RELLENO DE LAGUNAS DE LOS DATOS OBSERVADOS SUMINISTRADOS

PARA EL PROYECTO

"ELABORACIÓN DE ESCENARIOS LOCALES Y REGIONALES DE CAMBIO CLIMÁTICO ADAPTADOS AL SEXTO INFORME DEL GRUPO INTERGUBERNAMENTAL DE EXPERTOS SOBRE EL CAMBIO CLIMÁTICO (IPCC) Y SU DIFUSIÓN A TRAVÉS DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN"

INTRODUCCIÓN 2

1. ZONIFICACIÓN DE LOS OBSERVATORIOS 3

2. RELLENO DE LAGUNAS DE TEMPERATURA 4

2.1. NÚMERO DE VECINOS 4

2.2. CORRELACIÓN 7

2.3. METODOLOGÍA 9

2.4. RESULTADOS 10

3. RELLENO DE LAGUNAS DE PRECIPITACIÓN 15

3.1. MÉTODO 1 15

3.1.1. NÚMERO DE VECINOS 15

3.1.2. CORRELACIÓN 18

3.1.3. METODOLOGÍA 21

3.1.4. RESULTADOS 22

3.2. MÉTODO 2: REDDPREC PACKAGE 26

3.2.1. RESULTADOS 26

3.3. SELECCIÓN DEL MÉTODO 31

REFERENCIAS 32

INTRODUCCIÓN

El presente documento detalla el proceso de Relleno de Lagunas de los datos observados suministrados para el proyecto “Elaboración de escenarios locales y regionales de cambio climático adaptados al sexto informe del grupo intergubernamental de expertos sobre el cambio climático (IPCC) y su difusión a través de las tecnologías de la información y la comunicación” de la Junta de Andalucía.

El proceso del relleno de lagunas viene motivado por el hecho de que las series temporales diarias de datos observados (tanto de precipitación como de temperatura), precisamente las utilizadas en este proyecto, tienen huecos (lagunas), es decir, de algunos días y a veces de muchos no tenemos el valor diario correspondiente. El objetivo de este estudio es intentar rellenar esas lagunas con algunos valores que se puedan inferir a partir de los valores de observatorios meteorológicos lo suficientemente cercanos.

Lo que sigue es un intento de determinar una cierta metodología para ese relleno de lagunas después de realizar un estudio sobre tales series temporales y los observatorios a los que están asociadas.

1. ZONIFICACIÓN DE LOS OBSERVATORIOS

Para el siguiente proyecto, y como parte del propio pliego de condiciones técnicas, con el fin de poder estudiar con mayor exactitud determinadas variables se ha dividido Andalucía en cinco zonas climáticas atendiendo a criterios de altitud, orientación, temperatura media y distribución de precipitaciones. Estas cinco zonas son: Norte, Este, Guadalquivir, Cádiz y Málaga. Son cinco zonas que definen otras cinco zonas climáticas o, al menos, de influencia climática homogénea, y por eso en nuestro estudio del relleno de lagunas de las series observadas suministradas lo que haremos será intentar rellenar las lagunas de cada posible observatorio solamente con observatorios de su misma área climática. La figura 1.1 muestra esas cinco áreas climáticas en las que se divide el área de Andalucía.

Figura 1.1: Mapa de la división de Andalucía en zonas climáticas.

Para la realización del estudio, se han utilizado 789 observatorios cuyas series han sido anteriormente homogeneizadas y que cuentan con un mínimo de 2000 datos.

2. RELLENO DE LAGUNAS DE TEMPERATURA

2.1. NÚMERO DE VECINOS

El primer paso de cara a llevar a cabo el relleno de lagunas de temperatura ha sido estimar la proximidad de los observatorios con los que, en teoría, vamos a proceder a rellenar las lagunas. Para ello, se ha realizado un estudio en cada una de las cinco áreas climáticas calculando el número de vecinos que tiene cada observatorio hasta una distancia de 50 km, obteniéndose los siguientes resultados:

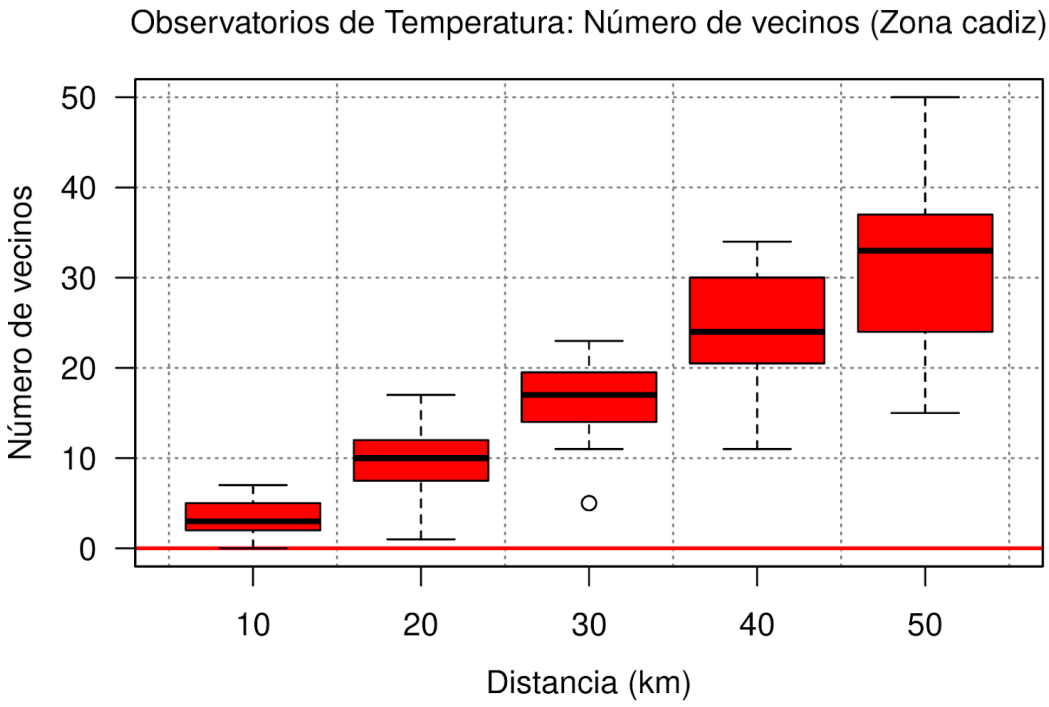


Figura X: Número de observatorios vecinos en función de la distancia en Cádiz.

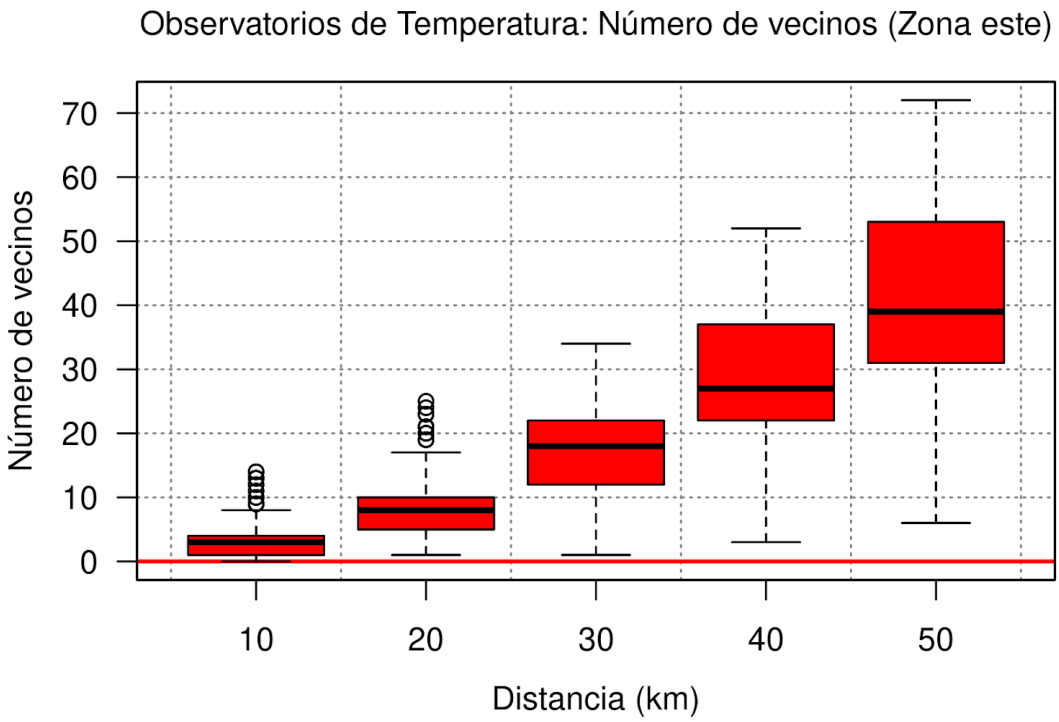


Figura X: Número de observatorios vecinos en función de la distancia en Este.

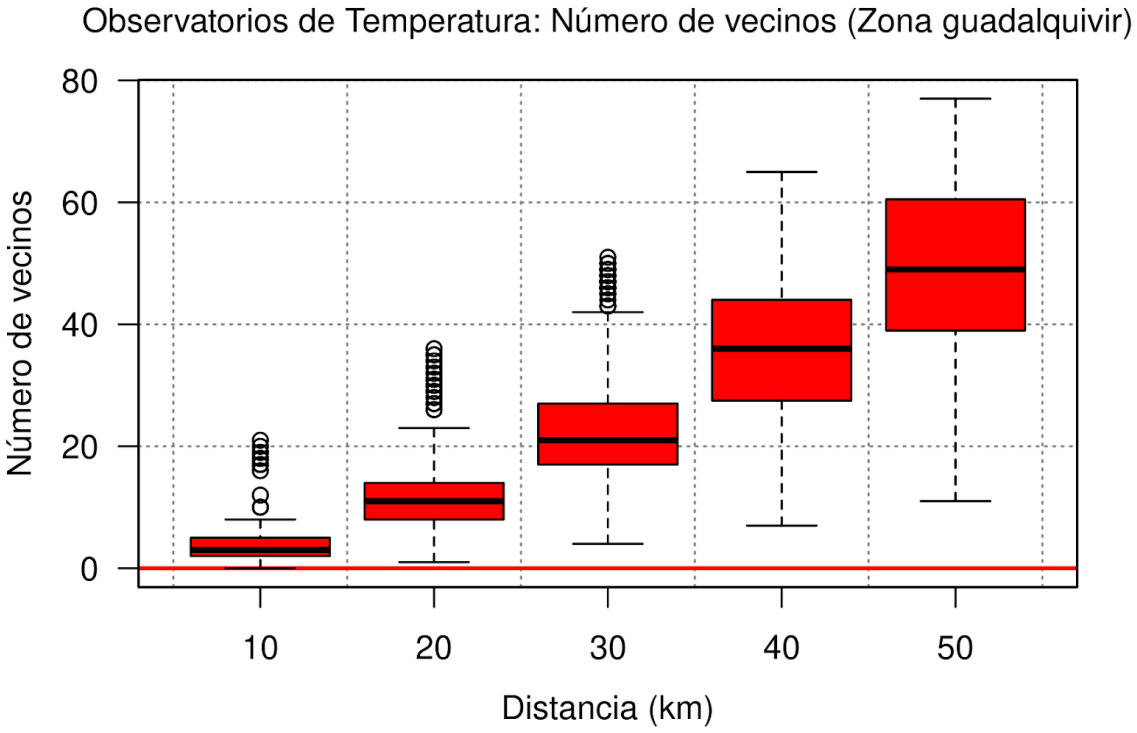


Figura X: Número de observatorios vecinos en función de la distancia en Guadalquivir.

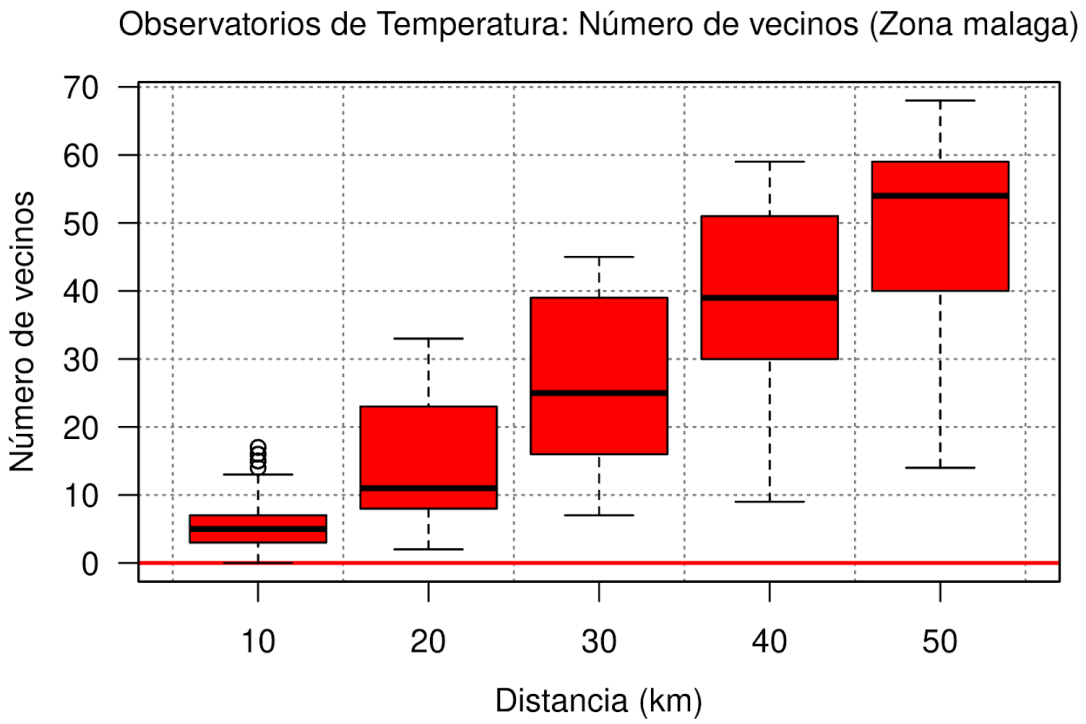


Figura X: Número de observatorios vecinos en función de la distancia en Málaga.

