrow = FALSE,dimnames = NULL)
xi[j].rep=data.frame(xi[j].rep)
colnames(xi[j].rep)=colnames(xi[j])
for(k in 1:dim(xi[j])[1])
{xi[j].rep[xi[j]$DIA[k],]= xi[j][k,]
}
}
}

```

```

}
for(k in 1:30)
{
if (xij.rep$PRECIPITAC[k]==0){ xij.rep[k,]= t(c(xij[1,1],xij[1,2],xij[1,3],k,0))}
}
x.comp=rbind(x.comp,xij.rep)
}

if(j==2 & floor((i-1960)/4)==(i-1960)/4){
xij.rep=matrix(data=rep(0,29*5), nrow =29, ncol =5, byrow = FALSE,dimnames = NULL)
xij.rep=data.frame(xij.rep)
colnames(xij.rep)=colnames(xij)
for(k in 1:dim(xij)[1])
{xij.rep[xij$DIA[k],]= xij[k,]}
for(k in 1:29)
{
if (xij.rep$PRECIPITAC[k]==0){ xij.rep[k,]= t(c(xij[1,1],xij[1,2],xij[1,3],k,0))}
}
x.comp=rbind(x.comp,xij.rep)
}

if(j==2 & floor((i-1960)/4)!= (i-1960)/4){
xij.rep=matrix(data=rep(0,28*5), nrow =28, ncol =5, byrow = FALSE,dimnames = NULL)
xij.rep=data.frame(xij.rep)
colnames(xij.rep)=colnames(xij)
for(k in 1:dim(xij)[1])
{xij.rep[xij$DIA[k],]= xij[k,]}
for(k in 1:28)
{
if (xij.rep$PRECIPITAC[k]==0){ xij.rep[k,]= t(c(xij[1,1],xij[1,2],xij[1,3],k,0))}
}
x.comp=rbind(x.comp,xij.rep)
}
}}

#Debe coincidir el número de días con la dimensión de los datos

dim(x.comp)[1]==nro.dias
x.comp[dim(x.comp)[1],]

#Promedio de días lluviosos mensuales por encima de 1 mm.

prom.dias.lluv=cbind(sort(rep(1960:2008,12)),rep(1:12,2009-1960),rep(0,588))

for (i in 1960:2008)
{xi=x.comp[x.comp$A.0==i,]
for(j in 1:12)
{xij=xi[xi$MES==j,]
prom=0
for(k in 1:dim(xij)[1]){
if(xij$PRECIPITAC[k]>0){prom=prom+1}}
prom.dias.lluv[(i-1960)*12+j,3]=round(prom/dim(xij)[1],3)
}}

write.table(x.comp ,file = "promdiaslluv.txt", dec =".")

#Además de la serie con los ceros colocamos una columna con los días de "no lluvia" y los días sin registro.

meses.sin.registro
x.comp=cbind(x.comp,rep(0,dim(x.comp)[1]),rep(0,dim(x.comp)[1]))
colnames(x.comp)[6]=c("DATO.FALTANTE")
colnames(x.comp)[7]=c("NO.LLUVIA")

for(i in 1:dim(meses.sin.registro)[1])
{for(j in 1:length(which(x.comp$A.0==meses.sin.registro[i,2] & x.comp$MES==meses.sin.registro[i,3])))
{x.comp[which(x.comp$A.0==meses.sin.registro[i,2] & x.comp$MES==meses.sin.registro[i,3])[j],6]=1 }}

for(i in 1:dim(meses.sin.lluvia)[1])

```

```

{for(j in 1:length(which(x.comp$A.0==meses.sin.lluvia[i,2] & x.comp$MES==meses.sin.lluvia[i,3]))
{if (x.comp[which(x.comp$A.0==meses.sin.lluvia[i,2] & x.comp$MES==meses.sin.lluvia[i,3])[j],5]==0)
{x.comp[which(x.comp$A.0==meses.sin.lluvia[i,2] & x.comp$MES==meses.sin.lluvia[i,3])[j],7]=1}}}

x.comp[x.comp$NO.LLUVIA==1,]
nro.datos.falt=dim(x.comp[x.comp$DATO.FALTANTE==1,])[1]
nro.no.lluvia=dim(x.comp[x.comp$NO.LLUVIA==1,])[1]

#Se crea una secuencia a modo de detectar saltos

x.comp=cbind(x.comp,seq(1,dim(x.comp)[1]))
colnames(x.comp)[8]=c("SECUENCIA")
x.comp1=x.comp[x.comp$DATO.FALTANTE==0,] # aquí está la secuencia con los saltos

#####
#Maxima lluvia en 5 días #
#####

max.5dias=0
pos=0
i=1
while(i<=(dim(x.comp1)[1]-6))
{
if(x.comp1$SECUENCIA[i+5]==x.comp1$SECUENCIA[i]+5)
{
if (sum(x.comp1[i:(i+4),5])>max.5dias)
{
max.5dias=sum(x.comp1[i:(i+4),5])
pos=i}
i=i+1}
if(x.comp1$SECUENCIA[i+5]>x.comp1$SECUENCIA[i]+5)
{
u=1
while(x.comp1$SECUENCIA[i+u]==x.comp1$SECUENCIA[i]+u & i+u<=(dim(x.comp1)[1]-5))
{u=u+1}
i=i+u}
}
max.5dias
pos
round(c(nro.datos.falt,nro.no.lluvia,nro.datos.falt*100/nro.dias,
nro.no.lluvia*100/nro.dias,max(x.comp$PRECIPITAC),max.5dias,x.comp1[pos,2],4)

write.table(x.comp ,file = "lluviadiaria.txt",row.names = FALSE, dec = ".")

```

## B.2. Análisis de datos faltantes.

```

x=read.table("datosfaltantes(analizar).txt",header=TRUE,dec=",") #Tabla que contiene el porcentaje de datos
#faltantes por mes para cada estación del uruguay

estac=read.table("estacionesDatosFaltantes.txt",header=TRUE)

MediasMens=rep(0,588)
cant.na=rep(0,588)
for (i in 1:588)
{MediasMens[i]=sum(x[i,3:165],na.rm=TRUE)/(163-sum(is.na(x[i,3:165])))
#Los primeros dos xs son año y mes es el porcentaje de datos faltantes dentro de las estaciones
#que tienen algún registro

cant.na[i]=sum(is.na(x[i,3:165]))}
#Los primeros dos xs son año y mes es el número de datos faltantes dentro de las estaciones
#que tienen algún registro

MediasMens=cbind(x[,c(1,2)],MediasMens,cant.na)
MediasMens=cbind(MediasMens,163-cant.na) #163-cant.na es la cantidad de estaciones disponibles

#####
#Grafico cantidad de estaciones por año#

```

```
#####

cant.est.anual=rep(0,49)
for (i in 1:49)
  {cant.est.anual[i]=mean(MediasMens[((i-1)*12+1):(i*12),5])}
cant.est.anual=t(as.matrix(cant.est.anual))
colnames(cant.est.anual)=unique(MediasMens[,1])
barplot(cant.est.anual,col="lightblue", ylab ="Nro de estaciones",xlab="Año",ylim=c(0,200))
#abline(h=150,col="red")

#####
#Grafico de porcentaje de meses faltantes por año#
#####

cant.dat.fal.anual=rep(0,49)
for (i in 1:49)
  {cant.dat.fal.anual[i]=mean(MediasMens[((i-1)*12+1):(i*12),3])}
cant.dat.fal.anual=t(as.matrix(cant.dat.fal.anual))
colnames(cant.dat.fal.anual)=unique(MediasMens[,1])
barplot(cant.dat.fal.anual,col="lightblue", ylab ="Porcentaje de datos faltantes",xlab="Año",ylim=c(0,.2))

abline(h=0.10,col="red")

#####
#Cantidad de datos faltantes por mes#
#####

cant.dat.fal.mes=rep(0,12)
for (i in 1:12)
  {MediasMensi=MediasMens[MediasMens[,2]==i,]
  cant.dat.fal.mes[i]=mean(MediasMensi[,3])}

cant.dat.fal.mes=t(as.matrix(cant.dat.fal.mes))
colnames(cant.dat.fal.mes)=c("Enero", "Febrero", "Marzo", "Abril", "Mayo", "Junio", "Julio", "Agosto",
, "Setiembre", "Octubre", "Noviembre", "Diciembre")
barplot(cant.dat.fal.mes,col="lightblue", ylab ="porcentaje de datos faltantes",xlab="mes",ylim=c(0,.08))
#abline(h=.07,col="red")

#####
#Cantidad de datos faltantes por mes por decada#
#####
par(mfrow=c(2,3))
decada=c("1960-1969", "1970-1979", "1980-1989", "1990-1999", "2000-2008")
inicio=c(1960,1970,1980,1990,2000)
fin=c(1969,1979,1989,1999,2008)
for(k in 1:5)
{MediasMensk=MediasMens[MediasMens$a.o>=inicio[k] & MediasMens$a.o<=fin[k],]
cant.dat.fal.mes=rep(0,12)
for (i in 1:12)
  {MediasMensik=MediasMensk[MediasMensk[,2]==i,]
  cant.dat.fal.mes[i]=mean(MediasMensik[,3])}
cant.dat.fal.mes=t(as.matrix(cant.dat.fal.mes))
colnames(cant.dat.fal.mes)=c("Ene", "Febr", "Mar", "Abr", "May", "Jun", "Jul", "Ago", "Set", "Oct", "Nov", "Dic")
barplot(cant.dat.fal.mes,col="lightblue", ylab ="porcentaje de datos faltantes por decada",
,xlab="mes",main=decada[k],ylim=c(0,0.14))
}

#####
#Separo las estaciones de cada una de las fuentes de informacion#
#####
inia=c("inia.lasbrujas", "inia.laestanz", "inia.salto", "inia.tacua", "inia.treinytre")
ute=c("X1147", "X1220", "X1224", "X1379", "X1405", "X1440", "X1454", "X1558", "X1880", "X1915", "X2016", "X2073",
, "X2215", "X1709", "X2486")
meteo=c("X2774", "X2206", "X2297", "X2583", "X2909", "X2887", "X2867", "X2179")

inia.df=x[,inia]
ute.df=x[,ute]
meteo.df=x[,meteo]
```

```

dnm.df=x[, (colnames(x) %in% meteo)==FALSE]
dnm.df=dnm.df[, (colnames(dnm.df) %in% inia)==FALSE]
dnm.df=dnm.df[, (colnames(dnm.df) %in% ute)==FALSE]

#####
#Datos faltantes por mes según fuente de información #
#####
par(mfrow=c(2,2))

#DNM
MediasMens=rep(0,588)
cant.na=rep(0,588)
for (i in 1:588)
  {MediasMens[i]=sum(dnm.df[i,3:dim(dnm.df)[2]],na.rm=TRUE)/(dim(dnm.df)[2]-sum(is.na(dnm.df[i,3:dim(dnm.df)[2]])))
  cant.na[i]=sum(is.na(dnm.df[i,3:dim(dnm.df)[2]]))}

MediasMens=cbind(x[,c(1,2)],MediasMens,cant.na)
MediasMens=cbind(MediasMens,(dim(dnm.df)[2]-2)-cant.na)

cant.dat.fal.mes=rep(0,12)
for (i in 1:12)
  {MediasMensi=MediasMens[MediasMens[,2]==i,]
  cant.dat.fal.mes[i]=mean(MediasMensi[,3])}

cant.dat.fal.mes=t(as.matrix(cant.dat.fal.mes))
colnames(cant.dat.fal.mes)=c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
,"Setiembre","Octubre","Noviembre","Diciembre")
barplot(cant.dat.fal.mes,col="blue",ylab="porcentaje de datos faltantes",xlab="mes",ylim=c(0,.08)
,main="Estaciones Red Pluviométrica DNM")

#UTE
MediasMens=rep(0,540) #va hasta 2004
cant.na=rep(0,540)
for (i in 1:540)
  {MediasMens[i]=sum(ute.df[i,1:dim(ute.df)[2]],na.rm=TRUE)/(dim(ute.df)[2]-sum(is.na(ute.df[i,1:dim(ute.df)[2]])))
  cant.na[i]=sum(is.na(ute.df[i,1:dim(ute.df)[2]]))}

MediasMens=cbind(x[1:540,c(1,2)],MediasMens,cant.na)
MediasMens=cbind(MediasMens,(dim(ute.df)[2]-cant.na))

cant.dat.fal.mes=rep(0,12)
for (i in 1:12)
  {MediasMensi=MediasMens[MediasMens[,2]==i,]
  cant.dat.fal.mes[i]=mean(MediasMensi[,3])}

cant.dat.fal.mes=t(as.matrix(cant.dat.fal.mes))
colnames(cant.dat.fal.mes)=c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
,"Setiembre","Octubre","Noviembre","Diciembre")
barplot(cant.dat.fal.mes,col="darkblue",ylab="porcentaje de datos faltantes",xlab="mes"
,ylim=c(0,.30),main="Estaciones Ute")