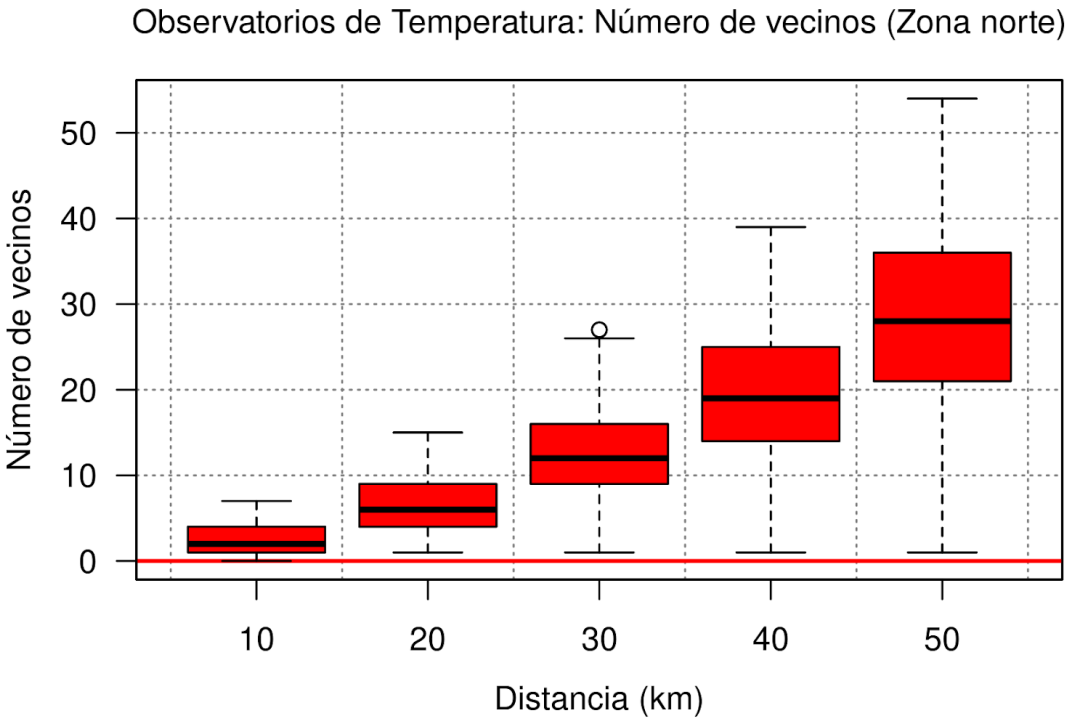


Figura X: Número de observatorios vecinos en función de la distancia en Norte.

Como se puede observar, cada observatorio tiene una media de unos 5 observatorios vecinos a una distancia inferior a 10 km. Hay alguno, sobre todo en grandes áreas urbanas, que tiene hasta 20 observatorios vecinos, pero lo relevante es que hay muchos que no tienen ninguno. Si se sube la distancia a 20km, se observa que el número medio de vecinos a una distancia inferior a esos 20km sube a 10 observatorios, y también se puede ver que, independientemente de la zona, todos los observatorios tienen al menos un vecino a una distancia inferior a esos 20km. Si la distancia se sube a 50 km, se observa que cada observatorio tiene una media de entre 30 y 60 observatorios vecinos en un radio inferior a esa distancia, observándose algunos cambios en función de la zona climática.

2.2. CORRELACIÓN

A la hora de tomar la decisión sobre con qué observatorios rellenar la serie de un determinado observatorio, es imprescindible calcular la correlación de los datos entre observatorios. Para ello, se ha realizado un estudio para determinar la correlación entre observaciones en función de su distancia. Para cada observatorio, se han calculado las distancias y la correlación respecto al resto de observatorios de la zona, tanto para temperatura máxima como para mínima, y se ha representado el número total de pares de observatorios agrupados por distancias frente a la correlación obtenida:

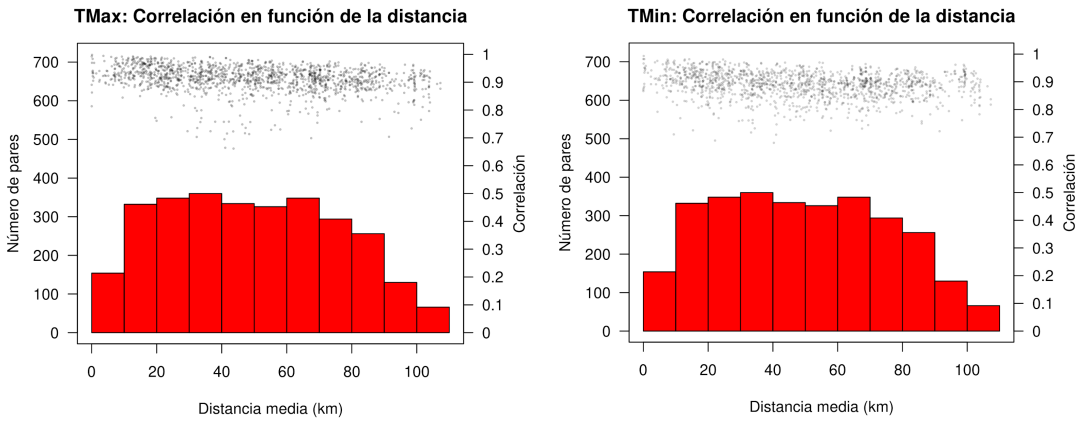


Figura 2.2.1: Correlación entre valores de distintos observatorios en función de su distancia para las variables TMax y TMin en Cádiz.

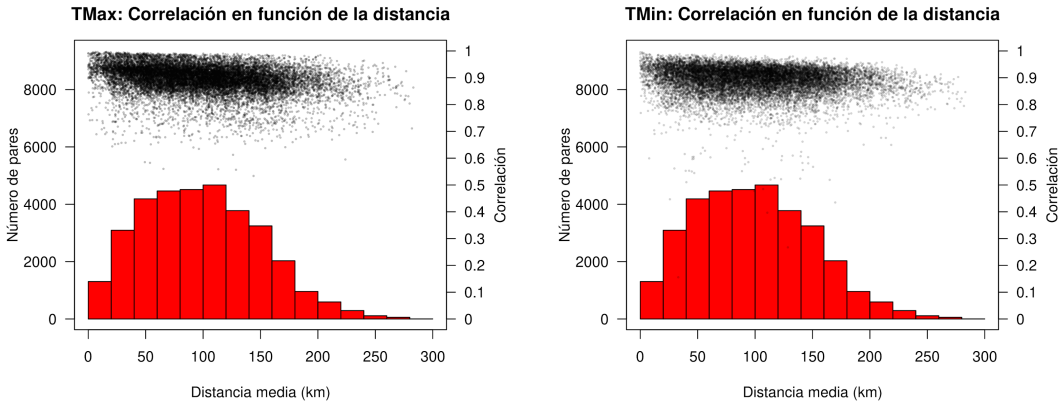


Figura 2.2.2: Correlación entre valores de distintos observatorios en función de su distancia para las variables TMax y TMin en Este.

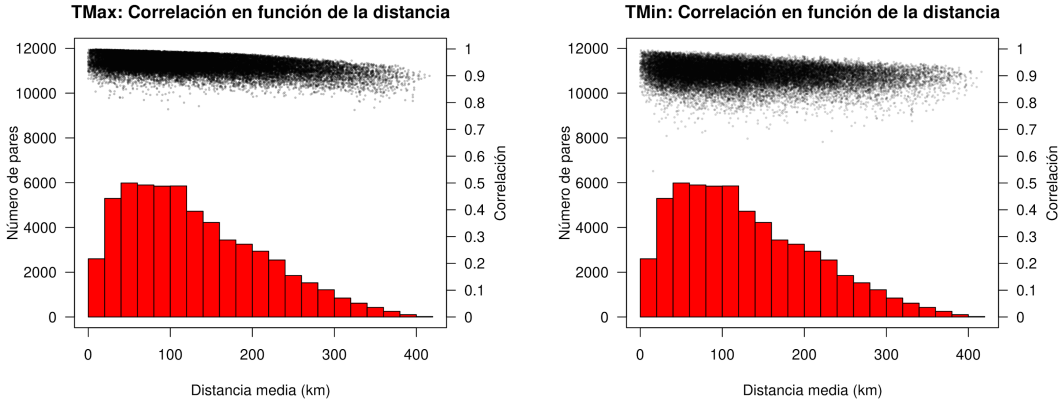


Figura 2.2.3: Correlación entre valores de distintos observatorios en función de su distancia para las variables TMax y TMin en Guadalquivir.

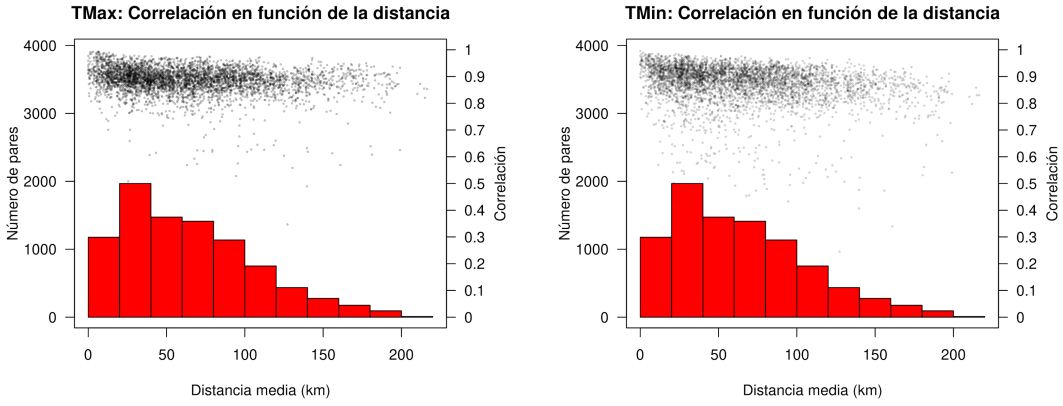


Figura 2.2.4: Correlación entre valores de distintos observatorios en función de su distancia para las variables TMax y TMin en Málaga.

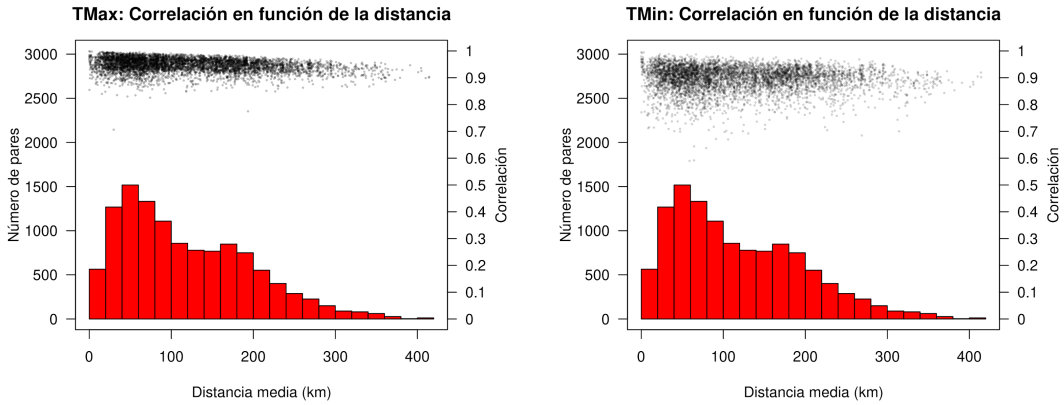


Figura 2.2.5: Correlación entre valores de distintos observatorios en función de su distancia para las variables TMax y TMin en Norte.

Lo más destacable de los resultados anteriores es que la correlación de las series, independientemente de la distancia entre observatorios, es muy alta, superior a 0.7. En observatorios muy cercanos la correlación es altísima, entre 0.95 y 1. Como es de esperar, a medida que aumenta la distancia entre observatorios, la nube de puntos de la correlación se va desplazando hacia valores inferiores, pero a distancias muy altas la correlación sigue siendo muy buena, por lo general superior a 0.8.

Se observa el mismo patrón tanto en temperatura máxima como en mínima, aunque se observa que la correlación es ligeramente inferior en temperatura mínima, incluso para distancias cortas. Esto puede deberse a que las temperaturas mínimas pueden llegar a ser muy diferentes en emplazamientos muy cercanos debido a las inversiones térmicas, algo que no ocurre con las temperaturas máximas.

2.3. METODOLOGÍA

Una vez analizados los resultados del número de vecinos cercanos y la correlación, se ha procedido a establecer la siguiente metodología para realizar el relleno de lagunas de temperatura:

1) Para cada observatorio, se seleccionan los observatorios situados a una distancia inferior a 20 km y que tengan una correlación superior a 0.7
2) En caso de que haya menos de 6 observatorios que cumplan estos requisitos, se ampliará de forma recursiva el radio de cercanía en 5km hasta que haya 6 observatorios cercanos.
3) De los 6 observatorios que cumplen los requisitos anteriores, se seleccionan los 3 que tengan mayor número de datos entre los años 1979 y 2020.
4) Se realiza una regresión lineal múltiple con los datos que tienen en común el observatorio que se quiere rellenar y sus 3 observatorios seleccionados anteriormente.
5) Con los parámetros de la regresión lineal, se rellenan los huecos del observatorio tipo que están vacíos y cuyos datos de los 3 observatorios seleccionados están completos.

No se rellenan huecos de un determinado día si cualquiera de los 3 observatorios con los que se ha realizado la regresión lineal no tienen datos para ese día.

6) Se repiten 6 veces todos los puntos anteriores con el objetivo de que, en cada iteración, se vayan rellenando cada vez más huecos. En las primeras 3 iteraciones, se rellenarán huecos de toda la serie, mientras que en las 3 últimas iteraciones, únicamente se rellenarán huecos entre los años 1979 y 2020.