#Meteorológica
MediasMens=rep(0,588)
cant.na=rep(0,588)
for (i in 1:588)
  {MediasMens[i]=sum(meteo.df[i,1:dim(meteo.df)[2]],na.rm=TRUE)/(dim(meteo.df)[2]-sum(is.na(meteo.df[i,1:dim(meteo.df)[2]])))
  cant.na[i]=sum(is.na(meteo.df[i,1:dim(meteo.df)[2]]))}

MediasMens=cbind(x[1:588,c(1,2)],MediasMens,cant.na)
MediasMens=cbind(MediasMens,(dim(meteo.df)[2]-cant.na))

cant.dat.fal.mes=rep(0,12)
for (i in 1:12)
  {MediasMensi=MediasMens[MediasMens[,2]==i,]
  cant.dat.fal.mes[i]=mean(MediasMensi[,3])}

```

```

cant.dat.fal.mes=t(as.matrix(cant.dat.fal.mes))
colnames(cant.dat.fal.mes)=c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
,"Setiembre","Octubre","Noviembre","Diciembre")
barplot(cant.dat.fal.mes,col="skyblue", ylab ="porcentaje de datos faltantes",xlab="mes",ylim=c(0,0.03)
,main="Estaciones Meteorológica")

#INIA
MediasMens=rep(0,(588-66))
cant.na=rep(0,(588-66))
for (i in 67:588)
  {MediasMens[i-67]=sum(inia.df[i,1:dim(inia.df)[2]],na.rm=TRUE)/(dim(inia.df)[2]-sum(is.na(inia.df[i,1:dim(inia.df)[2]])))
  cant.na[i-67]=sum(is.na(inia.df[i,1:dim(inia.df)[2]]))}

MediasMens=cbind(x[67:588,c(1,2)],MediasMens,cant.na)
MediasMens=cbind(MediasMens,(dim(inia.df)[2]-cant.na))

cant.dat.fal.mes=rep(0,12)
for (i in 1:12)
  {MediasMensi=MediasMens[MediasMens[,2]==i,]
  cant.dat.fal.mes[i]=mean(MediasMensi[,3])}

cant.dat.fal.mes=t(as.matrix(cant.dat.fal.mes))
colnames(cant.dat.fal.mes)=c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
,"Setiembre","Octubre","Noviembre","Diciembre")
barplot(cant.dat.fal.mes,col="lightblue", ylab ="porcentaje de datos faltantes",xlab="mes"
,main="Estaciones Inia",ylim=c(0,0.003))

```

### B.3. Mapas utilizados.

```

#Transforma coordenadas espaciales a cartesianas
rm(list=ls())
library(sp)
load(url("http://gadm.org/data/rda/URY_adm1.RData")) #se guarda como una variable llamada gadm
a=gadm

datos=read.table("distanciasabril2011.txt",header=T)
datos$latitud=datos$latitud/100000
datos$longitud=datos$longitud/100000
datos=cbind(datos,rep(0,dim(datos)[1]),rep(0,dim(datos)[1]))
colnames(datos)[c(6,7)]=c("existe","años.total")
#Esto se hace porque algunas estaciones que están en esta tabla en realidad no existen.

attach(datos)
for(i in 1:dim(datos)[1])
{if(is.na(datos$Año.inicio[i])==0)
{datos$existe[i]=1
datos$años.total[i]=Año.final[i]-Año.inicio[i]+1
}
}
detach(datos)
datos=datos[datos$existe==1,] # me quedo solo con los que existen

#####
#Distribución geográfica de los puntos en el mapa de Uruguay #
#####

par(oma=c(1,1,1,1),font=2,cex=0.5)
plot(a)
points(datos[,2],datos[,3],type="p",col="orange",pch=19)
grid(10,10)
text(datos[,2],datos[,3],datos[,1],cex=.8)

#####
#Distribución geográfica de los puntos en el mapa de Uruguay según años de registro #
#####

```

```

datos=cbind(datos,rep(0,dim(datos)[1]))
colnames(datos)[8]=c("cl_anual")
for (i in 1:dim(datos)[1])
{
if (datos$años.total[i]<16){datos$cl_anual[i]="Menor o igual a 15 años"}
if (datos$años.total[i]<31 & datos$años.total[i]>=16){datos$cl_anual[i]="Entre 16 y 30 años"}
if (datos$años.total[i]<50 & datos$años.total[i]>=31){datos$cl_anual[i]="Entre 31 y 50 años"}
if (datos$años.total[i]>=50 & datos$años.total[i]<98){datos$cl_anual[i]="Mas de 50 años"}
}

datos$cl_anual=factor(datos$cl_anual)

attach(datos)
plot(a)
points(datos[,2],datos[,3],col= factor(cl_anual),pch=19)
grid(12,12)
legend(-54,-30, unique(datos$cl_anual)[c(4,2,1,3)],col=c(unique(factor(cl_anual)))[c(4,2,1,3)],
text.col = "green4",bg = 'gray90',title.col="black", cex = 1, pch=19,title="Años de registro")

#####
#Distribución geográfica de los puntos en el mapa de Uruguay según fuente de información #
#####

detach(datos)
x=read.table("datos importantes a usarcn datos.txt",header=TRUE,sep = "\t",dec=",")
#x contiene varios datos importantes de cada estación como por ejemplo el periodo utilizado,
#el numero de datos faltantes, el porcentaje de días sin lluvia, el departamento,
#la máxima cantidad de precipitación acumulada en 5 días, Etc.

#1:143 DNM
#144:158 UTE
#159:164 inia
meteo=c("2774", "2206", "2297", "2583", "2909", "2887", "2867", "2179")
x=cbind(x,c(rep("DNM",143),rep("UTE",15),rep("INIA",5)))
colnames(x)[16]="Fuente_de_informacion"
x$Fuente_de_informacion=as.character(x$Fuente_de_informacion)
x$Fuente_de_informacion[x$Estacion %in% meteo]="Meteorológica"

#se agrega longitud y latitud a la base general
x=cbind(x,rep(0,dim(x)[1]),rep(0,dim(x)[1]))
colnames(x)[c(17,18)]=c("longitud","latitud")
for (i in 1:dim(x)[1])
{
aux=which(x$Estacion[i]==datos$estacion)
x$longitud[i]=datos$longitud[aux]
x$latitud[i]=datos$latitud[aux]
}

x1=x[,c(16,17,18)]
plot(a)
points(x1[,2],x1[,3],col= factor(x1[,1]),pch=19)
grid(12,12)
legend(-54,-30, unique(x1[,1]),col=c(unique(factor(x1[,1]))),
text.col = "green4",bg = 'gray90',title.col="black", cex = 0.5, pch=19,title="Fuente de información")

#####
#Distribución geográfica de los puntos en el mapa de Uruguay según % de datos faltantes #
#####
x=cbind(x,rep(0,dim(x)[1]))
colnames(x)[19]=c("cl_porc_dat_falt")
for (i in 1:dim(x)[1])
{
if (x$Porcentaje.de.datos.faltantes[i]<1){x$cl_porc_dat_falt[i]="Menor a 1%"}
if (x$Porcentaje.de.datos.faltantes[i]<2 & x$Porcentaje.de.datos.faltantes[i]>=1)

```

```

        {x$cl_porcentaje_faltante[i]="Entre 1 y 2 %"}
    if (x$Porcentaje.de.datos.faltantes[i]<5 & x$Porcentaje.de.datos.faltantes[i]>=2)
        {x$cl_porcentaje_faltante[i]="Entre 2 y 5 %"}
    if (x$Porcentaje.de.datos.faltantes[i]>=5 & x$Porcentaje.de.datos.faltantes[i]<10)
        {x$cl_porcentaje_faltante[i]="Entre 5 y 10 %"}
    if (x$Porcentaje.de.datos.faltantes[i]>=10){x$cl_porcentaje_faltante[i]="Mas de 10 %"}
}

attach(x)
plot(a)
points(x[,17],x[,18],col= factor(cl_porcentaje_faltante),pch=19)
grid(12,12)
par(oma=c(1,1,2,1),new=T,font=2,cex=0.5)
legend(-54,-30, unique(cl_porcentaje_faltante),col=c(unique(factor(cl_porcentaje_faltante))),
      text.col = "green4",bg = 'gray90',title.col="black", cex = 1 , pch=19,title="% de datos faltantes")
detach(x)

#####
#Distribución geográfica de los puntos según % de datos faltantes pero con menos de 10% de los mismos #
#####
x1=x[x$Porcentaje.de.datos.faltantes<10,]
x1=cbind(x1,rep(0,dim(x1)[1]))
colnames(x1)[19]=c("cl_porcentaje_faltante")
for (i in 1:dim(x1)[1])
{
    if (x1$Porcentaje.de.datos.faltantes[i]<1){x1$cl_porcentaje_faltante[i]="Menor a 1%"}
    if (x1$Porcentaje.de.datos.faltantes[i]<2 & x1$Porcentaje.de.datos.faltantes[i]>=1)
        {x1$cl_porcentaje_faltante[i]="Entre 1 y 2 %"}
    if (x1$Porcentaje.de.datos.faltantes[i]<5 & x1$Porcentaje.de.datos.faltantes[i]>=2)
        {x1$cl_porcentaje_faltante[i]="Entre 2 y 5 %"}
    if (x1$Porcentaje.de.datos.faltantes[i]>=5 & x1$Porcentaje.de.datos.faltantes[i]<10)
        {x1$cl_porcentaje_faltante[i]="Entre 5 y 10 %"}
}

attach(x1)
plot(a)
points(x1[,17],x1[,18],col= factor(cl_porcentaje_faltante),pch=19)
grid(12,12)
par(oma=c(1,1,2,1),new=T,font=2,cex=0.5)
legend(-54,-30, unique(cl_porcentaje_faltante),col=c(unique(factor(cl_porcentaje_faltante))),
      text.col = "green4",bg = 'gray90',title.col="black", cex = 1.3, pch=19,title="% de datos faltantes")
detach(x1)

table(x$Fuente_de_informacion,x$cl_porcentaje_faltante)[,c(5,1,2,3,4)]
library(xtable)
xtable(table(x$Fuente_de_informacion,x$cl_porcentaje_faltante)[,c(5,1,2,3,4)])

#####
#Distribución geográfica de los puntos según % de días sin lluvia #
#####
x=cbind(x,rep(0,dim(x)[1]),rep(0,dim(x)[1]))
colnames(x)[c(20,21)]=c("porcentaje_dias_sin_lluvia","cl_dias_sin_lluvia")
x[,20]=100*x$Nro.de.dias.no.lluvia/(x$Nro.dias.total-x$Nro.datos.faltantes)
x$porcentaje_dias_sin_lluvia=as.numeric(x$porcentaje_dias_sin_lluvia)

for (i in 1:dim(x)[1])
{
    if (x$porcentaje_dias_sin_lluvia[i]<75){x$cl_dias_sin_lluvia[i]="Menor a 75%"}
    if (x$porcentaje_dias_sin_lluvia[i]<80 & x$porcentaje_dias_sin_lluvia[i]>=75){x$cl_dias_sin_lluvia[i]="Entre 75 y 80 %"} #el 75 es el umbral
    if (x$porcentaje_dias_sin_lluvia[i]<85 & x$porcentaje_dias_sin_lluvia[i]>=80){x$cl_dias_sin_lluvia[i]="Entre 80 y 85 %"}
    if (x$porcentaje_dias_sin_lluvia[i]>=85){x$cl_dias_sin_lluvia[i]="Mas de 85 %"}
}

table(x$Fuente_de_informacion,x$cl_dias_sin_lluvia)[,c(4,1,2,3)]
xtable(table(x$Fuente_de_informacion,x$cl_dias_sin_lluvia)[,c(4,1,2,3)])

```

```

attach(x)
plot(a)
points(x[,17],x[,18],col= factor(cl_dias_sl),pch=19)
grid(12,12)
par(oma=c(1,1,2,1),new=T,font=2,cex=0.5)
legend(-54,-30, unique(cl_dias_sl)[c(4,3,1,2)],col=c(unique(factor(cl_dias_sl)))[c(4,3,1,2)],
      text.col = "green4",bg = 'gray90',title.col="black", cex = 1 ,
      pch=19,title="% de días sin lluvia")
detach(x)

table(x$cl_porcentaje_falt,x$cl_dias_sl)[c(5,1,2,3,4),c(4,1,2,3)]

table(x[x$Fuente_de_informacion=="DNM"],$cl_porcentaje_falt,x[x$Fuente_de_informacion=="DNM"],$cl_dias_sl)[c(5,1,2,3,4),]

x=cbind(x,0,0)
colnames(x)[c(22,23)]=c("años.total","cl_anual")
x$años.total=x$año.final-x$año.inicio+1

for (i in 1:dim(x)[1])
{
if (x$años.total[i]<16){x$cl_anual[i]="Menor o igual a 15 años"}
if (x$años.total[i]<31 & x$años.total[i]>=16){x$cl_anual[i]="Entre 16 y 30 años"}
if (x$años.total[i]<50 & x$años.total[i]>=31){x$cl_anual[i]="Entre 31 y 50 años"}
if (x$años.total[i]>=50 & x$años.total[i]<98){x$cl_anual[i]="Mas de 50 años"}
}
#Se quitan los datos que no cumplen con determinadas condiciones
limpieza=x[x$cl_porcentaje_falt!="Mas de 10 %" & x$cl_dias_sl!="Mas de 85 %" & x$cl_anual!="Menor o igual a 15 años",]
plot(a)
points(limpieza[,17],limpieza[,18],pch=19,col="darkblue")
grid(10,10)

xtable(table(limpieza$Fuente_de_informacion))

#####
# Mapa de Uruguay según densidad pluviométrica por departamento#
#####
#Vector de densidades pluv.
vec=c(0.84,0.59, 0.88,1.80,0.17,0.78,0.10,0.80,0.42, .59,0.36,0.97,0.11,0.57,0.71,1.20,1.00,0.71,0.42)
a$Shape_Area =vec
library(RColorBrewer)
mypalette<-brewer.pal(9,"YlGn")
mypalette=sort(rep(mypalette,3),decreasing=TRUE)
spplot(a,"Shape_Area",col.regions=mypalette,names.attr=a$NAME_1)
#####

lim=limpieza
limpieza$Fuente_de_informacion=factor(limpieza$Fuente_de_informacion)
levels(limpieza$Fuente_de_informacion)=c("black","red","red")
limpieza=cbind(limpieza,limpieza$Fuente_de_informacion)
colnames(limpieza)[24]="tipo"
levels(limpieza$tipo)=c(19,24)
plot(a)
points(limpieza[limpieza$Fuente_de_informacion=="red",17],limpieza[limpieza$Fuente_de_informacion=="red",18],
col="red",pch=17)
points(limpieza[limpieza$Fuente_de_informacion=="black",17],limpieza[limpieza$Fuente_de_informacion=="black",18],
col="black",pch=19)
grid(10,10)
legend(-55,-30, c("Estaciones Meteorológica o Inia","Estaciones DNM"),col=c("red","black"),
      text.col = "green4",bg = 'gray90',title.col="black", cex = .8, pch=c(17,19))

levels(limpieza[,2])[146]="Met. Colonia"
levels(limpieza[,2])[159]="La Estanzuela"
limpieza[limpieza$Departamento=="Colonia",2][10]="Met. Colonia"
limpieza[limpieza$Departamento=="Colonia",2][11]="La Estanzuela"
levels(limpieza[,2])[146]="Met. Colonia"

```

```

levels(limpieza[,2])[159]="La Estanzuela"

plot(a,xlim=c(-58.5,-57),ylim=c(-34.5,-33.5))
points(limpieza[limpieza$Fuente_de_informacion=="red",17],limpieza[limpieza$Fuente_de_informacion=="red",18]
,col="red",pch=17)
points(limpieza[limpieza$Fuente_de_informacion=="black",17],limpieza[limpieza$Fuente_de_informacion=="black",18]
,col="black",pch=19)
grid(10,10)
text(limpieza[limpieza$Departamento=="Colonia",17],limpieza[limpieza$Departamento=="Colonia",18],
limpieza[limpieza$Departamento=="Colonia",2],cex=.8,pos=2,col="red")

```

## B.4. Análisis de estaciones pluviométricas en colonia.

```

library(lattice)
require(grDevices)
data=read.table("BaseDatosColonial1est.txt",header=TRUE)
#Tabla que contiene todos los datos de precipitación para las estaciones de Colonia en el período 1960-2008
y=read.table("matrizdedistancias.txt",header=TRUE,dec=",")
#Matriz que contiene todas las distancias entre estaciones.

estacionesColonia=as.character(c("X2471", "X2476", "X2480C", "X2570", "X2611", "X2657A", "X2662A"
, "X2702", "X2737", "X2774", "inia.laestanz"))
#Va con X adelante para que lo iguale a la matriz de distancias
estacionesColoniaGr=as.character(c("Estación 2471", "Estación 2476", "Estación 2480C"
, "Estación 2570", "Estación 2611", "Estación 2657A"
, "Estación 2662A", "Estación 2702", "Estación 2737", "Estación 2774 Met."
, "Estación La Estanz", "total")) #Solo para los gráficos

estacionesColoniaSE=as.character(c("2471", "2476", "2480C", "X2570", "2611", "2657A", "2662A", "2702", "2737", "2774"))
#Sin la estanzuela

#####
#Se ubican todas las estaciones de Colonia en la matriz de distancias #
#####
y1=y[,-1] # Quitar fila de id
ubMatDis=NULL
for(i in 1:length(estacionesColonia)){ubMatDis=c(ubMatDis,which(estacionesColonia[i]==colnames(y1)))}
ubMatDis

#Ahora construyo la matriz de distancias mas chica

distCol=y1[ubMatDis,ubMatDis]
colnames(distCol)=estacionesColonia
rownames(distCol)=colnames(distCol)
distCol

nroest=length(estacionesColonia)
data=data[,-1] # saco la fila de id

#####
#Lluvia mensual, numero de días con lluvia, máximo mensual #
#####

a=unique(data[,c(1,2)])
lluvia.mensual=cbind(a,matrix(rep(0,dim(a)[1]*(nroest+1)),length(rep(0,dim(a)[1])),(nroest+1)))
lluvia.mensual=rbind(lluvia.mensual,rep(0,((nroest+3))))
colnames(lluvia.mensual)[c(3:(nroest+2))]=estacionesColonia
colnames(lluvia.mensual)[(nroest+3)]="Total"
dias.lluviosos=lluvia.mensual
maximo.mensual=lluvia.mensual
head(lluvia.mensual)
for(i in 1:(dim(lluvia.mensual)[1]-1))
{data1=data[data[,1]==lluvia.mensual[i,1] & data[,2]==lluvia.mensual[i,2],]
for(j in 1:nroest)
{if (length(data1[data1[,3+2*j]>=0,3+2*j])!=dim(data1)[1])
{lluvia.mensual[i,2+j]=NaN
dias.lluviosos[i,2+j]=NaN

```

```

maximo.mensual[i,2+j]=NaN}
if (length(data1[data1[,3+2*j]>=0,3+2*j])==dim(data1)[1])
{lluvia.mensual[i,2+j]=sum(data1[data1[,3+2*j]>0,3+2*j])
dias.lluviosos[i,2+j]=length(data1[data1[,3+2*j]>0,3+2*j])
maximo.mensual[i,2+j]=max(data1[,3+2*j])}
}}
dias.lluviosos$Total=rowMeans(dias.lluviosos[,c(3:(nroest+2))], na.rm =TRUE)
dias.lluviosos[dim(dias.lluviosos)[1,]=colMeans(dias.lluviosos[-dim(dias.lluviosos)[1,], na.rm = TRUE)
dias.lluviosos[dim(dias.lluviosos)[1],c(1,2)] =c(0,0)

lluvia.mensual$Total=rowMeans(lluvia.mensual[,c(3:12)], na.rm =TRUE)
lluvia.mensual[dim(lluvia.mensual)[1,]=colMeans(lluvia.mensual[-dim(lluvia.mensual)[1,], na.rm = TRUE)
lluvia.mensual[dim(lluvia.mensual)[1],c(1,2)] =c(0,0)

maximo.mensual$Total=rowMeans(maximo.mensual[,c(3:12)], na.rm =TRUE)
maximo.mensual[dim(maximo.mensual)[1,]=colMeans(maximo.mensual[-dim(maximo.mensual)[1,], na.rm = TRUE)
maximo.mensual[dim(maximo.mensual)[1],c(1,2)] =c(0,0)

#####
#Cantidad de precipitación mensual #
#####
lluvia.por.mes=matrix(rep(0,4*12),4,12)
for (i in 1:12)
{a=lluvia.mensual[lluvia.mensual$MES==i,]
lluvia.por.mes[1,i]=mean(a$Total,na.rm=TRUE)
lluvia.por.mes[2,i]=sd(a$Total,na.rm=TRUE)
lluvia.por.mes[c(3,4),i]=quantile(a$Total,na.rm=TRUE,c(0.05,0.95))}
colnames(lluvia.por.mes)=c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio",
,"Agosto","Setiembre","Octubre","Noviembre","Diciembre")
rownames(lluvia.por.mes)=c("Media Lluvia Mensual","SD Lluvia Mensual","IC inferior","IC superior")

#Días lluviosos por mes.
dias.lluviosos.por.mes=matrix(rep(0,4*12),4,12)
for (i in 1:12)
{a=dias.lluviosos[dias.lluviosos$MES==i,]
dias.lluviosos.por.mes[1,i]=mean(a$Total,na.rm=TRUE)
dias.lluviosos.por.mes[2,i]=sd(a$Total,na.rm=TRUE)
dias.lluviosos.por.mes[c(3,4),i]=quantile(a$Total,na.rm=TRUE,c(0.05,0.95))}
colnames(dias.lluviosos.por.mes)=c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio",
,"Agosto","Setiembre","Octubre","Noviembre","Diciembre")
rownames(dias.lluviosos.por.mes)=c("Media Dias Lluviosos Mensual",
,"SD Dias Lluviosos Mensual","IC inferior","IC superior")

boxplot(dias.lluviosos[-dim(dias.lluviosos)[1],14]~dias.lluviosos[-dim(dias.lluviosos)[1],2], col="darkgreen",
names=c("Ene","Feb","Mar","Abr","May","Jun","Jul","Ago","Set","Oct","Nov","Dic"),varwidth=TRUE
,ylab="Nro. dias lluviosos",sub="Periodo 1960-2008",pch=19)

boxplot(lluvia.mensual[-dim(lluvia.mensual)[1],14]~lluvia.mensual[-dim(dias.lluviosos)[1],2], col="lightblue",
names=c("Ene","Feb","Mar","Abr","May","Jun","Jul","Ago","Set","Oct","Nov","Dic"),varwidth=TRUE,
ylab="mm",sub="Periodo 1960-2008",pch=19)

#####
#Gráfico de barras de días lluviosos por mes según estación y total en Colonia #
#####

par(mfrow=c(3,floor(nroest/3)+1))
lim=rep(11,12)
for(i in 1:(nroest+1))
{dias.lluviososi=dias.lluviosos[-dim(dias.lluviosos)[1],c(1,2,i+2)]
dias.lluviosos.por.mesi=matrix(rep(0,12),1,12)
for(j in 1:12)
{a=dias.lluviososi[dias.lluviososi$MES==j,]
dias.lluviosos.por.mesi[1,j]=mean(a[,3],na.rm=TRUE)
}
}
if(i>=9)
{colnames(dias.lluviosos.por.mesi)=c("En","Fe","Ma","Ab","Ma","Jun","Jul","Ag","Se","Oc","No","Di")}
etiqueta=c(' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ',' ')
rownames(dias.lluviosos.por.mesi)=c("Media Lluvia Mensual")

```

```

if(i==1 | i==5 | i==9)
{barplot(dias.lluviosos.por.mesi,col="darkblue", axes = TRUE, ylab ="Numero de dias"
        ,xlab=etiqueta[i],ylim=c(0,lim[i]))
}
if(i!=1 & i!=5 & i!=9)
{barplot(dias.lluviosos.por.mesi,col="darkblue", axes = TRUE,ylim=c(0,lim[i]) ,xlab=etiqueta[i])}
title(main=list(estacionesColoniaGr[i],cex=1.5,col="red", font=2))
}

#####
#Gráficos qq plot para ver si hay diferencias entre estaciones de DNM respecto a la meteorológica y INIA #
#####

par(mfrow=c(2,floor((nroest-2)/2)+1))
for(i in 1:(nroest-2))
{
  qqplot(dias.lluviosos[,i+2],dias.lluviosos[,12],xlim=c(0,20),ylim=c(0,20),xlab=estacionesColoniaGr[i]
        ,ylab=estacionesColoniaGr[10],pch=19)
  lines(seq(0,20),seq(0,20), col = 2,cex=5, lwd=3)
}

###Comparado con la del INIA.
par(mfrow=c(2,floor((nroest-1)/2)))
for(i in 1:(nroest-1))
{
  qqplot(dias.lluviosos[,i+2],dias.lluviosos[,13],xlim=c(0,20),ylim=c(0,20),xlab=estacionesColoniaGr[i]
        ,ylab=estacionesColoniaGr[11],pch=19)
  lines(seq(0,20),seq(0,20), col = 2,cex=5, lwd=3)
}