2.4. RESULTADOS

En primer lugar, se ha realizado una prueba para obtener el error medio al efectuar el relleno de lagunas con la metodología propuesta. Para ello, a cada serie a rellenar (observatorio tipo) se le ha eliminado un tercio de los datos y se le ha aplicado la metodología de relleno de lagunas para obtener esos mismos datos. A continuación, se ha comparado la serie original con la serie obtenida con datos rellenos y se han calculado dos parámetros para determinar su error: el BIAS y el MAE. Los resultados de esta prueba han sido los siguientes:

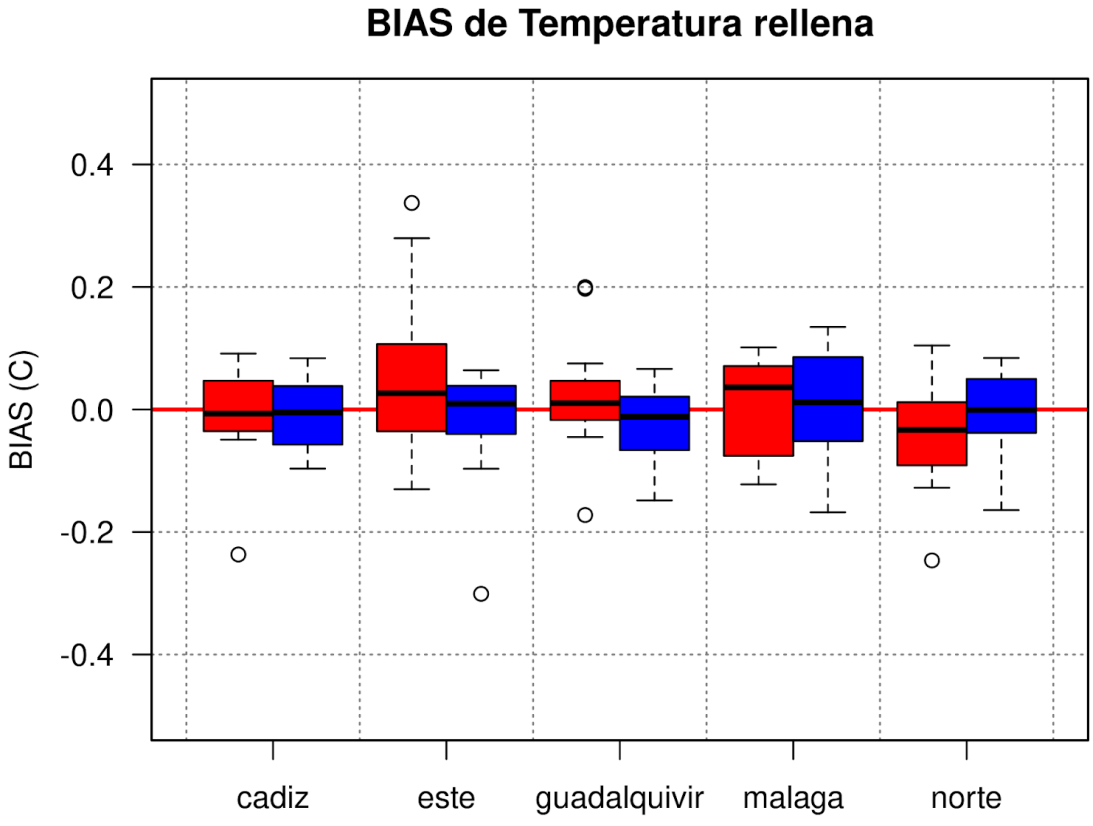


Figura 2.4.1: BIAS de los datos rellenos de temperatura máxima (en rojo) y mínima (en azul) por zonas.

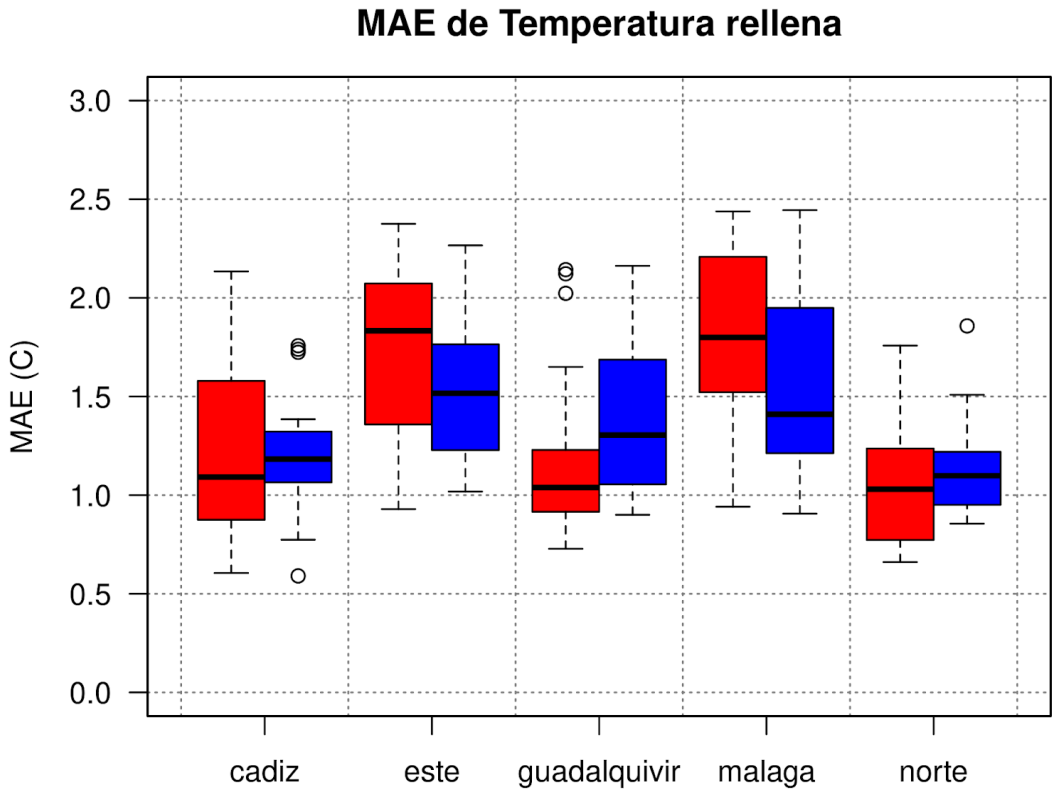


Figura 2.4.2: MAE de los datos rellenos de temperatura máxima (en rojo) y mínima (en azul) por zonas.

Como se puede observar, los datos rellenos presentan muy poco BIAS respecto a los valores reales. Hay mayor variabilidad en algunas zonas geográficas, como puede ser Este o Málaga, muy probablemente debido a la gran diferencia en altitud que puede haber entre observatorios cercanos, pero lo importante es que la mediana está muy cercana a 0 tanto en temperatura máxima como mínima y en las cinco áreas geográficas. Esto nos indica que la simulación de los datos es muy buena y que el método utilizado no genera un error sistemático.

En cuanto al MAE, lo más destacable es que, en tres de las cinco zonas (Cádiz, Guadalquivir y Norte) la mediana del error está por debajo de 1.5º tanto en temperatura máxima como en temperatura mínima. En temperaturas máximas, el error mediano es incluso cercano a 1º. Teniendo en cuenta que el error sistemático de los sensores se suele situar en 1º, estos resultados son excelentes, ya que prácticamente no habría distinción entre el error cometido al rellenar huecos y el error sistemático del termómetro. El error cometido al rellenar temperatura mínima es ligeramente superior (entre 1.2º y 1.5º) debido a lo que ya se ha comentado de que, debido a las inversiones térmicas, puede haber diferencias de temperatura muy significativas en observatorios que están a poca distancia. En el caso de las áreas geográficas de Este y Málaga, el error en temperatura máxima es superior (cercano a 2º) debido seguramente a que son zonas con terreno más escarpado y, por tanto, en observatorios cercanos hay mayor diferencia de temperatura debido a diferencias en altitud o en orientación.

Una vez realizado el proceso de relleno de lagunas en 6 iteraciones, en primer lugar se ha obtenido el número de observatorios que han rellenado algún hueco en la iteración 1 y en la iteración 6, como podemos ver en la tabla 2.4.1.

Tabla 2.4.1: Número de observatorios que rellenan algún hueco tras las primera y última iteración.

Zona	Observatorios totales	Observatorios rellenos iteración 1	Observatorios rellenos iteración 6
Cádiz	60	58	50
Este	213	187	187
Guadalquivir	276	266	242
Málaga	111	106	105
Norte	129	121	102

A continuación, se ha calculado el porcentaje de huecos que hay en el periodo 1979-2020 tras cada iteración:

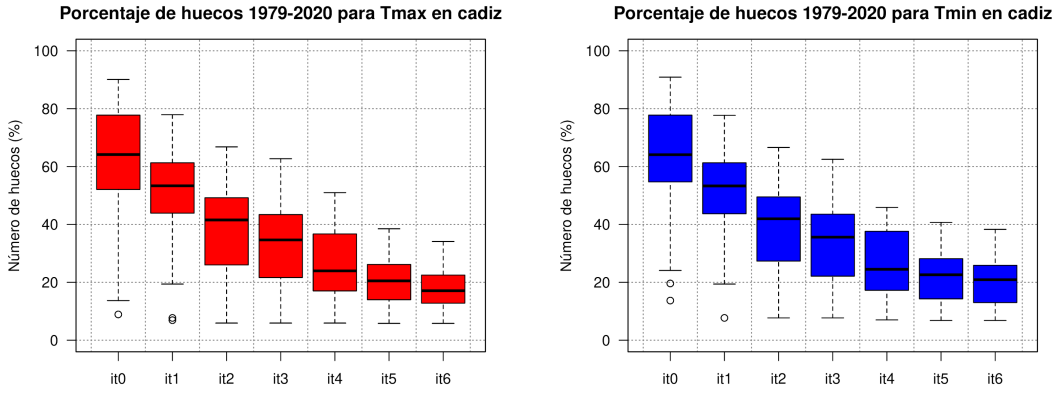


Figura 2.4.3: Porcentaje de huecos entre 1979 y 2020 en los observatorios de Cádiz.

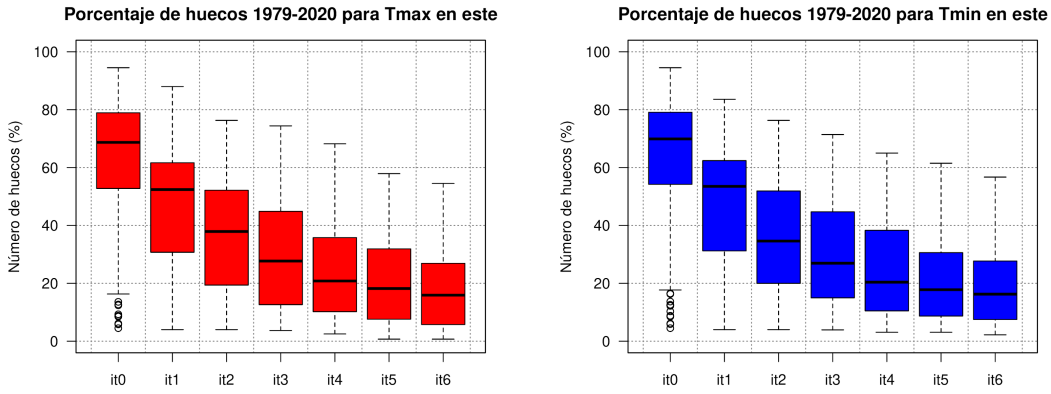


Figura 2.4.4: Porcentaje de huecos entre 1979 y 2020 en los observatorios de Este.

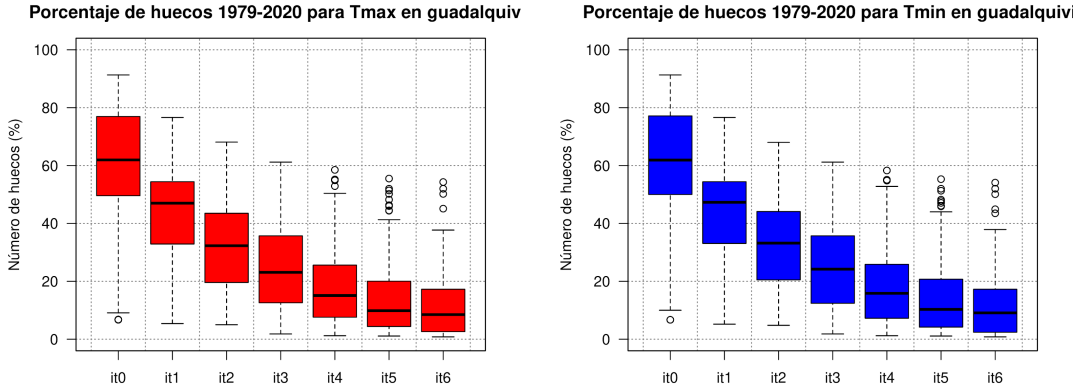


Figura 2.4.5: Porcentaje de huecos entre 1979 y 2020 en los observatorios de Guadalquivir.

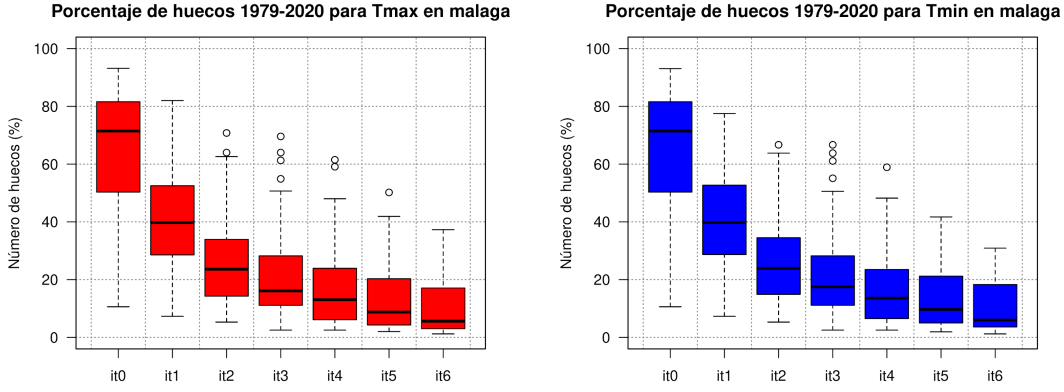


Figura 2.4.6: Porcentaje de huecos entre 1979 y 2020 en los observatorios de Málaga.

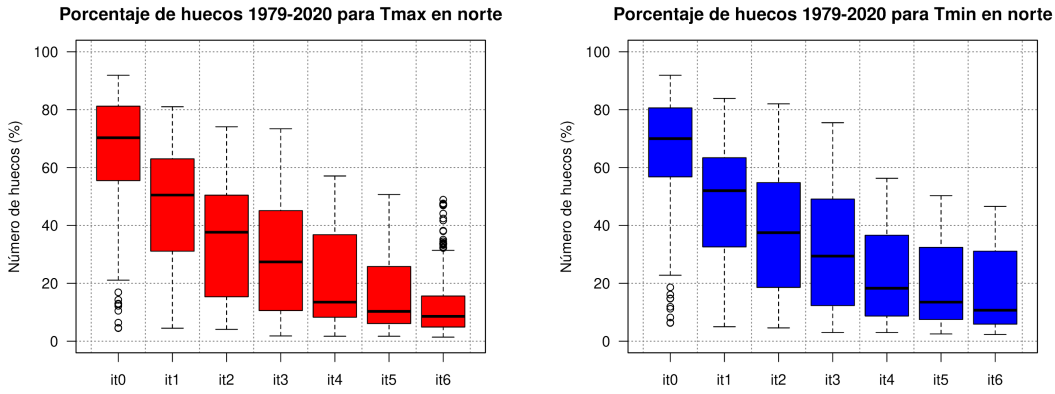


Figura 2.4.7: Porcentaje de huecos entre 1979 y 2020 en los observatorios de Norte.