#####
#Gráfico de barras de lluvia mensual según estación y total en Colonia #
#####

#Hacemos unos barplot con la lluvia.mensual
par(mfrow=c(3,floor(nroest/3)+1))
for(i in 1:(nroest+1))
{lluvia.mensuali=lluvia.mensual[-dim(lluvia.mensual)[1],c(1,2,i+2)]
  lluvia.por.mesi=matrix(rep(0,12),1,12)
  for(j in 1:12)
  {a=lluvia.mensuali[lluvia.mensuali$MES==j,]
    lluvia.por.mesi[1,j]=mean(a[,3],na.rm=TRUE)
  }
  if(i>=9)
  {colnames(lluvia.por.mesi)=c("En","Fe","Ma","Ab","Ma","Jun","Jul","Ag","Se","Oc","No","Di")}
  etiqueta=c('','','','','','','','','Mes','Mes','Mes','Mes')
  rownames(lluvia.por.mesi)=c("Media Dias Lluviosos Mensual")
  if(i==1 | i==5 | i==9)
  {barplot(lluvia.por.mesi,col="darkblue", axes = TRUE,ylab ="mm",xlab=etiqueta[i],ylim=c(0,140))
  }
  if(i!=1 & i!=5 & i!=9)
  {barplot(lluvia.por.mesi,col="darkblue", axes = TRUE,ylim=c(0,140) ,xlab=etiqueta[i])}
  title(main=list(estacionesColoniaGr[i],cex=1.5,col="red", font=2))
  }

#####
#Gráfico qq plot entre lluvia acumulada en las estaciones del año en Colonia #
#####

panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE,breaks=c(0,50,100,150,200,250,300,250,300,350,400,450,500,550,600))
  breaks <- h$breaks;
  nB <- length(breaks)
}

```

```

y <- h$counts; y <- y/max(y)
rect(breaks[-nB], 0, breaks[-1], y, col="cyan")
#para dibujar los histogramas
}
pairs(lluvia.mensual.res1, panel=panel.smooth, cex = 1.5,ylim=c(0,max(lluvia.mensual.res1))
,xlim=c(0,max(lluvia.mensual.res1)),labels=c("Verano","Otoño","Invierno","Primavera"),pch = 19, bg="lightblue",
diag.panel=panel.hist, cex.labels = 1, font.labels=1)

par(oma=c(1,1,1,1),new=T,font=2,cex=0.5)

#####
#Lluvia mensual #
#####

a=unique(data[,c(1,2)])
lluvia.mensualaux=cbind(a,matrix(rep(0,dim(a)[1]*(nroest+1)),length(rep(0,dim(a)[1])),(nroest+1)))
lluvia.mensualaux=rbind(lluvia.mensualaux,rep(0,(nroest+3)))
colnames(lluvia.mensualaux)[c(3:(nroest+2))]=estacionesColonia
colnames(lluvia.mensualaux)[(nroest+3)]="Total"
dias.lluviososaux=lluvia.mensualaux
for(i in 1:(dim(lluvia.mensualaux)[1]-1))
  {data1=data[data[,1]==lluvia.mensualaux[i,1] & data[,2]==lluvia.mensualaux[i,2],]
  for(j in 1:nroest)
    {if (length(data1[data1[,3+2*j]>=0,3+2*j])!=dim(data1)[1])
      {lluvia.mensualaux[i,2+j]==-99.9
      dias.lluviososaux[i,2+j]==-99.9}
    if (length(data1[data1[,3+2*j]>=0,3+2*j])==dim(data1)[1])
      {lluvia.mensualaux[i,2+j]=sum(data1[data1[,3+2*j]>0,3+2*j])
      dias.lluviososaux[i,2+j]=length(data1[data1[,3+2*j]>0,3+2*j])}
    }}

#####
# Correlación entre cantidad lluvia mensual y días lluviosos #
#####

correlaciones=rep(0,nroest)
for(i in 1:nroest)
  {lluvia.mensualauxi=lluvia.mensualaux[lluvia.mensualaux[,i+2]==-99.9,]
  lluvia.mensualauxi=lluvia.mensualauxi[-dim(lluvia.mensualauxi)[1],]
  dias.lluviososauxi=dias.lluviososaux[dias.lluviososaux[,i+2]==-99.9,]
  dias.lluviososauxi=dias.lluviososauxi[-dim(dias.lluviososauxi)[1],]
  correlaciones[i]=cor(lluvia.mensualauxi[i+2],dias.lluviososauxi[i+2],)
  }
correlaciones=t(cbind(correlaciones,rep(0,nroest),rep(0,nroest)))
colnames(correlaciones)=estacionesColonia
rownames(correlaciones)=c("correlaciones lluvia mensual y días lluviosos por estación","IC sup","IC inf")
#transf fisher: (ver libro capitulo 8.2.3)
trans.fish=0.5*log((1+correlaciones[1,])/(1-correlaciones[1,]))
icsuptf=trans.fish+1.96*(1/(dim(lluvia.mensualauxi)[1]-3))
icinf=trans.fish-1.96*(1/(dim(lluvia.mensualauxi)[1]-3))
#deshago la trans
correlaciones[2,]=(exp(2*icsuptf)-1)/(exp(2*icsuptf)+1)
correlaciones[3,]=(exp(2*icinf)-1)/(exp(2*icinf)+1)

#####
#Influencia del mes en la correlación entre cantidad lluvia mensual y días lluviosos #
#####

correlaciones=matrix(rep(0,12*nroest),nroest,12)
for(j in 1:12)
  {lluvia.mensualauxi=lluvia.mensualaux[lluvia.mensualaux$MES==j,]
  dias.lluviososauxi=dias.lluviososaux[dias.lluviososaux$MES==j,]
  for(i in 1:nroest)
    {lluvia.mensualauxi=lluvia.mensualauxi[lluvia.mensualauxi[,i+2]==-99.9,]
    lluvia.mensualauxi=lluvia.mensualauxi[-dim(lluvia.mensualauxi)[1],]
    dias.lluviososauxi=dias.lluviososauxi[dias.lluviososauxi[,i+2]==-99.9,]
  }
}

```

```

dias.lluviososauxi=dias.lluviososauxi[-dim(dias.lluviososauxi)[1],]
  correlaciones[i,j]=cor(lluvia.mensualaux[,i+2],dias.lluviososauxi[,i+2])
}}
colnames(correlaciones)=c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
,"Setiembre","Octubre","Noviembre","Diciembre")
rownames(correlaciones)=estacionesColonia
correlaciones

#####
#Correlacion mensual entre estaciones #
#####

matcorrmen=diag(1,nroest,nroest)
for(i in 1:nroest-1)
{for (j in (i+1):nroest)
  {lluvia.mensualij=lluvia.mensualaux[lluvia.mensualaux[,i+2]!=-99.9 & lluvia.mensualaux[,j+2]!=-99.9,]
lluvia.mensualij=lluvia.mensualij[-dim(lluvia.mensualij)[1],]
  matcorrmen[i,j]=cor(lluvia.mensualij[i+2],lluvia.mensualij[j+2],)
matcorrmen[j,i]=matcorrmen[i,j]
}
}
colnames(matcorrmen)=estacionesColonia
rownames(matcorrmen)=estacionesColonia

#####
#Correlacion diaria, y concordancia entre estaciones #
#####

matcorr=diag(1,nroest,nroest) #matriz de correlación
matconcor=diag(1,nroest,nroest)
for(i in 1:nroest-1)
{for (j in (i+1):nroest)
{data1=cbind(data[,c(i*2+3-1,i*2+3,j*2+3-1,j*2+3)])
data1=data1[data1[2]>=0 & data1[4]>=0,] #me quedo solo con los datos mayores a cero
data1=cbind(data1,rep(0,dim(data1)[1]),rep(0,dim(data1)[1]),rep(0,dim(data1)[1]))
data1[data1[2]>0,5]=1
data1[data1[4]>0,6]=1 #le hago como si llovio o si no llovio para los 2
data1[data1[5]==data1[6],7]=1
matconcor[i,j]=sum(data1[7])/dim(data1)[1]
matconcor[j,i]=matconcor[i,j]
matcorr[i,j]=cor(data1[2],data1[4])
matcorr[j,i]=matcorr[i,j]
}
}
colnames(matconcor)=estacionesColonia
rownames(matconcor)=colnames(matconcor)
colnames(matcorr)=estacionesColonia
rownames(matcorr)=colnames(matcorr)

#####Se crea una matriz que contenga estos 3 elementos para hacer grafico#####
DistCorrCon=t(c(0,0,0,0,0,0))
for(i in 1:(nroest-1))
{for (j in (i+1):nroest)
  {DistCorrCon=rbind(DistCorrCon,c(distCol[i,j],matcorr[i,j],matconcor[i,j],matcorrmen[i,j],i,j))
  }}
DistCorrCon=DistCorrCon[-1,]
colnames(DistCorrCon)=c("distCol","matcorr","matconcor","matcorrmen","i","j")
DistCorrCon=DistCorrCon[order(DistCorrCon[,5],DistCorrCon[,6]),] ###ordena por distancia a la matriz
distancias=cbind(DistCorrCon[,1],rep(1,length(DistCorrCon[,1])))

#####
#Gráficos #
#####

par(mfrow=c(1,3),mar=c(4,4,7,4))
plot(DistCorrCon[,1],DistCorrCon[,2],xlab="Distancia entre estaciones en km"
,ylab="Correlación diaria entre estaciones" ,main="Distancia vs correlación diaria.",pch=19)
plot(DistCorrCon[,1],DistCorrCon[,3],xlab="Distancia entre estaciones en km"

```

```

,ylab="Concordancia diaria entre estaciones",main="Distancia vs concordancia diaria.",pch=19)
plot(DistCorrCon[,1],DistCorrCon[,4],xlab="Distancia entre estaciones en km"
,ylab="Correlación mensual estaciones",main="Distancia vs correlación mensual.",pch=19)
mtext(side=3, line=-1.3, cex=1.3, outer=T,"Gráfico de dispersión
relacionando distancia con distintos valores en Colonia.")

cor(DistCorrCon[,1],DistCorrCon[,3]) #correlación entre distancia y concordancia
cor(DistCorrCon[,1],DistCorrCon[,2]) #correlación entre distancia y correlación diaria
cor(DistCorrCon[,1],DistCorrCon[,4]) #correlación entre distancia y correlación mensual

DistCorrConAux=rbind(DistCorrCon,DistCorrCon[,c(1,2,3,4,6,5)])
DistCorrConAux=DistCorrConAux[,-6]#sacamos la j

#####
#Saco la identificación, y convierto a binario los datos #
#####

data=read.table("BaseDatosColonial1est.txt",header=TRUE)
data=data[,-1]
databin=data[,seq(5,25,2)] #tiene todas las precipitaciones
for(i in 1:11)
{databin[databin[,i]>0,i]=1}#aca ya quedo binario

#####
#Estimación de densidades #
#####

data=data[,-1]
data=data[,seq(5,25,2)]
par(mfrow=c(1,2))
for (i in 1:nroest)
{
x=data[data[,i]>5,i]
a=density(x,from=5)
if(i==1){plot(a,ylab="Densidad",main="densidades estimadas dado que llovió
mas de 5mm para estaciones de colonia",xlab="Precipitación diaria en mm",ylim=c(0,0.05),col=i)}
if(i>1){lines(a,col=i)}
}
for (i in 1:nroest)
{
x=data[data[,i]>10,i]
a=density(x,from=10)
if(i==1){plot(a,ylab="Densidad",main="densidades estimadas dado que llovió
mas de 10mm para estaciones de colonia",xlab="Precipitación diaria en mm",ylim=c(0,0.05),col=i)}
if(i>1){lines(a,col=i)}
}

par(mfrow=c(4,3))
for (i in 1:nroest)
{
x=data[data[,i]>5,i]#condiciono a que llovio mas de 5 mm
a=density(x,from=5.1)
plot(a,ylab="Densidad",main=estacionesColonias[i],xlab="Precipitación diaria en mm",ylim=c(0,.05))
}

#Lluvia suave
estacionesColonias=as.character(c("Estación 2480C","Estación 2657A","Estación 2774 Met","Estación La Estanz"))
vec=c(3,6,10,11)
par(mfrow=c(1,4))
for (i in 1:4)
{
x=data[data[,vec[i]]>=0.1 & data[,vec[i]]<=5,vec[i]] #
a=density(x,from=0.01,to=5.0)
plot(a,ylab="Densidad",main=estacionesColonias[i],xlab="Precipitación diaria(mm)",ylim=c(0,0.7))
}

```

```

#Lluvia moderada
estacionesColonia=as.character(c("Estación 2480C", "Estación 2657A", "Estación 2774 Met", "Estación La Estanz"))
vec=c(3,6,10,11)
par(mfrow=c(1,4))
for (i in 1:4)
{
x=data[data[,vec[i]]>5.01 & data[,vec[i]]<=15,vec[i]] #
a=density(x,from=5.01,to=15.0)
plot(a,ylab="Densidad",main=estacionesColonia[i],xlab="Precipitación diaria(mm)",ylim=c(0,0.2))
}

#lluvia suave
par(mfrow=c(4,3))
ylab1=c("Densidad","", "", "Densidad","", "", "Densidad","", "", "Densidad","", "")
xlab1=c("", "", "", "", "", "", "", "", "Precipitación diaria en mm", "Precipitación diaria en mm", "Precipitación diaria en mm")
for (i in 1:nroest)
{
x=data[data[,i]>=0.1 & data[,i]<=5,i] #condiciono a que llovio suave
a=density(x,from=0.01,to=5.0)
plot(a,ylab=ylab1[i],main=estacionesColoniaGr[i],xlab=xlab1[i],ylim=c(0,.5),cex.lab=1.8, cex.axis=1.8,cex.main=1.8)
}

#lluvia severa
par(mfrow=c(4,3))
ylab1=c("Densidad","", "", "Densidad","", "", "Densidad","", "", "Densidad","", "")
xlab1=c("", "", "", "", "", "", "", "", "Precipitación diaria en mm", "Precipitación diaria en mm", "Precipitación diaria en mm")
for (i in 1:nroest)
{
x=data[data[,i]>=80.1,i] #condiciono a que llovio suave
a=density(x,from=80.01,to=250.0)
plot(a,ylab=ylab1[i],main=estacionesColoniaGr[i],xlab=xlab1[i],ylim=c(0,.04),cex.lab=1.8, cex.axis=1.8,cex.main=1.8)
}

#####
#Cantidad de días lluviosos y diferencia con el que tiene mas días lluviosos #
#####
cant=rep(0,nroest)
for (i in 1:nroest)
{cant[i]=sum(databin[databin[,i]==1,i])
}
names(cant)=estacionesColoniaGr[1:11]

(abso=abs(cant-max(cant)))

mean(abso[1:9]) #1546 dias mas
mean(abso[1:9])/365.25 #4.234 años
sd(abso[1:9])/365.25

#####
#Por década #
#####
databin=databin[,1:nroest]
databin=cbind(data[,c(2,3,4)],databin)

#Número de días lluviosos
cortes=seq(1960,2010,10)
cant=matrix(rep(0,5*nroest),5,nroest)
colnames(cant)=estacionesColoniaGr[1:11]
rownames(cant)=c("[1960,1970)", "[1970,1980)", "[1980,1990)", "[1990,2000)", "[2000,2008]")
for(j in 1:5)
{databin1=subset(databin,databin$A.0>=cortes[j] & databin$A.0<cortes[j+1])
for (i in 1:nroest)
{cant[j,i]=sum(databin1[databin1[,i+3]==1,i+3])
}
}

```

```

}

#Cantidad de datos faltantes
cortes=seq(1960,2010,10)
cant2=matrix(rep(0,5*nroest),5,nroest)
colnames(cant2)=estacionesColoniaGr[1:11]
rownames(cant2)=c("[1960,1970)","[1970,1980)","[1980,1990)","[1990,2000)","[2000,2008]")
for(j in 1:5)
{databin1=subset(databin,databin$A.0>=cortes[j] & databin$A.0<cortes[j+1])
for (i in 1:nroest)
{cant2[j,i]=length(databin1[databin1[,i+3]<0,i+3])
}
}
datosfalt=cant2

#Diferencia con la estación que llovió mas
diference=cant
for(i in 1:5)
{diference[i,]=abs(diference[i,]-max(diference[i,]))}

diference=t(diference)
mean(rowSums(diference)[1:9]) #1674.4
mean(rowSums(diference)[1:9])-mean(abso[1:9])#son 128 dias mas con lo de comparar por décadas.
diference=cbind(diference,rowSums(diference))

#####
#Número de días lluviosos en cada escala de medición #
#####
escala=matrix(rep(0,6*nroest),nroest,6)
rownames(escala)=estacionesColoniaGr[1:11]
colnames(escala)=rep(c("suave","moderada","moderada a fuerte","fuerte","severa","Faltante"),1)
for (i in 1:nroest)
{escala[i,1]=length(dataimp[dataimp[,i]>0 & dataimp[,i]<=5,i])
escala[i,2]=length(dataimp[dataimp[,i]>5 & dataimp[,i]<=15,i])
escala[i,3]=length(dataimp[dataimp[,i]>15 & dataimp[,i]<=50,i])
escala[i,4]=length(dataimp[dataimp[,i]>50 & dataimp[,i]<=100,i])
escala[i,5]=length(dataimp[dataimp[,i]>100,i])
escala[i,6]=length(dataimp[dataimp[,i]<0,i])
}
xtable(escala)
prom=rep(0,11)
for (i in 1:nroest)
{prom[i]=mean(dataimp[dataimp[,i]>=4,i])}

prom

```

## B.5. Modelado para medir la ocurrencia de lluvia.

### B.5.1. Prueba de independencia.

```

rm(list=ls())
library(xtable)

data=read.table("BaseDatosColonia11est.txt",header=TRUE)
data=data[,-1]
databin=data[,seq(5,25,2)] #tiene todas las precipitaciones
for(i in 1:11)
{databin[databin[,i]>0,i]=1}#aca ya quedo binario
#####
##Son los datos iid##
#####
head(databin)
databinest=databin[,10]

p_hat=mean(databinest[databinest>=0])

```

```

n_cambios=0
for (i in 1:(length(databinest)-1))
{if (databinest[i]>=0 & databinest[i+1]>=0){n_cambios=n_cambios+1}}
IndependenciaMatriz=matrix(c(n_cambios*(1-p_hat)*(1-p_hat),n_cambios*p_hat*(1-p_hat)
,n_cambios*p_hat*(1-p_hat),n_cambios*p_hat*p_hat),2,2,
dimnames = list(c("Ayer Seco", "Ayer Llovio"),c("Hoy Seco", "Hoy Llueve")))

#####
#Matrices de transición de orden 1 #
#####
nroest=11
Ptranstotal=matrix(rep(0,4*nroest),nroest,4)
PtransTOTAL=matrix(rep(0,4*nroest),nroest,4)
Ptran=matrix(c(0,0,0,0),2,2) #matriz de transición P para cada estación las filas corresponden al día anterior
Ptot=matrix(c(0,0,0,0),2,2) #esta es con los totales las filas corresponden al día anterior
colnames(Ptot)=c(0,1)
rownames(Ptot)=c(0,1)
colnames(Ptran)=c(0,1)
rownames(Ptran)=c(0,1)
pluv=databin[,10]
for (j in 1:(length(pluv)-1))
{if(pluv[j]>=0 & pluv[j+1]>=0)
{if(pluv[j]>0 & pluv[j+1]>0){Ptot[2,2]=Ptot[2,2]+1}
if(pluv[j]==0 & pluv[j+1]==0) {Ptot[1,1]=Ptot[1,1]+1}
if(pluv[j]>0 & pluv[j+1]==0) {Ptot[2,1]=Ptot[2,1]+1}
if(pluv[j]==0 & pluv[j+1]>0) {Ptot[1,2]=Ptot[1,2]+1}
}
}
Ptran[1,1]=Ptot[1,1]/(Ptot[1,1]+Ptot[1,2])
Ptran[1,2]=Ptot[1,2]/(Ptot[1,1]+Ptot[1,2])
Ptran[2,1]=Ptot[2,1]/(Ptot[2,2]+Ptot[2,1])
Ptran[2,2]=Ptot[2,2]/(Ptot[2,2]+Ptot[2,1])
Ptranstotal=matrix(c(Ptran[1,1],Ptran[1,2],Ptran[2,1],Ptran[2,2]),2,2,
dimnames = list(c("Ayer Seco", "Ayer Llovio"),c("Hoy Seco", "Hoy Llueve")), byrow = TRUE)

PtransTOTAL=matrix(c(Ptot[1,1],Ptot[1,2],Ptot[2,1],Ptot[2,2]),2,2,
dimnames = list(c("Ayer Seco", "Ayer Llovio"),c("Hoy Seco", "Hoy Llueve")), byrow = TRUE)

xtable(PtransTOTAL,caption="matriz de transicion para cada estacion de Colonia")

PtransTOTAL ##es la de la practica
IndependenciaMatriz ##es la de la teoria

sum((PtransTOTAL-IndependenciaMatriz)^2/IndependenciaMatriz)

```

## B.5.2. Prueba de estacionariedad.

```

rm(list=ls())
data=read.table("BaseDatosColonial1est.txt",header=TRUE)
head(data)
data=data[,c(2,3,4,24)] #se elige la estación deseada
Partv=matrix(0,61,4) #hace un partición para ver las fechas que agarra
colnames(Partv)=c("MES_I", "DIA_I", "MES_F", "DIA_F")
datav=data[data[,1]==1960,]

for (i in 1:61)
{Partv[i,c(1,2)]=as.numeric(datav[(i-1)*6+1,c(2,3)])
Partv[i,c(3,4)]=as.numeric(datav[i*6,c(2,3)])}

centrosv=datav[seq(3,366,6),c(2,3)]

PartNv=Partv
PartNv[10,4]=28

```

```

##matriz de transición
probabilidadesest=matrix(0,61,4)
colnames(probabilidadesest)=c("P00", "P01", "P10", "P11")

for (i in 1960:2008)
{if (i/4==floor(i/4))
{for (j in 1:dim(centrosv)[1])
{minimo=which(data$A.0==i & data$MES==Partv[j,1] & data$DIA==Partv[j,2])
maximo=which(data$A.0==i & data$MES==Partv[j,3] & data$DIA==Partv[j,4])
subdata=data[seq(minimo,maximo),4]
{if (min(subdata)>=0)
{for (t in 1:(length(subdata)-1))
{if(subdata[t]>0 & subdata[t+1]>0){probabilidadesest[j,4]=probabilidadesest[j,4]+1}
if(subdata[t]==0 & subdata[t+1]==0) {probabilidadesest[j,1]=probabilidadesest[j,1]+1}
if(subdata[t]>0 & subdata[t+1]==0) {probabilidadesest[j,3]=probabilidadesest[j,3]+1}
if(subdata[t]==0 & subdata[t+1]>0) {probabilidadesest[j,2]=probabilidadesest[j,2]+1}
}
}
}
}
}
{if (i/4!=floor(i/4))
{for (j in 1:dim(centrosv)[1])
{minimo=which(data$A.0==i & data$MES==PartNv[j,1] & data$DIA==PartNv[j,2])
maximo=which(data$A.0==i & data$MES==PartNv[j,3] & data$DIA==PartNv[j,4])
subdata=data[seq(minimo,maximo),4]
{if (min(subdata)>=0)
{for (t in 1:(length(subdata)-1))
{if(subdata[t]>0 & subdata[t+1]>0){probabilidadesest[j,4]=probabilidadesest[j,4]+1}
if(subdata[t]==0 & subdata[t+1]==0) {probabilidadesest[j,1]=probabilidadesest[j,1]+1}
if(subdata[t]>0 & subdata[t+1]==0) {probabilidadesest[j,3]=probabilidadesest[j,3]+1}
if(subdata[t]==0 & subdata[t+1]>0) {probabilidadesest[j,2]=probabilidadesest[j,2]+1}
}
}
}
}
}
}

rowSums(probabilidadesest) #da bien
ProbGraf=cbind(probabilidadesest[,1]/(probabilidadesest[,1]+probabilidadesest[,2])
,probabilidadesest[,2]/(probabilidadesest[,1]+probabilidadesest[,2]),
probabilidadesest[,3]/(probabilidadesest[,3]+probabilidadesest[,4])
,probabilidadesest[,4]/(probabilidadesest[,3]+probabilidadesest[,4]))
time=seq(3,366,6)

ProbGraf
casos=colSums(probabilidadesest)
probST=cbind(casos[1]/(casos[1]+casos[2]),casos[2]/(casos[1]+casos[2])
,casos[3]/(casos[3]+casos[4]),casos[4]/(casos[3]+casos[4]))
cociente=cbind(probST[1]/ProbGraf[,1],probST[2]/ProbGraf[,2],probST[3]/ProbGraf[,3],probST[4]/ProbGraf[,4])
coeficientes=cociente^probabilidadesest
lambda=prod(coeficientes[,1])*prod(coeficientes[,2])*prod(coeficientes[,3])*prod(coeficientes[,4])
-2*log(lambda)

```

### B.5.3. Modelado para medir la ocurrencia de lluvia.