Inicialmente, se observa que la mediana de los huecos en todos los observatorios supera el 60%, situándose entre el 60%y 65% en las zonas de Cádiz y Guadalquivir y entre el 70% y el 75% en Este, Málaga y Norte. En la primera iteración, el número de huecos se desploma casi un 20% en todas las zonas, y a medida que pasan las iteraciones, se van rellenando más huecos aunque a un ritmo decreciente. De esta manera, después de la sexta iteración, el número de huecos en las series de Guadalquivir, Málaga y Norte no supera el 10% mientras que en los casos de Cádiz y Este, el número final de huecos se sitúa entre el 15% y el 20%.

La conclusión final de este estudio es que se puede reducir drásticamente el número de lagunas de una serie cometiendo un error muy pequeño, inferior a 1,5º.

3. RELLENO DE LAGUNAS DE PRECIPITACIÓN

Para el relleno de lagunas de precipitación se emplean dos métodos diferentes con el objetivo de seleccionar el que mejor funcione, pues el proceso de relleno de lagunas diarias resulta más complejo en este caso. Para evaluar la efectividad de cada método, se seleccionan los 20 observatorios con mayor longitud de datos diarios para cada una de las 5 zonas y se elimina de forma aleatoria un tercio de los datos. Así, se comparan los datos originales con los datos rellenados según los diferentes métodos para evaluar la efectividad de cada uno de ellos.

3.1. MÉTODO 1

3.1.1. NÚMERO DE VECINOS

Al igual que en temperatura, el primer paso de cara a llevar a cabo el relleno de lagunas de precipitación ha sido estimar la proximidad de los observatorios. Para ello, se ha realizado un estudio en cada una de las cinco zonas calculando el número de vecinos que tiene cada observatorio hasta una distancia de 50 km, obteniéndose los siguientes resultados:

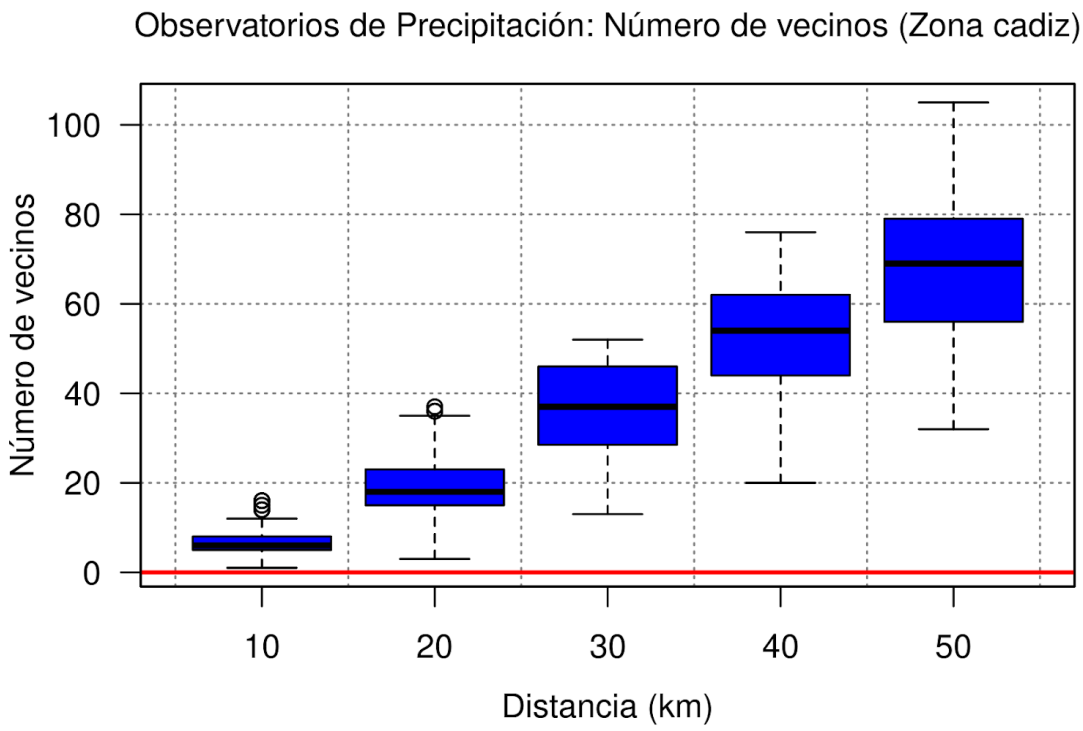


Figura X: Número de observatorios de precipitación vecinos en función de la distancia en Cádiz.

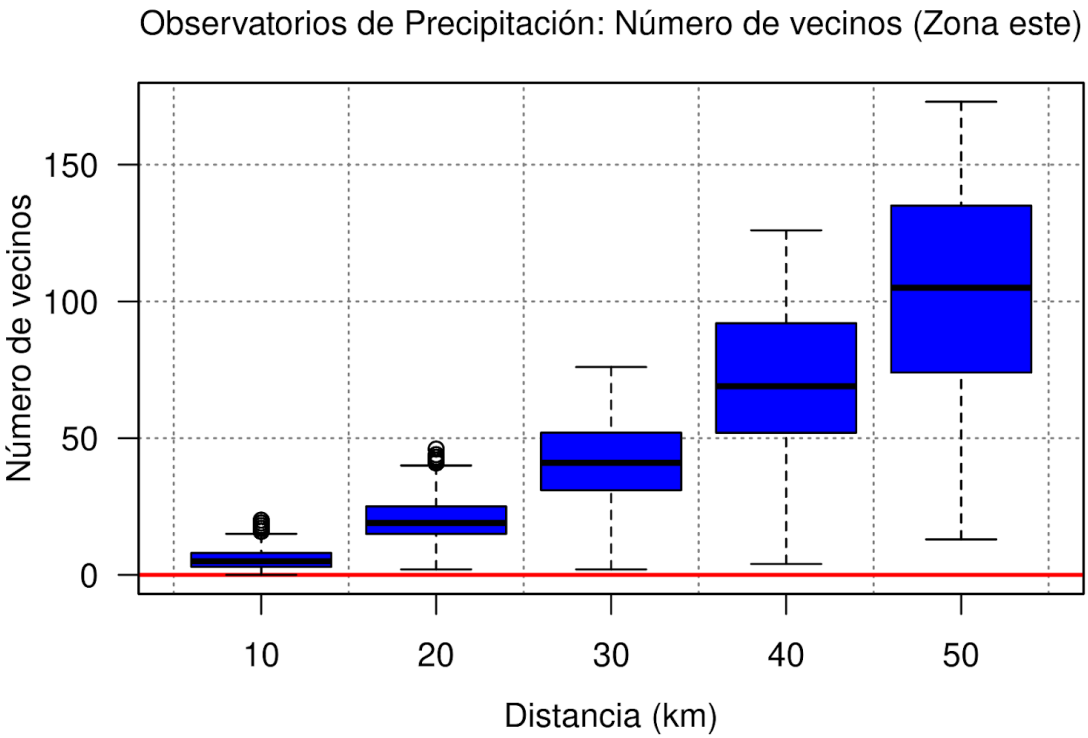


Figura X: Número de observatorios de precipitación vecinos en función de la distancia en Este.

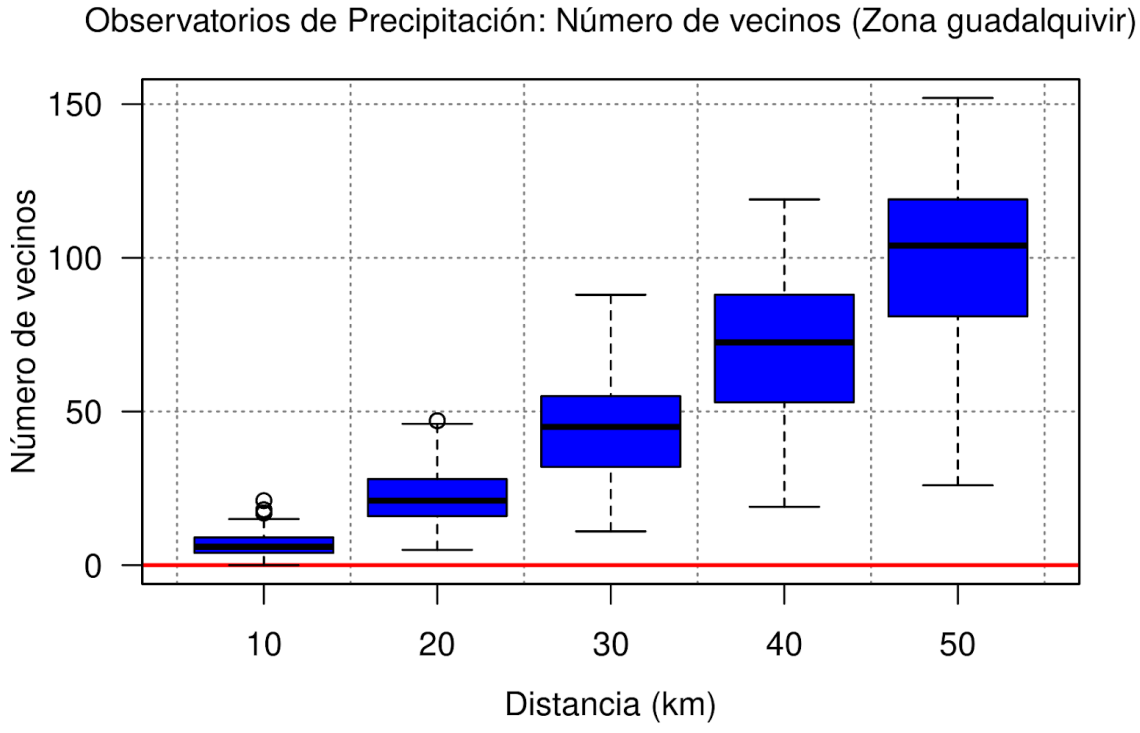


Figura X: Número de observatorios de precipitación vecinos en función de la distancia en Guadalquivir.

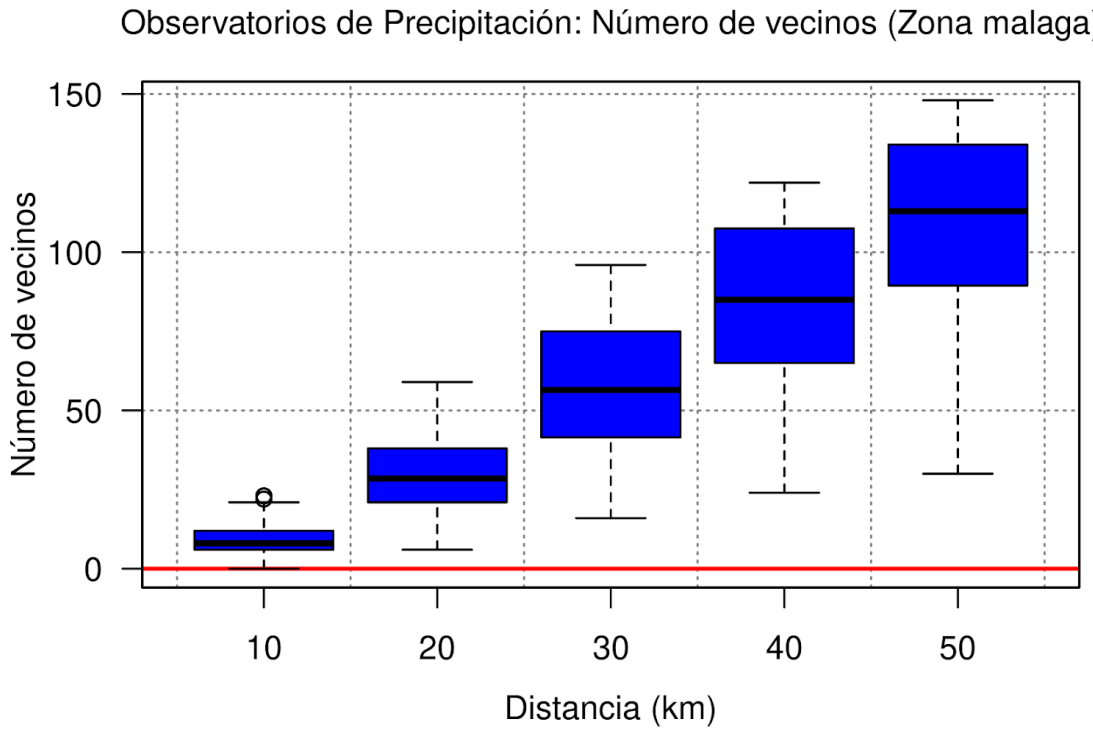
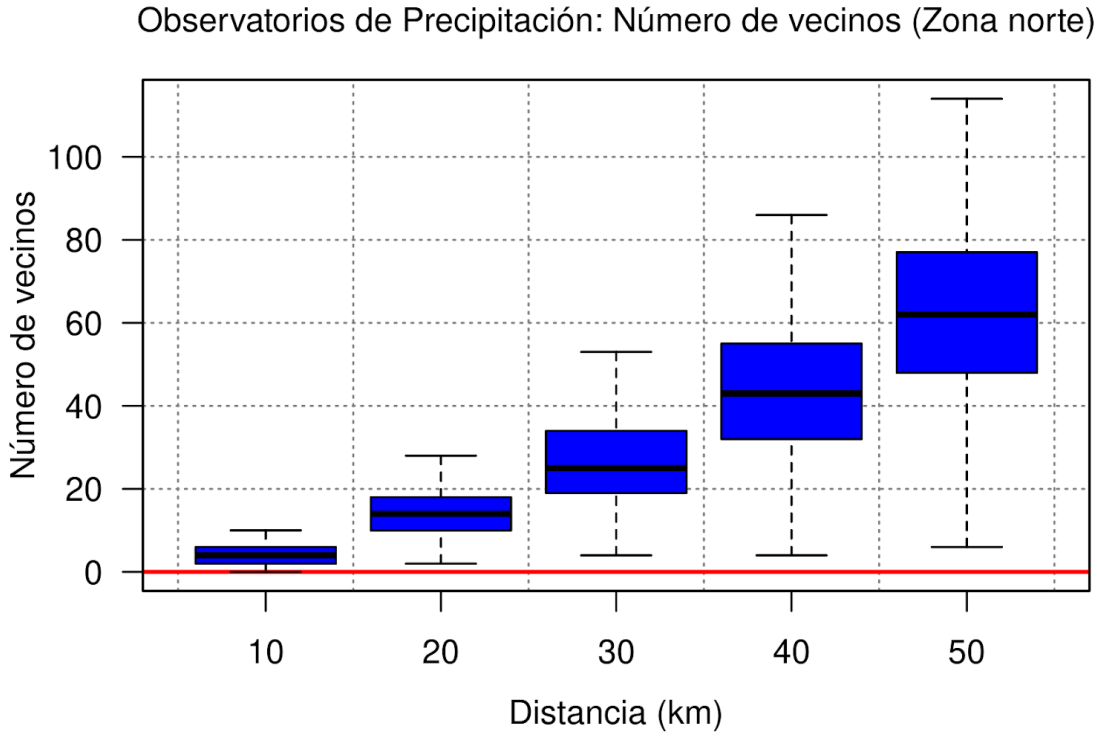


Figura X: Número de observatorios de precipitación vecinos en función de la distancia en Málaga



.