```

data=read.table("BaseDatosColonial11est.txt",header=TRUE)
head(data)
dataEst=data[,c(2,3,4,24)] # agarro solo la meteorológica
dataEst=dataEst[dataEst[,1]>=1960,]
año.inicial=1960
año.final=2008

```

```

#####
#Agrego t que corresponde al día en cuestión      #
#####
dataEst=cbind(dataEst,0)
colnames(dataEst)[5]="time"
for (i in (año.inicial):año.final)
{
if (i/4==floor(i/4)) {dataEst[dataEst[,1]==i,5]=seq(1,366)}
if (i/4!=floor(i/4)) {dataEst[dataEst[,1]==i,5]=c(seq(1,59),seq(61,366))}
}
a=NULL

for (i in 1:366)
{a=c(a,length(which(dataEst$time==i)))}

dataEstbin=dataEst
dataEstbin[dataEstbin$PRECIPITAC2774>0,4]=1

#####
#Función para crear la cadena de Markov no estacionaria      #
#####

cmarkovreal=function(datos1=dataEstbin,dos.dias.ant=0,dia.ant=0)
{exitos=cbind(1:366,0)
casos=cbind(1:366,0)
for (i in 1:366)
{if (i>=3 & i!=61 & i!=62 & i!=60) #dias comunes
{largoNum=0
largoDen=0
for (j in año.inicial:año.final)
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}
}
}
exitos[i,2]=largoNum
casos[i,2]=largoDen
}

if (i==1) #1er dia del año
{largoNum=0
largoDen=0
for (j in año.inicial:año.final)
{if(all(datos1[datos1$A.0==j & (datos1$time==365 | datos1$time==366 | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==365 | datos1$time==366 | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}
}
}
exitos[i,2]=largoNum
casos[i,2]=largoDen
}

if (i==2) #2do dia del año
{largoNum=0
largoDen=0
for (j in año.inicial:año.final)
{if(all(datos1[datos1$A.0==j & (datos1$time==366 | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==366 | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}
}
}
exitos[i,2]=largoNum
casos[i,2]=largoDen
}
}

```

```

if (i==60) #29 febrero
{largoNum=0
largoDen=0
for (j in año.inicial:año.final)
{if(j/4==floor(j/4))
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}}
}
 exitos[i,2]=largoNum
 casos[i,2]=largoDen
}
if (i==61) #1er dia de marzo
{largoNum=0
largoDen=0
for (j in año.inicial:año.final)
{if(j/4==floor(j/4))
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}}
if(j/4!=floor(j/4))
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-3) | datos1$time==(i-2) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-3) | datos1$time==(i-2) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}}
}}
 exitos[i,2]=largoNum
 casos[i,2]=largoDen
}
if (i==62) #2do dia de marzo
{largoNum=0
largoDen=0
for (j in año.inicial:año.final)
{if(j/4==floor(j/4))
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}}
if(j/4!=floor(j/4))
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-3) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-3) | datos1$time==(i-1) | datos1$time==i),4]==c(dos.dias.ant,dia.ant,0)))
{largoDen=1+largoDen}}
}}
 exitos[i,2]=largoNum
 casos[i,2]=largoDen
}
}
return(cbind(exitos,casos[,2]))
}
#####
#Coloco a cada día de lluvia o no lluvia con su estado real #
#####

dataEstbin=cbind(dataEstbin,0)
colnames(dataEstbin)[6]="Grupo"
for (i in 3:dim(dataEstbin)[1])
{if (dataEstbin[i-2,4]==0 & dataEstbin[i-1,4]==0){dataEstbin[i,6]=1}
if (dataEstbin[i-2,4]==1 & dataEstbin[i-1,4]==0){dataEstbin[i,6]=2}
if (dataEstbin[i-2,4]==0 & dataEstbin[i-1,4]==1){dataEstbin[i,6]=3}

```

```

if (dataEstbin[i-2,4]==1 & dataEstbin[i-1,4]==1){dataEstbin[i,6]=4}
}

##

mc=cmarkovreal(dos.dias.ant=0,dia.ant=0)
mc1=cmarkovreal(dos.dias.ant=1,dia.ant=0)
mc2=cmarkovreal(dos.dias.ant=0,dia.ant=1)
mc3=cmarkovreal(dos.dias.ant=1,dia.ant=1)

#####
#Calculo de verosimilitud #
#####
mc0=mc#dia, exitos y casos
probest00=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest00[i,1]
probest00[i,2]=(sum(mc00[(t-2):(t+3),3])-sum(mc00[(t-2):(t+3),2]))/sum(mc00[(t-2):(t+3),3])
probest00[i,3]=(sum(mc00[(t-2):(t+3),3])-sum(mc00[(t-2):(t+3),2]))
probest00[i,4]=sum(mc00[(t-2):(t+3),2])/sum(mc00[(t-2):(t+3),3])
probest00[i,5]=sum(mc00[(t-2):(t+3),2])
}
colnames(probest00)=c("t","p000","n000","p001","n001")
mc01=mc2
probest01=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest01[i,1]
probest01[i,2]=(sum(mc01[(t-2):(t+3),3])-sum(mc01[(t-2):(t+3),2]))/sum(mc01[(t-2):(t+3),3])
probest01[i,3]=(sum(mc01[(t-2):(t+3),3])-sum(mc01[(t-2):(t+3),2]))
probest01[i,4]=sum(mc01[(t-2):(t+3),2])/sum(mc01[(t-2):(t+3),3])
probest01[i,5]=sum(mc01[(t-2):(t+3),2])
}
colnames(probest01)=c("t","p010","n010","p011","n011")

mc10=mc1
probest10=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest10[i,1]
probest10[i,2]=(sum(mc10[(t-2):(t+3),3])-sum(mc10[(t-2):(t+3),2]))/sum(mc10[(t-2):(t+3),3])
probest10[i,3]=(sum(mc10[(t-2):(t+3),3])-sum(mc10[(t-2):(t+3),2]))
probest10[i,4]=sum(mc10[(t-2):(t+3),2])/sum(mc10[(t-2):(t+3),3])
probest10[i,5]=sum(mc10[(t-2):(t+3),2])
}
colnames(probest10)=c("t","p100","n100","p101","n101")

mc11=mc3
probest11=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest11[i,1]
probest11[i,2]=(sum(mc11[(t-2):(t+3),3])-sum(mc11[(t-2):(t+3),2]))/sum(mc11[(t-2):(t+3),3])
probest11[i,3]=(sum(mc11[(t-2):(t+3),3])-sum(mc11[(t-2):(t+3),2]))
probest11[i,4]=sum(mc11[(t-2):(t+3),2])/sum(mc11[(t-2):(t+3),3])
probest11[i,5]=sum(mc11[(t-2):(t+3),2])
}
colnames(probest11)=c("t","p110","n110","p111","n111")

vero2=sum(log(probest11[,2])*probest11[,3]+log(probest11[,4])*probest11[,5])+
sum(log(probest10[,2])*probest10[,3]+log(probest10[,4])*probest10[,5])+
sum(log(probest00[,2])*probest00[,3]+log(probest00[,4])*probest00[,5])+
sum(log(probest01[,2])*probest01[,3]+log(probest01[,4])*probest01[,5])

#####
#Estado 00 #
#####

#Grupo cada 6 días
probest=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))

```

```

{t=probest[i,1]
probest[i,2]=sum(mc[(t-2):(t+3),2])/sum(mc[(t-2):(t+3),3])
probest[i,3]=sum(mc[(t-2):(t+3),3])
}

#ghit=ahio+[ahik*sin(k*2*pi*t/366)+bihk*sin(k*2*pi*t/366)]
#phi()=exp(ghit)/(1+exp(ghit)) # por lo tanto ghit=log(p/(1-p))

probest=cbind(probest,log(probest[,2]/(1-probest[,2])),2*pi*probest[,1]/366)
colnames(probest)=c("time","p00","nhi","ghit","rt")

#Acá pasamos de la p gorro a la g gorro.
probest=as.data.frame(probest)

Aic=matrix(0,nrow=6,ncol=4)
colnames(Aic)=c('00','10','01','11')
rownames(Aic)=0:5
Loglik=Aic

probest2=cbind(probest,sin(probest$rt),cos(probest$rt),sin(2*probest$rt),cos(2*probest$rt)
,sin(3*probest$rt),cos(3*probest$rt),
sin(4*probest$rt),cos(4*probest$rt),sin(5*probest$rt),cos(5*probest$rt))

# Con m=0
#En este ejemplo como es p00 la hi valen 0
regresion01 <- lm(ghit ~ 1,weights=nhi,data=probest2)
pr0=predict (regresion01, interval="confidence")

# Con m=1, aqui k vale solo 1
regresion11 <- lm(ghit ~sin(probest$rt) + cos(probest$rt),weights=nhi,data=probest2)
pr1=predict (regresion11, interval="confidence")

# Con m=2, aca k vale 1 o 2
regresion21 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
,weights=nhi,data=probest2)
pr2=predict (regresion21, interval="confidence")

# Con m=3, aca k vale 1 o 2
regresion31 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt),weights=nhi,data=probest2)
pr3=predict (regresion31, interval="confidence")

# Con m=4,
regresion41 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt)+ sin(4 * probest$rt) + cos(4 * probest$rt),weights=nhi,data=probest2)
pr4=predict (regresion41, interval="confidence")

# Con m=5,
regresion51 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt)+ sin(4 * probest$rt) + cos(4 * probest$rt) + sin(5 * probest$rt)
+ cos(5 * probest$rt),weights=nhi,data=probest2)
pr5=predict (regresion51, interval="confidence")

#gráfico
par(mfrow=c(1,2))

p00est1=exp(pr0)/(1+exp(pr0))
#regresión 0
plot(probest[,1],probest[,2],pch=19,ylim=c(0,1))
lines(probest[,1],p00est1[,1])
lines(probest[,1],p00est1[,2],col="red",lty=2)
lines(probest[,1],p00est1[,3],col="red",lty=2)

p00est2=exp(pr1)/(1+exp(pr1))
#regresión 1
plot(probest[,1],probest[,2],pch=19,ylim=c(0,1))
lines(probest[,1],p00est2[,1])
lines(probest[,1],p00est2[,2],col="red",lty=2)

```

```

lines(probest[,1],p00est2[,3],col="red",lty=2)

p00est3=exp(pr2)/(1+exp(pr2))
#regresión 2
plot(probest[,1],probest[,2],pch=19,ylim=c(0,1))
lines(probest[,1],p00est3[,1])
lines(probest[,1],p00est3[,2],col="red",lty=2)
lines(probest[,1],p00est3[,3],col="red",lty=2)

p00est4=exp(pr3)/(1+exp(pr3))
#regresión 3
plot(probest[,1],probest[,2],pch=19,ylim=c(0,1))
lines(probest[,1],p00est4[,1])
lines(probest[,1],p00est4[,2],col="red",lty=2)
lines(probest[,1],p00est4[,3],col="red",lty=2)
probest00=probest ##aca le digo que es el optimo

p00est5=exp(pr4)/(1+exp(pr4))
#regresión 4
plot(probest[,1],probest[,2],pch=19,ylim=c(0,1))
lines(probest[,1],p00est5[,1])
lines(probest[,1],p00est5[,2],col="red",lty=2)
lines(probest[,1],p00est5[,3],col="red",lty=2)

p00est6=exp(pr5)/(1+exp(pr5))
#regresión 5
plot(probest[,1],probest[,2],pch=19,ylim=c(0,1))
lines(probest[,1],p00est6[,1])
lines(probest[,1],p00est6[,2],col="red",lty=2)
lines(probest[,1],p00est6[,3],col="red",lty=2)

probest1=cbind(probest[,2]*probest[,3],probest[,3]*(1-probest[,2]),p00est1[,1],p00est2[,1],p00est3[,1],
,p00est4[,1],p00est5[,1],p00est6[,1])
colnames(probest1)=c("n1","n0","0","1","2","3","4","5")
loglik=c(
sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]),
sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]),
sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]),
sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]),
sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]),
sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))

AIC=c(
-2*(sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]))+1,
-2*(sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]))+3,
-2*(sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]))+5,
-2*(sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]))+7,
-2*(sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]))+9,
-2*(sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))+11
)

Aic[,1]=AIC
Bic[,1]=BIC
Loglik[,1]=loglik

#####
#Ahora queda para los demás modelos el 011 111 101 #
#####

#####
#Estado 10 #
#####

```

```

#Agrupo cada 6 días
probest=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest[i,1]
probest[i,2]=sum(mc1[(t-2):(t+3),2])/sum(mc1[(t-2):(t+3),3])
probest[i,3]=sum(mc1[(t-2):(t+3),3])
}

probest=cbind(probest,log(probest[,2]/(1-probest[,2])),2*pi*probest[,1]/366)
colnames(probest)=c("time","p10","nhi","ghit","rt")
probest=as.data.frame(probest)

##Intervalo de confianza
probest2=cbind(probest,sin(probest$rt),cos(probest$rt),sin(2*probest$rt),cos(2*probest$rt),sin(3*probest$rt),
cos(3*probest$rt),sin(4*probest$rt),cos(4*probest$rt),sin(5*probest$rt),cos(5*probest$rt))

probest10=probest

regresion02 <- lm(ghit ~ 1,weights=nhi,data=probest2)
pr0=predict (regresion02, interval="confidence")

#Con m=1, aquí k vale solo 1
regresion12 <- lm(ghit ~sin(probest$rt) + cos(probest$rt),weights=nhi,data=probest2)
pr1=predict (regresion12, interval="confidence")

# Con m=2, acá k vale 1 o 2
regresion22 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
,weights=nhi,data=probest2)
pr2=predict (regresion22, interval="confidence")

# Con m=3, acá k vale 1, 2 o 3
regresion32 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt),weights=nhi,data=probest2)
pr3=predict (regresion32, interval="confidence")

# Con m=4,
regresion42 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt)+ sin(4 * probest$rt) + cos(4 * probest$rt),weights=nhi,data=probest2)
pr4=predict (regresion42, interval="confidence")

# Con m=5,
regresion52 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt)+ sin(4 * probest$rt) + cos(4 * probest$rt) + sin(5 * probest$rt)
+ cos(5 * probest$rt),weights=nhi,data=probest2)
pr5=predict (regresion52, interval="confidence")

p10est1=exp(pr0)/(1+exp(pr0))
p10est2=exp(pr1)/(1+exp(pr1))
p10est3=exp(pr2)/(1+exp(pr2))
p10est4=exp(pr3)/(1+exp(pr3))
p10est5=exp(pr4)/(1+exp(pr4))
p10est6=exp(pr5)/(1+exp(pr5))

probest1=cbind(probest[,2]*probest[,3],probest[,3]*(1-probest[,2]),p10est1[,1],p10est2[,1],p10est3[,1],
p10est4[,1],p10est5[,1],p10est6[,1])
colnames(probest1)=c("n1","n0","0","1","2","3","4","5")
loglik=c(
sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]),
sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]),
sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]),
sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]),
sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]),
sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))

```

```

AIC=c(
-2*(sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]))+1,
-2*(sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]))+3,
-2*(sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]))+5,
-2*(sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]))+7,
-2*(sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]))+9,
-2*(sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))+11
)
Aic[,2]=AIC
Loglik[,2]=loglik

#####
#Estado 01 #
#####

#Agrupo cada 6 días
probest=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest[i,1]
probest[i,2]=sum(mc2[(t-2):(t+3),2])/sum(mc2[(t-2):(t+3),3])
probest[i,3]=sum(mc2[(t-2):(t+3),3])
}

probest=cbind(probest,log(probest[,2]/(1-probest[,2])),2*pi*probest[,1]/366)
colnames(probest)=c("time","p01","nhi","ghit","rt")
probest=as.data.frame(probest)

##Intervalo de confianza
probest2=cbind(probest,sin(probest$rt),cos(probest$rt),sin(2*probest$rt),cos(2*probest$rt),sin(3*probest$rt),
,cos(3*probest$rt),sin(4*probest$rt),cos(4*probest$rt),sin(5*probest$rt),cos(5*probest$rt))

probest01=probest

regresion03 <- lm(ghit ~ 1,weights=nhi,data=probest2)
pr0=predict (regresion03, interval="confidence")

#Con m=1, aquí k vale solo 1
regresion13 <- lm(ghit ~sin(probest$rt) + cos(probest$rt),weights=nhi,data=probest2)
pr1=predict (regresion13, interval="confidence")

# Con m=2, acá k vale 1 o 2
regresion23 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
,weights=nhi,data=probest2)
pr2=predict (regresion23, interval="confidence")

# Con m=3, acá k vale 1, 2 o 3
regresion33 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt),weights=nhi,data=probest2)
pr3=predict (regresion33, interval="confidence")

# Con m=4,
regresion43 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt) + sin(4 * probest$rt) + cos(4 * probest$rt),weights=nhi,data=probest2)
pr4=predict (regresion43, interval="confidence")

# Con m=5,
regresion53 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt)+ sin(4 * probest$rt) + cos(4 * probest$rt) + sin(5 * probest$rt)
+ cos(5 * probest$rt) ,weights=nhi,data=probest2)
pr5=predict (regresion53, interval="confidence")

p01est1=exp(pr0)/(1+exp(pr0))

p01est2=exp(pr1)/(1+exp(pr1))

p01est3=exp(pr2)/(1+exp(pr2))

```

```

p01est4=exp(pr3)/(1+exp(pr3))

p01est5=exp(pr4)/(1+exp(pr4))

p01est6=exp(pr5)/(1+exp(pr5))

probest1=cbind(probest[,2]*probest[,3],probest[,3]*(1-probest[,2]),p01est1[,1],p01est2[,1],p01est3[,1],
,p01est4[,1],p01est5[,1],p01est6[,1])
colnames(probest1)=c("n1","n0","0","1","2","3","4","5")
loglik=c(
sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]),
sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]),
sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]),
sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]),
sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]),
sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))

AIC=c(
-2*(sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]))+1,
-2*(sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]))+3,
-2*(sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]))+5,
-2*(sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]))+7,
-2*(sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]))+9,
-2*(sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))+11
)

Aic[,3]=AIC
Loglik[,3]=loglik

#####
#Estado 01 #
#####

#Grupo cada 6 días
probest=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest[i,1]
probest[i,2]=sum(mc3[(t-2):(t+3),2])/sum(mc3[(t-2):(t+3),3])
probest[i,3]=sum(mc3[(t-2):(t+3),3])
}

probest=cbind(probest,log(probest[,2]/(1-probest[,2])),2*pi*probest[,1]/366)
colnames(probest)=c("time","p11","nhi","ghit","rt")
probest=as.data.frame(probest)

## Intervalos de confianza
probest2=cbind(probest,sin(probest$rt),cos(probest$rt),sin(2*probest$rt),cos(2*probest$rt),sin(3*probest$rt),
,cos(3*probest$rt),sin(4*probest$rt),cos(4*probest$rt),sin(5*probest$rt),cos(5*probest$rt))

probest11=probest

regresion04 <- lm(ghit ~ 1,weights=nhi,data=probest2)
pr0=predict (regresion04, interval="confidence")

# Con m=1, aqui k vale solo 1
regresion14<- lm(ghit ~sin(probest$rt) + cos(probest$rt),weights=nhi,data=probest2)
pr1=predict (regresion14, interval="confidence")

# Con m=2, aca k vale 1 o 2
regresion24 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
,weights=nhi,data=probest2)
pr2=predict (regresion24, interval="confidence")

# Con m=3, aca k vale 1 o 2
regresion34 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt),weights=nhi,data=probest2)
pr3=predict (regresion34, interval="confidence")

```

```

# Con m=4,
regresion44 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt)+ sin(4 * probest$rt) + cos(4 * probest$rt),weights=nhi,data=probest2)
pr4=predict (regresion44, interval="confidence", weights = probest2$nhi)

# Con m=5,
regresion54 <- lm(ghit ~sin(probest$rt) + cos(probest$rt) + sin(2 * probest$rt) + cos(2 * probest$rt)
+ sin(3 * probest$rt) + cos(3 * probest$rt)+ sin(4 * probest$rt) + cos(4 * probest$rt) + sin(5 * probest$rt)
+ cos(5 * probest$rt),weights=nhi,data=probest2)
pr5=predict (regresion54, interval="confidence")

p11est1=exp(pr0)/(1+exp(pr0))

p11est2=exp(pr1)/(1+exp(pr1))

p11est3=exp(pr2)/(1+exp(pr2))

p11est4=exp(pr3)/(1+exp(pr3))

p11est5=exp(pr4)/(1+exp(pr4))

p11est6=exp(pr5)/(1+exp(pr5))

probest1=cbind(probest[,2]*probest[,3],probest[,3]*(1-probest[,2]),p11est1[,1],p11est2[,1],p11est3[,1],p11est4[,1]
,p11est5[,1],p11est6[,1])
colnames(probest1)=c("n1","n0","0","1","2","3","4","5")
loglik=c(
sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]),
sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]),
sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]),
sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]),
sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]),
sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))

AIC=c(
-2*(sum(log(probest1[,3])*probest1[,1]+log(1-probest1[,3])*probest1[,2]))+1,
-2*(sum(log(probest1[,4])*probest1[,1]+log(1-probest1[,4])*probest1[,2]))+3,
-2*(sum(log(probest1[,5])*probest1[,1]+log(1-probest1[,5])*probest1[,2]))+5,
-2*(sum(log(probest1[,6])*probest1[,1]+log(1-probest1[,6])*probest1[,2]))+7,
-2*(sum(log(probest1[,7])*probest1[,1]+log(1-probest1[,7])*probest1[,2]))+9,
-2*(sum(log(probest1[,8])*probest1[,1]+log(1-probest1[,8])*probest1[,2]))+11
)

Aic[,4]=AIC
Loglik[,4]=loglik

##Aca hago el gráfico de intervalos de confianza en cadenas de Markov de orden 2
par(mfrow=c(2,2))
#00
n=1.5*(sum(probest00[,3])/max(probest00[,3]))
plot(probest00[,1],probest00[,2],pch=19,main="00",cex=n*(probest00[,3]/sum(probest00[,3]))
,ylim=c(0,1),xlab="Dias",ylab=expression(p[00](t)))
lines(probest00[,1],p00est4[,1])
lines(probest00[,1],p00est4[,2],col="red",lty=2)
lines(probest00[,1],p00est4[,3],col="red",lty=2)

#01
n=1.5*(sum(probest01[,3])/max(probest01[,3]))
plot(probest01[,1],probest01[,2],pch=19,cex=n*(probest01[,3]/sum(probest01[,3])),main="01"
,ylim=c(0,1),xlab="Dias",ylab=expression(p[01](t)))
lines(probest01[,1],p01est4[,1])
lines(probest01[,1],p01est4[,2],col="red",lty=2)
lines(probest01[,1],p01est4[,3],col="red",lty=2)

#10
n=1.5*(sum(probest10[,3])/max(probest10[,3]))
plot(probest10[,1],probest10[,2],main="10",cex=n*(probest10[,3]/sum(probest10[,3])),pch=19

```

```

,ylim=c(0,1),xlab="Dias",ylab=expression(p[10](t)))
lines(probest10[,1],p10est4[,1])
lines(probest10[,1],p10est4[,2],col="red",lty=2)
lines(probest10[,1],p10est4[,3],col="red",lty=2)

#11
n=1.5*(sum(probest11[,3])/max(probest11[,3]))
plot(probest11[,1],probest11[,2],cex=n*(probest11[,3]/sum(probest11[,3])),main="11",pch=19
,ylim=c(0,1),xlab="Dias",ylab=expression(p[11](t)))
lines(probest11[,1],p11est5[,1])
lines(probest11[,1],p11est5[,2],col="red",lty=2)
lines(probest11[,1],p11est5[,3],col="red",lty=2)

library(xtable)
xtable(Aic)
xtable(Bic)
xtable(jicudad)

#####
#Se repite el procedimiento pero con cadenas de Markov de orden 1 #
#####

#Programa para generar la cadena de Markov de primer orden
cmarkovorden1real=funcion(datos1=dataEstbin,dia.ant=0)
{exitos=cbind(1:366,0)
casos=cbind(1:366,0)
for (i in 1:366)
{if (i>=2 & i!=61 & i!=60) #dias comunes
{largoNum=0
largoDen=0
for (j in (año.inicial+1):año.final)
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-1) | datos1$time==i),4]==c(dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-1) | datos1$time==i),4]==c(dia.ant,0)))
{largoDen=1+largoDen}
}
}
exitos[i,2]=largoNum
casos[i,2]=largoDen
}

if (i==1) #1er dia del año
{largoNum=0
largoDen=0
for (j in (año.inicial+1):año.final)
{if(all(datos1[datos1$A.0==j & (datos1$time==366 | datos1$time==i),4]==c(dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==366 | datos1$time==i),4]==c(dia.ant,0)))
{largoDen=1+largoDen}
}
}
exitos[i,2]=largoNum
casos[i,2]=largoDen
}

if (i==60) #29 febrero
{largoNum=0
largoDen=0
for (j in (año.inicial+1):año.final)
{if(j/4==floor(j/4))
{if(all(datos1[datos1$A.0==j & ( datos1$time==(i-1) | datos1$time==i),4]==c(dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-1) | datos1$time==i),4]==c(dia.ant,0)))
{largoDen=1+largoDen}}
}
}
exitos[i,2]=largoNum
casos[i,2]=largoDen
}

```

```

}
if (i==61) #1er dia de marzo
{largoNum=0
largoDen=0
for (j in (año.inicial+1):año.final)
{if(j/4==floor(j/4))
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-1) | datos1$time==i),4]==c(dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-1) | datos1$time==i),4]==c(dia.ant,0)))
{largoDen=1+largoDen}}
if(j/4!=floor(j/4))
{if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==i),4]==c(dia.ant,1)))
{largoNum=1+largoNum
largoDen=1+largoDen}
if(all(datos1[datos1$A.0==j & (datos1$time==(i-2) | datos1$time==i),4]==c(dia.ant,0)))
{largoDen=1+largoDen}
}}
exitos[i,2]=largoNum
casos[i,2]=largoDen
}
}
return(cbind(exitos,casos[,2]))
}

mcorden1_0=cmarkovorden1real(dia.ant=0)
mcorden1_1=cmarkovorden1real(dia.ant=1)