Figura X: Número de observatorios de precipitación vecinos en función de la distancia en Norte.

Como se puede observar, el número de vecinos cercanos es mucho mayor que en el caso de la temperatura (más del doble). En el caso de la precipitación, cada observatorio tiene una media de unos 10 observatorios vecinos a una distancia inferior a 10 km. Hay alguno, sobre todo en grandes áreas urbanas, que tiene hasta 25 observatorios vecinos, pero lo relevante sigue siendo en este caso que hay muchos que no tienen ninguno. Si se sube la distancia a 20km, se observa que el número medio de vecinos a una distancia inferior a esos 20km sube a 20 observatorios, y también se puede ver que, independientemente de la zona, todos los observatorios tienen al menos un vecino a una distancia inferior a esos 20km. Si la distancia se sube a 50 km, se observa que cada observatorio tiene una media de entre 60 y 100 observatorios vecinos en un radio inferior a esa distancia, observándose algunos cambios en función de la zona climática.

3.1.2. CORRELACIÓN

Al igual que en temperatura, a la hora de tomar la decisión sobre con qué observatorios rellenar la serie de un determinado observatorio, es imprescindible calcular la correlación de los datos entre observatorios. Para ello, se ha realizado un estudio para determinar la correlación entre observaciones de precipitación en función de su distancia. Para cada observatorio, se han calculado las distancias y la correlación respecto al resto de observatorios de la zona, y se ha representado el número total de pares de observatorios agrupados por distancias frente a la correlación obtenida:

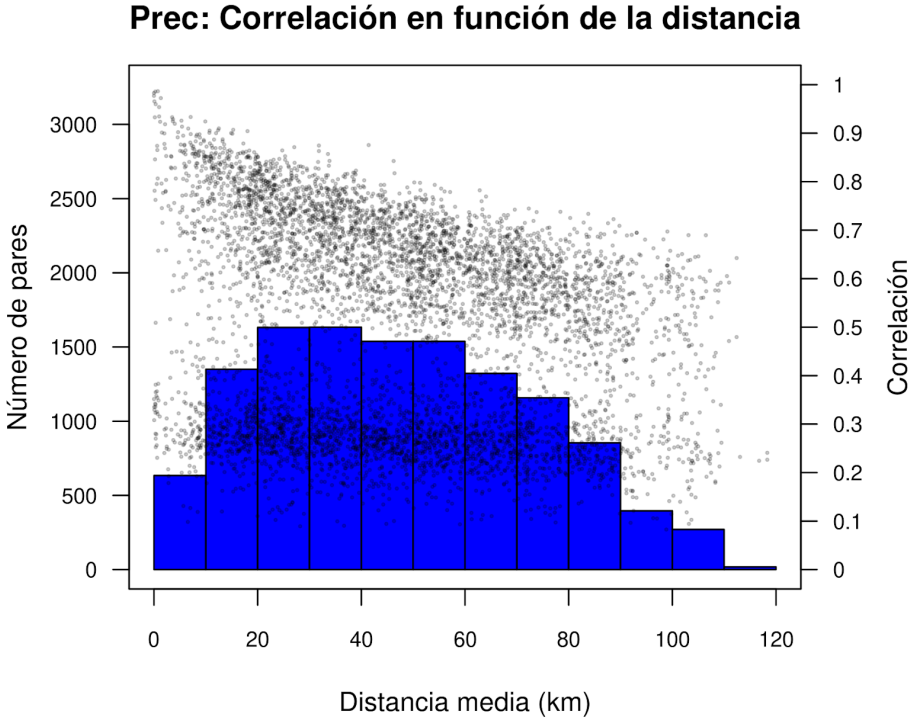


Figura 3.1.2.1: Correlación entre valores de distintos observatorios en función de su distancia para la precipitación en Cádiz.

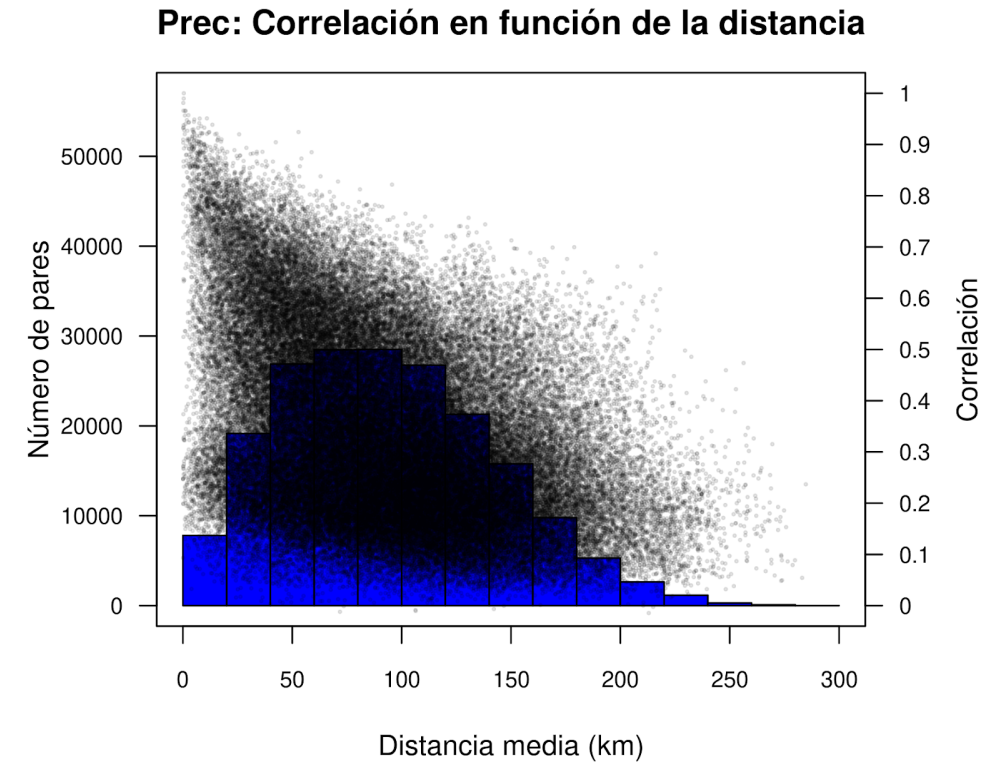


Figura 3.1.2.2: Correlación entre valores de distintos observatorios en función de su distancia para la precipitación en Este.

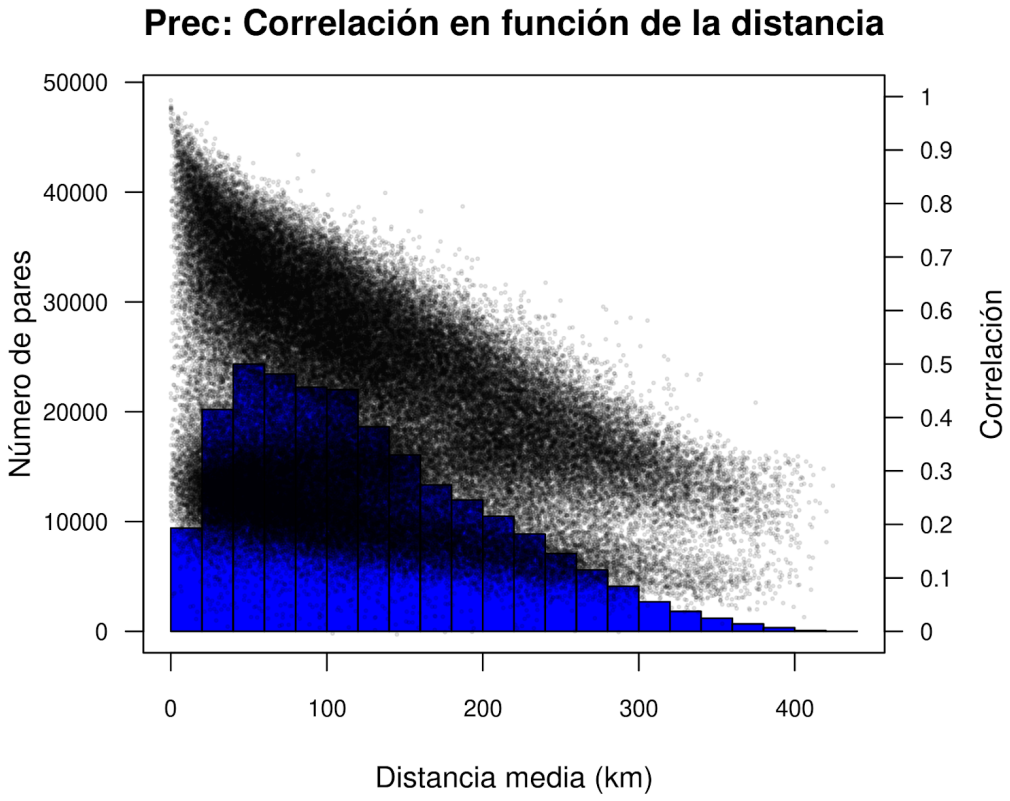


Figura 3.1.2.3: Correlación entre valores de distintos observatorios en función de su distancia para la precipitación en Guadalquivir.

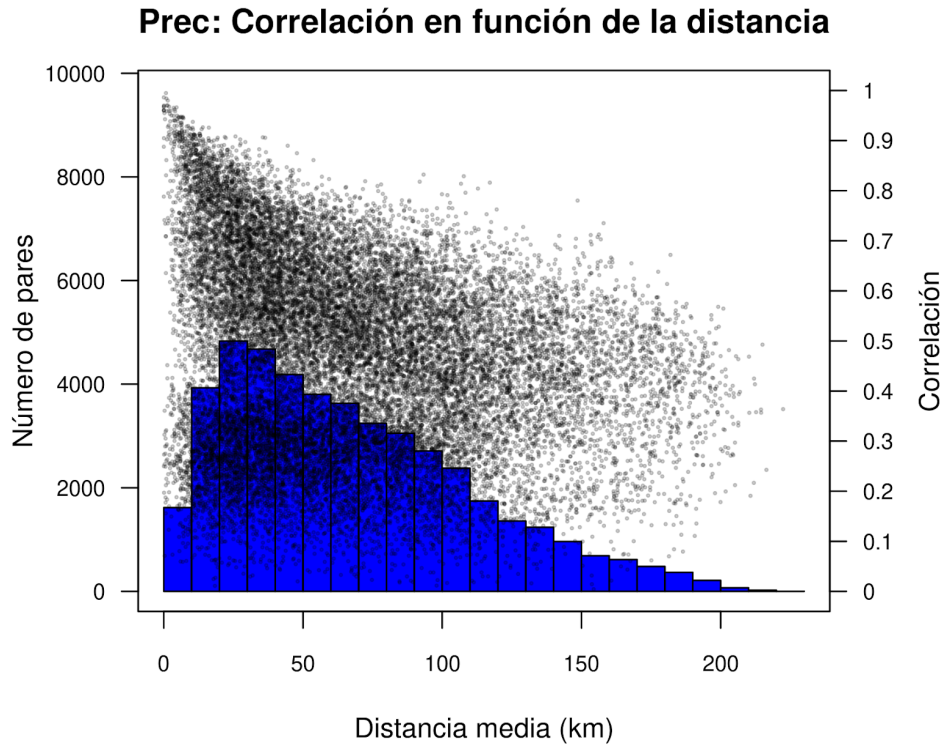


Figura 3.1.2.4: Correlación entre valores de distintos observatorios en función de su distancia para la precipitación en Málaga.

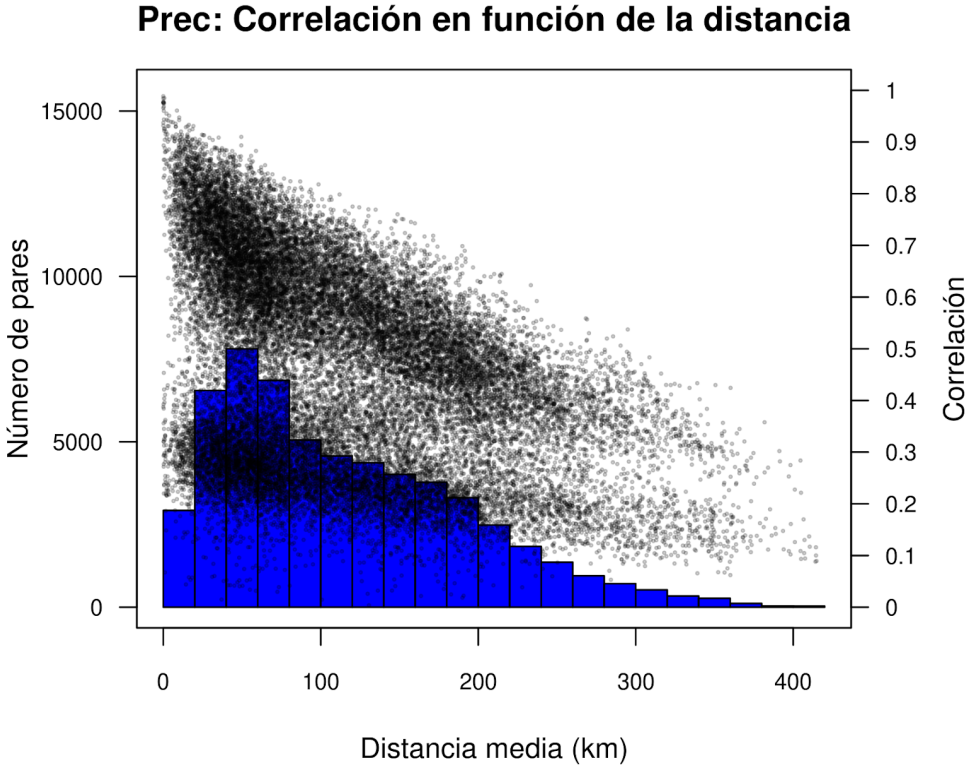


Figura 3.1.2.5: Correlación entre valores de distintos observatorios en función de su distancia para la precipitación en Norte.