#####
#Verosimilitud #
#####
probest1=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest1[i,1]
probest1[i,2]=(sum(mcorden1_1[(t-2):(t+3),3])-sum(mcorden1_1[(t-2):(t+3),2]))/sum(mcorden1_1[(t-2):(t+3),3])
probest1[i,3]=(sum(mcorden1_1[(t-2):(t+3),3])-sum(mcorden1_1[(t-2):(t+3),2]))
probest1[i,4]=sum(mcorden1_1[(t-2):(t+3),2])/sum(mcorden1_1[(t-2):(t+3),3])
probest1[i,5]=sum(mcorden1_1[(t-2):(t+3),2])
}
colnames(probest1)=c("t","p110","n110","p111","n111")

probest0=cbind(seq(3,366,by=6),rep(0,366/6),rep(0,366/6),rep(0,366/6),rep(0,366/6))
for (i in 1:(366/6))
{t=probest0[i,1]
probest0[i,2]=(sum(mcorden1_0[(t-2):(t+3),3])-sum(mcorden1_0[(t-2):(t+3),2]))/sum(mcorden1_0[(t-2):(t+3),3])
probest0[i,3]=(sum(mcorden1_0[(t-2):(t+3),3])-sum(mcorden1_0[(t-2):(t+3),2]))
probest0[i,4]=sum(mcorden1_0[(t-2):(t+3),2])/sum(mcorden1_0[(t-2):(t+3),3])
probest0[i,5]=sum(mcorden1_0[(t-2):(t+3),2])
}
colnames(probest0)=c("t","p110","n110","p111","n111")

vero1=sum(log(probest1[,2])*probest1[,3]+log(probest1[,4])*probest1[,5])+
sum(log(probest0[,2])*probest0[,3]+log(probest0[,4])*probest0[,5])

#H0 es de 1er orden
(-2)*vero1+2*vero2

```

#### B.5.4. Simulación del modelo de ocurrencia de lluvia.

```

set.seed(1)
probabilidadesFinales=cbind(rep(p00est4[,1],each = 6),rep(p01est3[,1],each = 6)
,rep(p10est4[,2],each = 6),rep(p11est5[,1],each = 6))

FrecuenciasSimuladasDiaria=matrix(data = rep(0,366*100), nrow = 366, ncol = 100, byrow = FALSE,

```

```

        dimnames = list(1:366,0:99)
dataEstbin=dataEstbin[,1:5]

for (i in 1:500)
{vectorInicial=dataEstbin[c(1,2),4]
if(all(vectorInicial[2]!=1))
{Empieza=0
j=3
while(Empieza==0)
{if(dataEstbin[j,4]!=1){j=j+1}
if(dataEstbin[j,4]==1)
        {Empieza=1
          j=j+1}
}
Contador=0}
if(vectorInicial[2]==1){j=3}
estadoanterior1=dataEstbin[j-1,4]
estadoanterior2=dataEstbin[j-2,4]
diautilizado=dataEstbin[j,5]
while(j<=(dim(dataEstbin)[1]-1))
{simulacion=runif(1)
  if(estadoanterior1==0 & estadoanterior2==0)
    {if (simulacion>probabilidadesFinales[dataEstbin[j,5],1]){Contador=Contador+1}
      if (simulacion<=probabilidadesFinales[dataEstbin[j,5],1])
        {FrecuenciasSimuladasDiaria[diautilizado,Contador]=FrecuenciasSimuladasDiaria[diautilizado,Contador]+1
          diautilizado=dataEstbin[j+1,5]
          Contador=0}}
    if(estadoanterior1==1 & estadoanterior2==0)
      {if (simulacion>probabilidadesFinales[dataEstbin[j,5],2]){Contador=Contador+1}
        if (simulacion<=probabilidadesFinales[dataEstbin[j,5],2])
          {FrecuenciasSimuladasDiaria[dataEstbin[j,5],Contador]=FrecuenciasSimuladasDiaria[dataEstbin[j,5],Contador]+1
            diautilizado=dataEstbin[j+1,5]
            Contador=0}}
    if(estadoanterior1==0 & estadoanterior2==1)
      {if (simulacion>probabilidadesFinales[dataEstbin[j,5],3]){Contador=Contador+1}
        if (simulacion<=probabilidadesFinales[dataEstbin[j,5],3])
          {FrecuenciasSimuladasDiaria[dataEstbin[j,5],Contador]=FrecuenciasSimuladasDiaria[dataEstbin[j,5],Contador]+1
            diautilizado=dataEstbin[j+1,5]
            Contador=0}}
    if(estadoanterior1==1 & estadoanterior2==1)
      {if (simulacion>probabilidadesFinales[dataEstbin[j,5],4]){Contador=Contador+1}
        if (simulacion<=probabilidadesFinales[dataEstbin[j,5],4])
          {FrecuenciasSimuladasDiaria[dataEstbin[j,5],Contador]=FrecuenciasSimuladasDiaria[dataEstbin[j,5],Contador]+1
            diautilizado=dataEstbin[j+1,5]
            Contador=0}}
    j=j+1
    estadoanterior2=estadoanterior1
    if(Contador==0){estadoanterior1=1}
    if(Contador!=0){estadoanterior1=0}
  }
}

write.table(FrecuenciasSimuladasDiaria,file = "simulaciones.txt", dec = ",",sep="\t",row.names =FALSE)
FrecuenciasSimuladasDiariaAGR=NULL
for (i in 1:61)
{FrecuenciasSimuladasDiariaAGR=rbind(FrecuenciasSimuladasDiariaAGR
,colSums(FrecuenciasSimuladasDiaria[seq(6*(i-1)+1,6*i,1),]))
}

ProbabilidadesSimuladasDiariaAGR=NULL
for (i in 1:61)
{ProbabilidadesSimuladasDiariaAGR=rbind(ProbabilidadesSimuladasDiariaAGR
,colSums(FrecuenciasSimuladasDiaria[seq(6*(i-1)+1,6*i,1),])/sum(FrecuenciasSimuladasDiaria[seq(6*(i-1)+1,6*i,1),]))
}

distribucionsimulada=colMeans(ProbabilidadesSimuladasDiariaAGR)

```

```

#####
#Una vez hecha la teórica, tengo que simular la real. #
#####

frecuenciasReales=matrix(data = rep(0,366*100), nrow = 366, ncol = 100, byrow = FALSE,dimnames = list(1:366,0:99))

vectorInicialR=dataEstbin[c(1,2),4]
if(all(vectorInicialR[2]!=1))
{EmpiezaR=0
j=3
while(EmpiezaR==0)
{if(dataEstbin[j,4]!=1){j=j+1}
if(dataEstbin[j,4]==1)
{EmpiezaR=1
j=j+1}
}
ContadorR=0}
if(vectorInicialR[2]==1){j=3}
estadoanteriorR=dataEstbin[j-1,4]
diautilizadoR=dataEstbin[j,5]
while(j<dim(dataEstbin)[1])
{if(estadoanteriorR<0){EmpiezaR=0}
if(dataEstbin[j,4]==1 & estadoanteriorR>=0 & EmpiezaR==1)
{frecuenciasReales[diautilizadoR,ContadorR]=frecuenciasReales[diautilizadoR,ContadorR]+1
diautilizadoR=dataEstbin[j+1,5]
ContadorR=0}
if(dataEstbin[j,4]==0 & estadoanteriorR>=0 & EmpiezaR==1){ContadorR=ContadorR+1}
if(dataEstbin[j,4]==1 & EmpiezaR==0)
{EmpiezaR=1
diautilizadoR=dataEstbin[j,5]
ContadorR=0}
estadoanteriorR=dataEstbin[j,4]
j=j+1
}

ProbabilidadesRealesDiariaAGR=NULL
for (i in 1:61)
{ProbabilidadesRealesDiariaAGR=rbind(ProbabilidadesRealesDiariaAGR
,colSums(frecuenciasReales[seq(6*(i-1)+1,6*i,1),])/sum(frecuenciasReales[seq(6*(i-1)+1,6*i,1),]))
}

SumaFrecuenciasReales=colSums(frecuenciasReales)[-1]
SumaFrecuenciasSimuladas=colSums(FrecuenciasSimuladasDiaria)[-1]

sum(SumaFrecuenciasReales)/100

ProbFrecuenciasReales=SumaFrecuenciasReales/sum(SumaFrecuenciasReales)
ProbFrecuenciasSimuladas=SumaFrecuenciasSimuladas/sum(SumaFrecuenciasSimuladas)

par(mfrow=c(1,2))
barplot(ProbFrecuenciasReales[1:35],col="lightblue", axes = TRUE,ylim=c(0,0.2),ylab="Frecuencia relativa"
,xlab="Días esperados",cex.lab=1.5, cex.axis=1.5,main="Frecuencias relativas Reales")
barplot(ProbFrecuenciasSimuladas[1:35],col="lightblue", axes = TRUE,ylim=c(0,0.2),ylab=""
,xlab="Días esperados",cex.lab=1.5, cex.axis=1.5,main="Frecuencias relativas Simuladas")

ProbFrecuenciasAcuReales=NULL
ProbFrecuenciasAcuSimuladas=NULL
for (i in 1:99)
{ProbFrecuenciasAcuReales=c(ProbFrecuenciasAcuReales,sum(ProbFrecuenciasReales[1:i]))
ProbFrecuenciasAcuSimuladas=c(ProbFrecuenciasAcuSimuladas,sum(ProbFrecuenciasSimuladas[1:i]))
}

```

```

names(ProbFrecuenciasAcuReales)=1:99
names(ProbFrecuenciasAcuSimuladas)=1:99

par(mfrow=c(1,2))
barplot(ProbFrecuenciasAcuReales[1:35],col="lightblue", axes = TRUE,ylim=c(0,1),ylab="Frecuencia relativa acumulada",
,xlab="Días esperados",cex.lab=1.5, cex.axis=1.5,main="Frecuencias relativas acumuladas Reales")
barplot(ProbFrecuenciasAcuSimuladas[1:35],col="lightblue", axes = TRUE,ylim=c(0,1),
,ylab="",xlab="Días esperados",cex.lab=1.5, cex.axis=1.5,main="Frecuencias relativas acumuladas Simuladas")

ProbabilidadesSimuladasAcuDiariaAGR=ProbabilidadesSimuladasDiariaAGR[,1]
for (i in 2:99)
{ProbabilidadesSimuladasAcuDiariaAGR=cbind(ProbabilidadesSimuladasAcuDiariaAGR
,rowSums(ProbabilidadesSimuladasDiariaAGR[,1:i]))
}
colnames(ProbabilidadesSimuladasAcuDiariaAGR)=1:99

write.table(SumaFrecuenciasReales,"frecuenciasMeteo.txt", dec = ",",sep="\t",row.names =FALSE)

#####
#Puntos problemáticos #
#####

ValoresAtipicos=NULL ##Ir  el a o,mes y d a
umbral=0.99 #el valor del umbral para considerar una racha como racha seca

vectorInicialR=dataEstbin[c(1,2),4]
if(all(vectorInicialR[2]!=1))
{EmpiezaR=0
j=3
while(EmpiezaR==0)
{if(dataEstbin[j,4]!=1){j=j+1}
if(dataEstbin[j,4]==1)
{EmpiezaR=1
j=j+1}
}
ContadorR=0}
if(vectorInicialR[2]==1)
{j=3
ContadorR=0
EmpiezaR=1}
estadoanteriorR=dataEstbin[j-1,4]
diautilizadoR=dataEstbin[j,5]
diaInicial=dataEstbin[j,1:3]
while(j<dim(dataEstbin)[1])
{if(estadoanteriorR<0){EmpiezaR=0}
if(dataEstbin[j,4]==1 & estadoanteriorR>=0 & ContadorR>0 & EmpiezaR==1)
{if(ProbabilidadesSimuladasAcuDiariaAGR[ceiling(diautilizadoR/6),ContadorR]>umbral)
{ValoresAtipicos=rbind(ValoresAtipicos,c(diaInicial,ContadorR))}
diautilizadoR=dataEstbin[j+1,5]
diaInicial=dataEstbin[j+1,1:3]
ContadorR=0}
if(dataEstbin[j,4]==1 & estadoanteriorR>=0 & EmpiezaR==1)
{diautilizadoR=dataEstbin[j+1,5]
diaInicial=dataEstbin[j+1,1:3]
ContadorR=0}
if(dataEstbin[j,4]==0 & estadoanteriorR>=0 & EmpiezaR==1){ContadorR=ContadorR+1}
if(dataEstbin[j,4]==1 & EmpiezaR==0)
{EmpiezaR=1
diautilizadoR=dataEstbin[j,5]
ContadorR=0}
estadoanteriorR=dataEstbin[j,4]
j=j+1
}

xtable(ValoresAtipicos)

```