En el caso de la precipitación, hay tres diferencias fundamentales respecto a lo que se observó en el caso de la temperatura: la correlación es menor, baja muy rápidamente con la distancia y se forman nubes de puntos diferenciados, o “clusters”.

En general, la correlación de las series de precipitación empieza siendo muy alta, alrededor de 0.9 en distancias muy cortas, pero se observa que baja muy rápidamente: en algunas zonas, a los 50 km se sitúa por debajo de 0.7 y a los 100 km por debajo de 0.5. Este resultado es coherente, ya que a una misma altitud y sin barreras orográficas, la temperatura puede cambiar muy poco pero la precipitación puede ser totalmente diferente, sobre todo en el caso de la precipitación convectiva.

En cuanto a los clusters dentro de una misma área geográfica, la posible explicación es la siguiente: en las zonas de Cádiz o Málaga hay dos distribuciones de precipitación totalmente diferentes: costa y montaña. Los observatorios cercanos a la costa van a tener entre ellos una correlación alta, lo mismo sucede con los observatorios de montaña cercanos entre sí. Estos puntos tenderán a concentrarse en una nube de puntos de correlación alta, configurando una nube. En cambio, la correlación entre un punto de costa respecto a un punto de montaña va a tener una correlación muy baja, y todos estos pares de puntos tenderán a concentrarse en una segunda nube de correlación baja. Lo mismo ocurre en la zona Norte con los observatorios que quedan a ambas vertientes de Sierra Morena, y en Guadalquivir con los observatorios que están cerca de las sierras de Cádiz y Málaga. En el caso de la zona Este es particular: su nube de puntos es muy extensa y dispersa debido a la irregularidad en las precipitaciones en toda zona y en la irregularidad del relieve que introducen las cordilleras Béticas.

3.1.3. METODOLOGÍA

Una vez analizados los resultados del número de vecinos cercanos y la correlación, se ha procedido a establecer la siguiente metodología para realizar el relleno de lagunas de precipitación (análoga al caso de la variable temperatura pero con disminución en el valor de correlación mínimo seleccionado):

1) Para cada observatorio, se seleccionan los observatorios situados a una distancia inferior a 20 km y que tengan una correlación superior a 0.5
2) En caso de que haya menos de 6 observatorios que cumplan estos requisitos, se ampliará de forma recursiva el radio de cercanía en 5km hasta que haya 6 observatorios cercanos.
3) De los 6 observatorios que cumplen los requisitos anteriores, se seleccionan los 3 que tengan mayor número de datos entre los años 1979 y 2020.
4) Se realiza una regresión lineal múltiple con los datos que tienen en común el observatorio que se quiere rellenar y sus 3 observatorios seleccionados anteriormente.
5) Con los parámetros de la regresión lineal, se rellenan los huecos del observatorio tipo que están vacíos y cuyos datos de los 3 observatorios seleccionados están completos.

No se rellenan huecos de un determinado día si cualquiera de los 3 observatorios con los que se ha realizado la regresión lineal no tienen datos para ese día.

6) Con 1 sola iteración, el proceso muestra que ya se han rellenado todos los observatorios para cada zona.

3.1.4. RESULTADOS

En primer lugar, se ha realizado una prueba para obtener el error medio al efectuar el relleno de lagunas con la metodología propuesta. Para ello, a cada serie a rellenar (observatorio tipo) se le ha eliminado un tercio de los datos y se le ha aplicado la metodología de relleno de lagunas para obtener esos mismos datos. A continuación, se ha comparado la serie original con la serie obtenida con datos rellenos y se han calculado parámetros para determinar su error: accuracy, BIAS, F1 score, MAE, percentiles 95 y 99 de la serie rellenada para compararlos con los de la serie observada, RMSE y el cociente de la desviación estándar de la serie rellenada entre la observada. Los resultados de esta prueba han sido los siguientes:

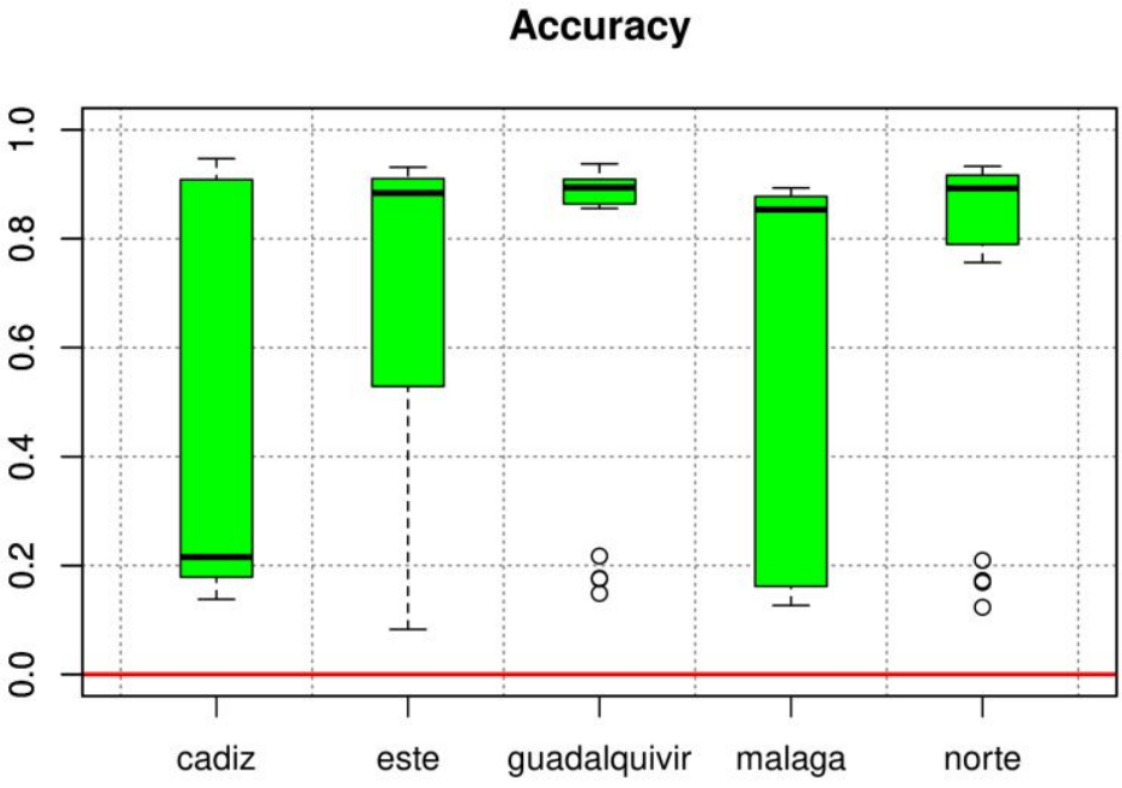


Figura 3.1.4.1: Accuracy de los datos rellenos de precipitación empleando el método 1 por zonas.

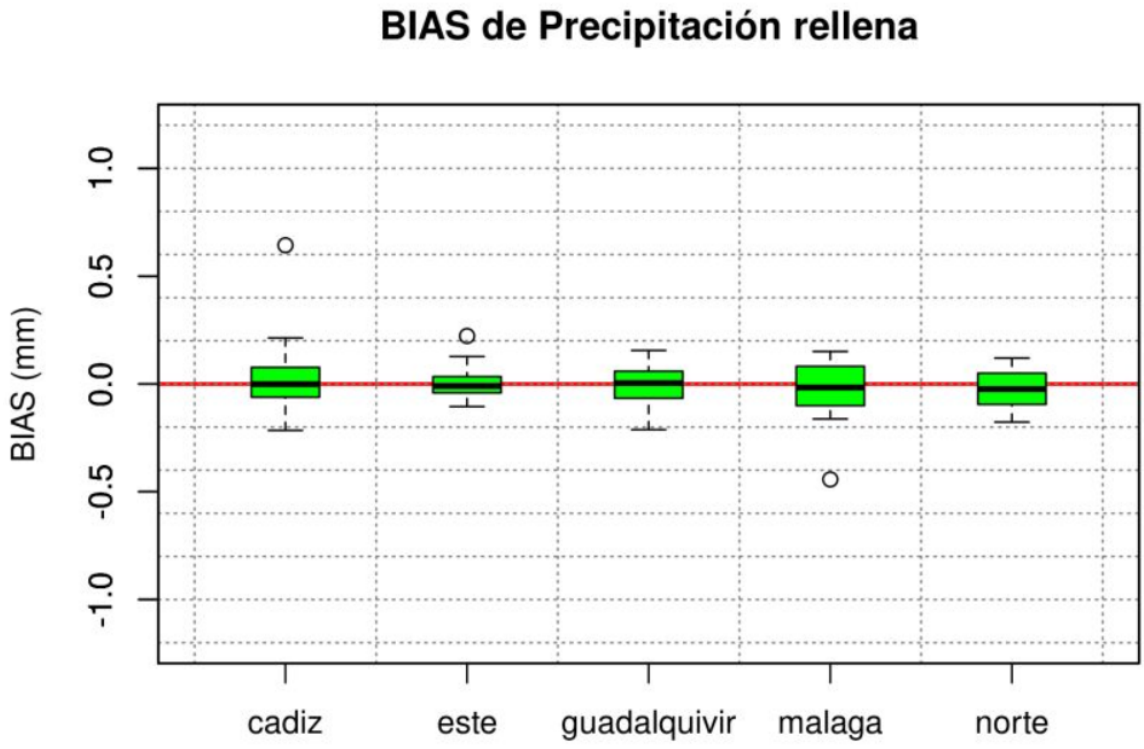


Figura 3.1.4.2: BIAS de los datos rellenos de precipitación empleando el método 1 por zonas.

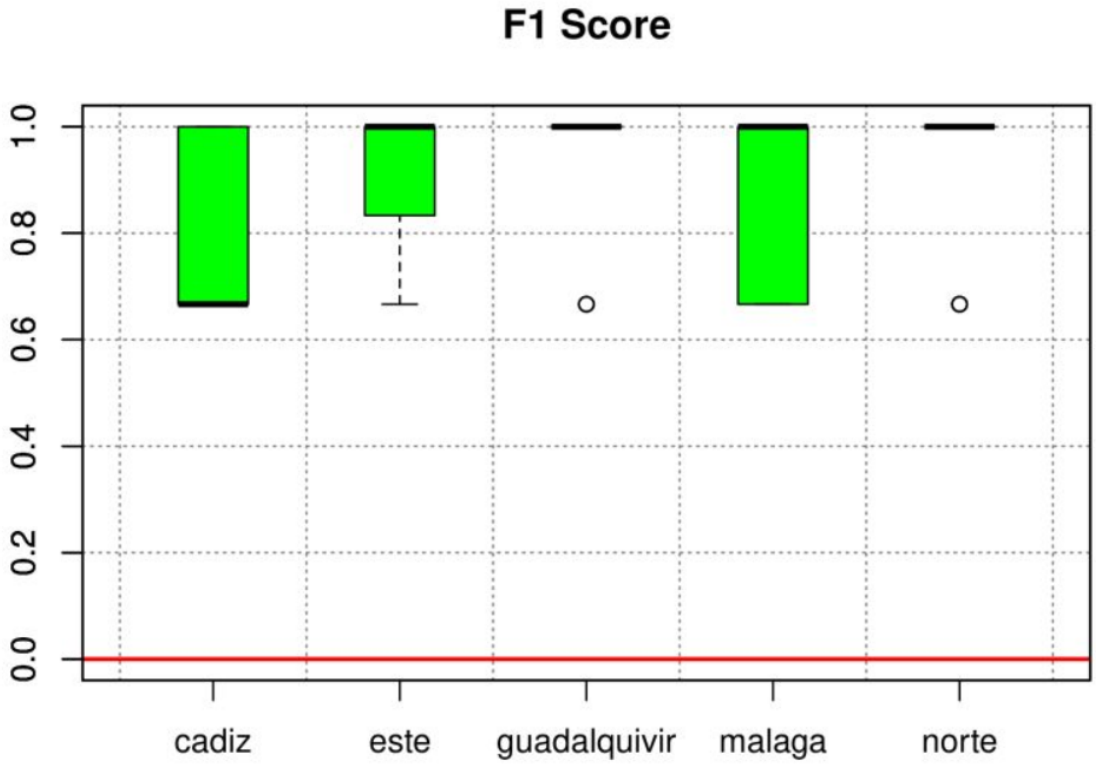


Figura 3.1.4.3: F1 Score de los datos rellenos de precipitación empleando el método 1 por zonas.

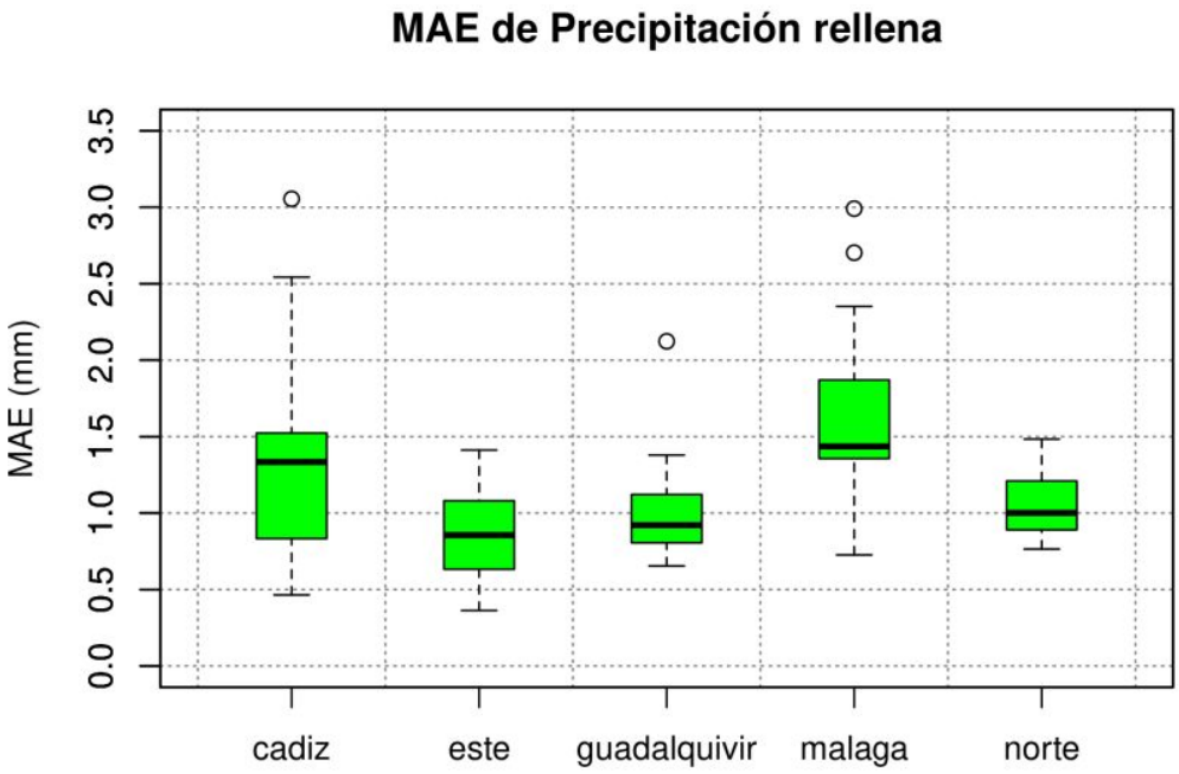


Figura 3.1.4.4: MAE de los datos rellenos de precipitación empleando el método 1 por zonas.

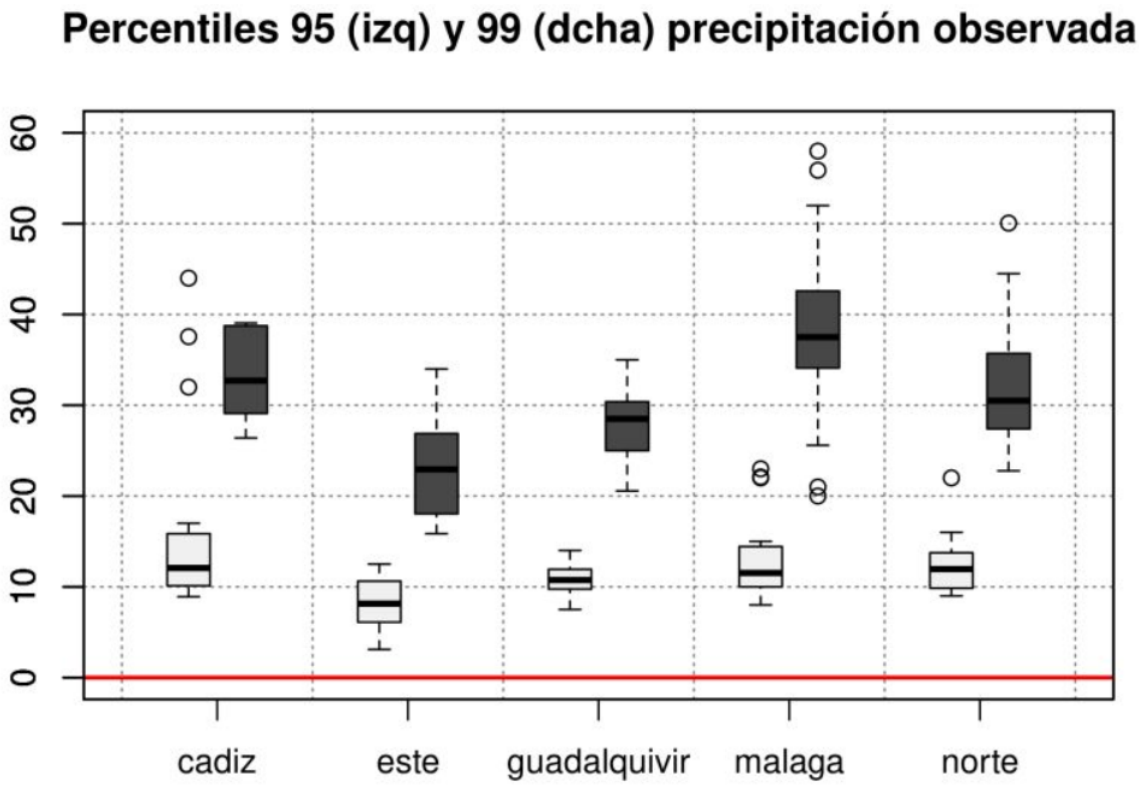


Figura 3.1.4.5: Percentiles 95 y 99 de los datos observados de precipitación por zonas.

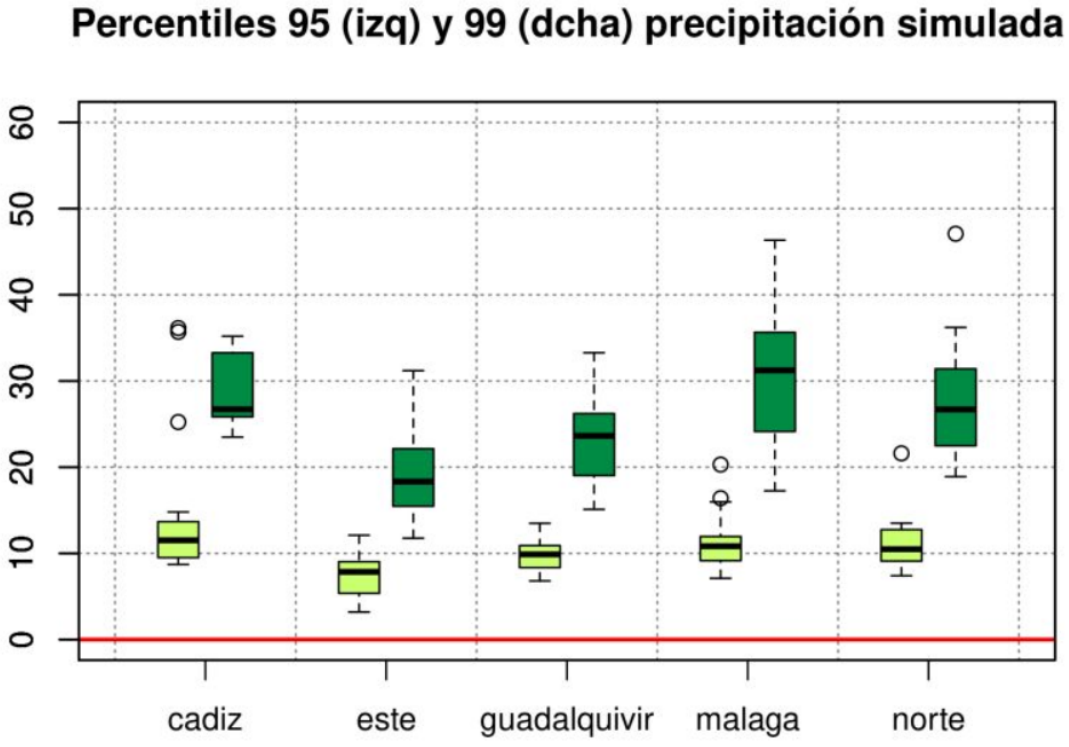


Figura 3.1.4.6: Percentiles 95 y 99 de los datos rellenos de precipitación empleando el método 1 por zonas.

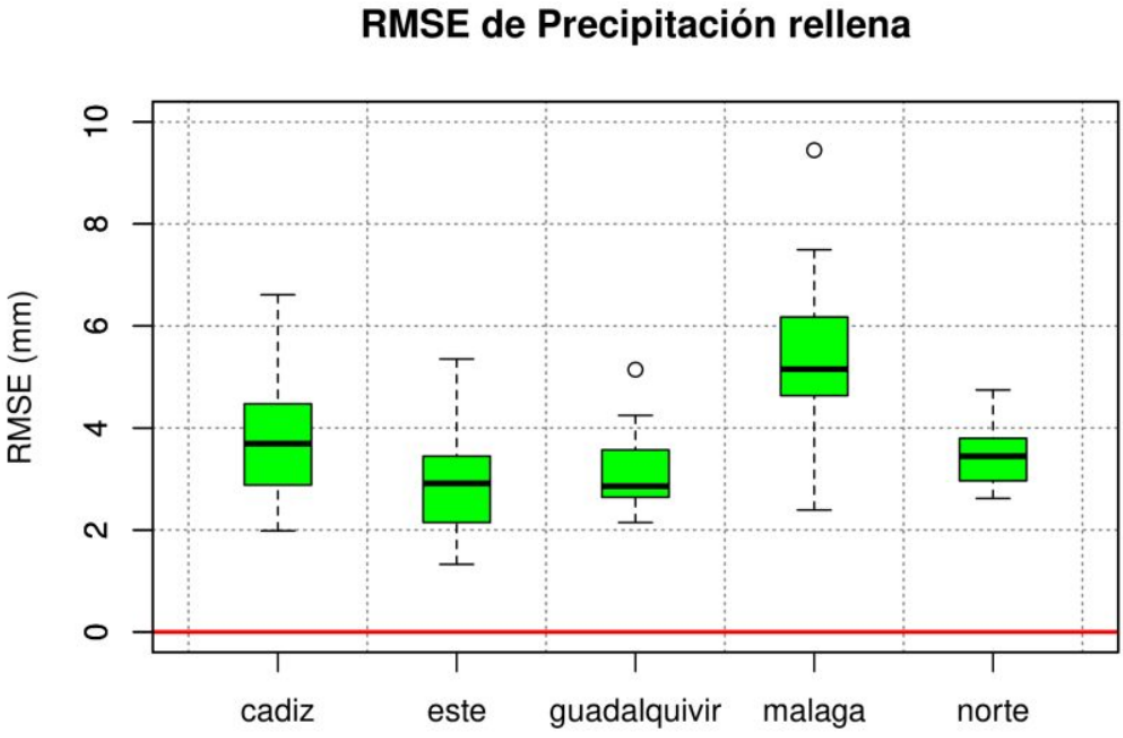


Figura 3.1.4.7: RMSE de los datos rellenos de precipitación empleando el método 1 por zonas.

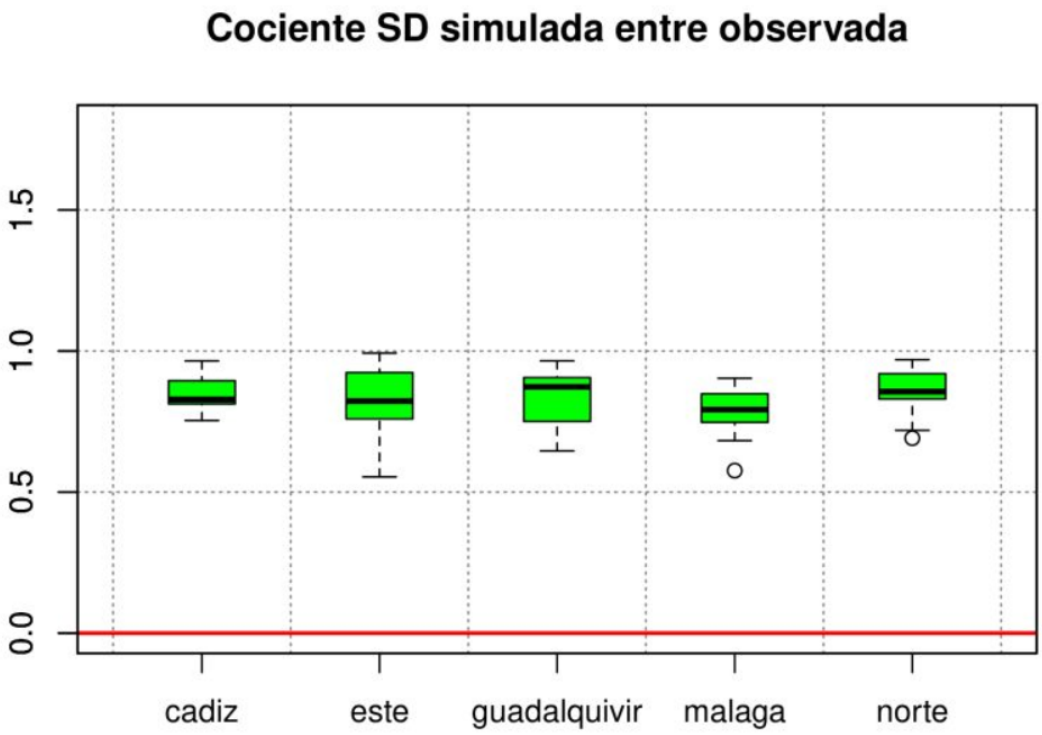


Figura 3.1.4.8: Cociente de la desviación estándar de los datos de precipitación rellenos entre los observados empleando el método 1 por zonas.

Los resultados son comparados en el apartado 3.3. con los obtenidos por el método 2, presentado en la sección 3.2.

3.2. MÉTODO 2: REDDPREC PACKAGE

El segundo método consiste en la aplicación de la función *gapFilling*, específicamente elaborada para el relleno de lagunas de datos diarios de precipitación y perteneciente al paquete *reddPrec* (Serrano-Notivoli et al., 2016) disponible para el software estadístico R.

3.2.1. RESULTADOS

De nuevo se ha realizado una prueba para obtener el error medio al efectuar el relleno de lagunas con esta metodología. Para ello, a cada serie a rellenar (observatorio tipo) se le ha eliminado un tercio de los datos y se le ha aplicado la metodología de relleno de lagunas para obtener esos mismos datos. A continuación, se ha comparado la serie original con la serie obtenida con datos rellenos y se han calculado parámetros para determinar su error: accuracy, BIAS, F1 score, MAE, percentiles 95 y 99 de la serie rellenada para compararlos con los de la serie observada, RMSE y el cociente de la desviación estándar de la serie rellenada entre la observada. Los resultados de esta prueba se pueden ver en las siguientes figuras.

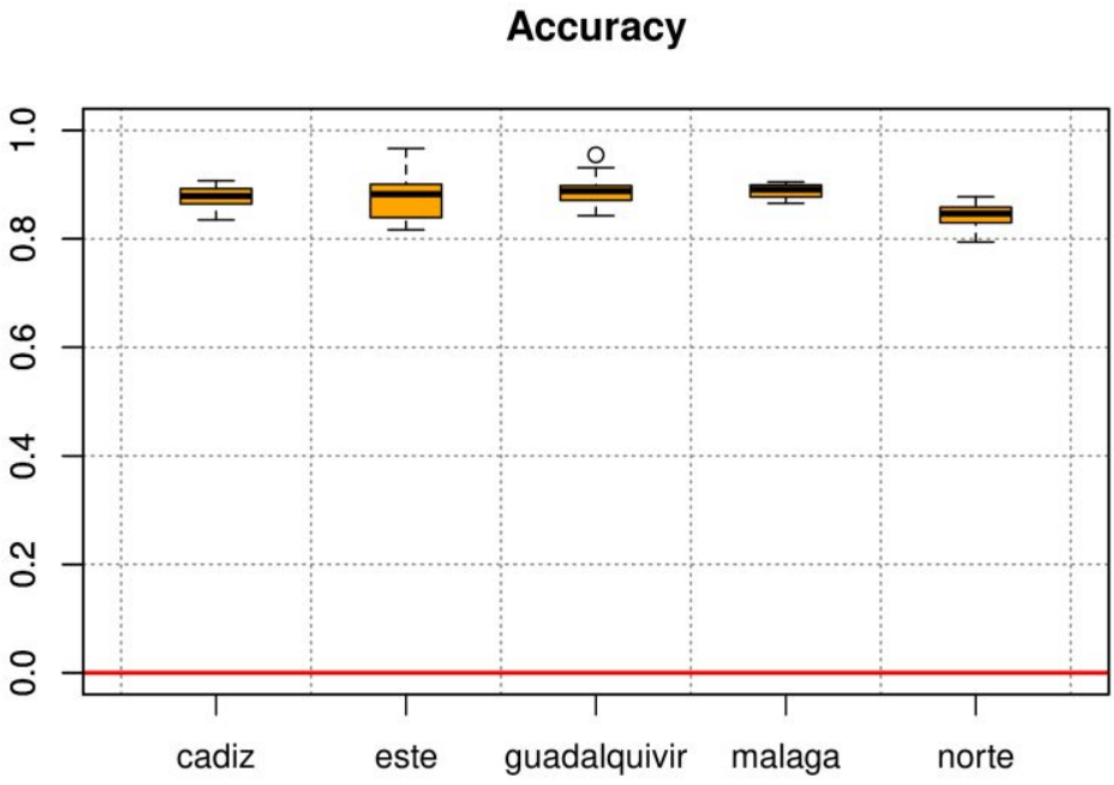


Figura 3.2.1.1: Accuracy de los datos de precipitación rellenos empleando el método 2 por zonas.

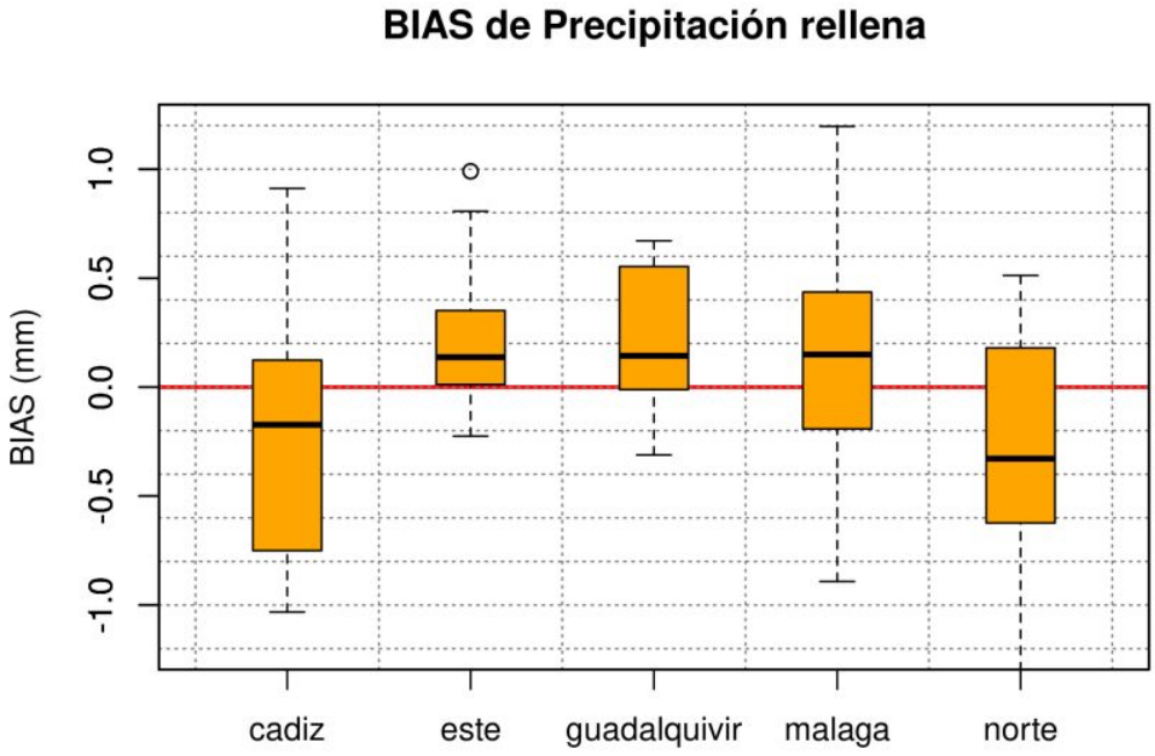


Figura 3.2.1.2: BIAS de los datos de precipitación rellenos empleando el método 2 por zonas.

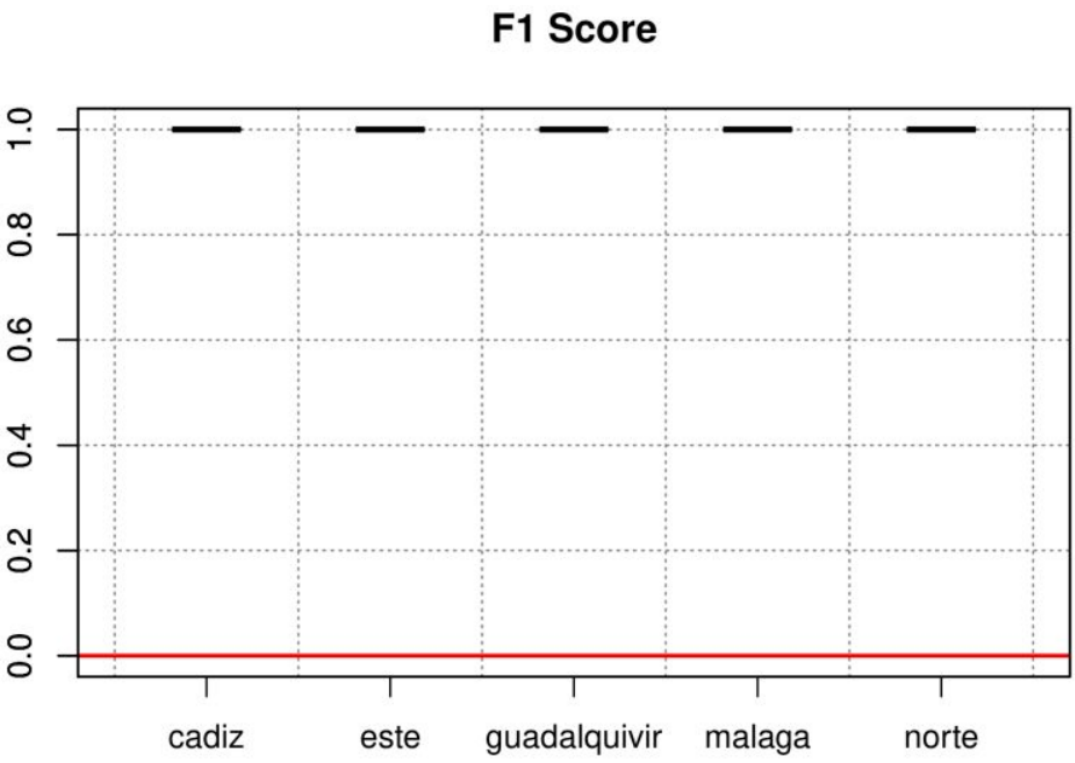


Figura 3.2.1.3: F1 Score de los datos de precipitación rellenos empleando el método 2 por zonas.

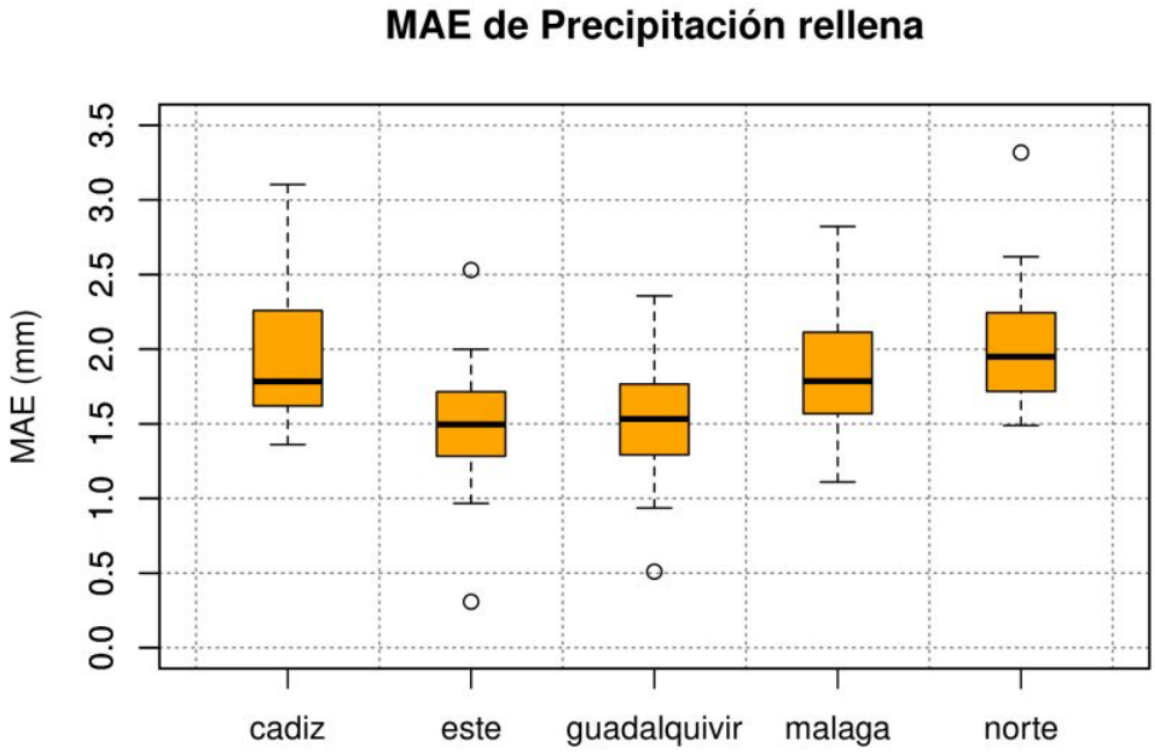


Figura 3.2.1.4: MAE de los datos de precipitación rellenos empleando el método 2 por zonas.

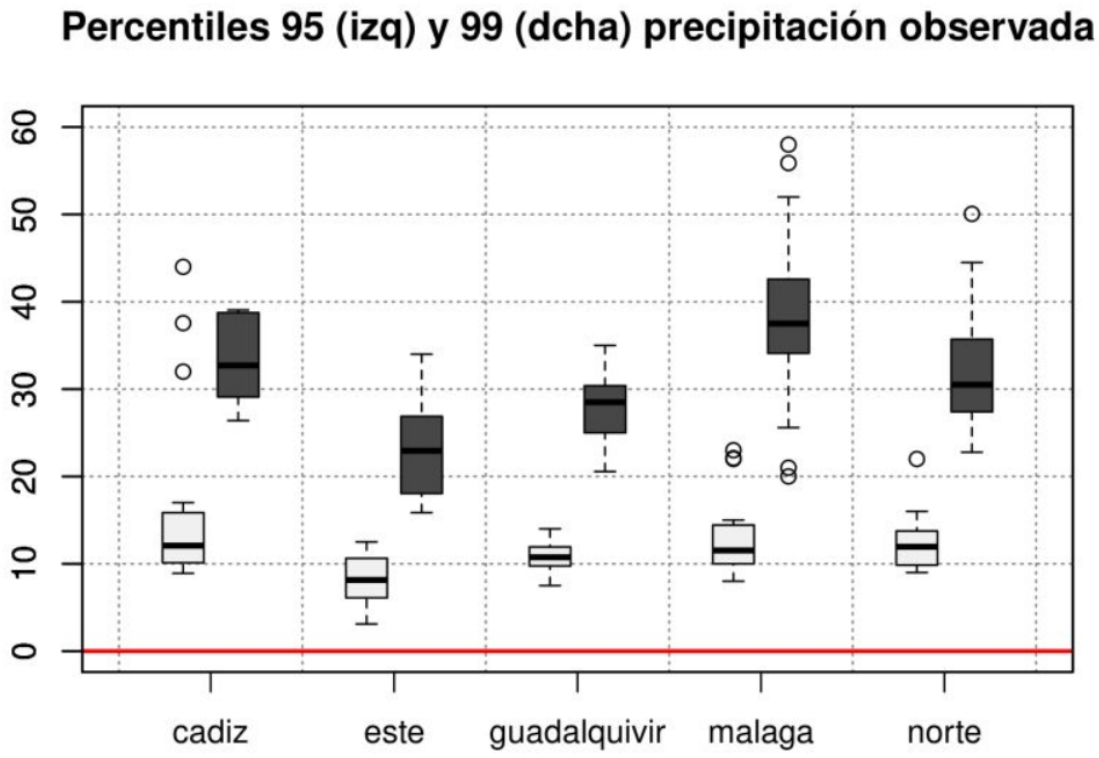


Figura 3.2.1.5: Percentiles 95 y 99 de los datos observados de precipitación por zonas.

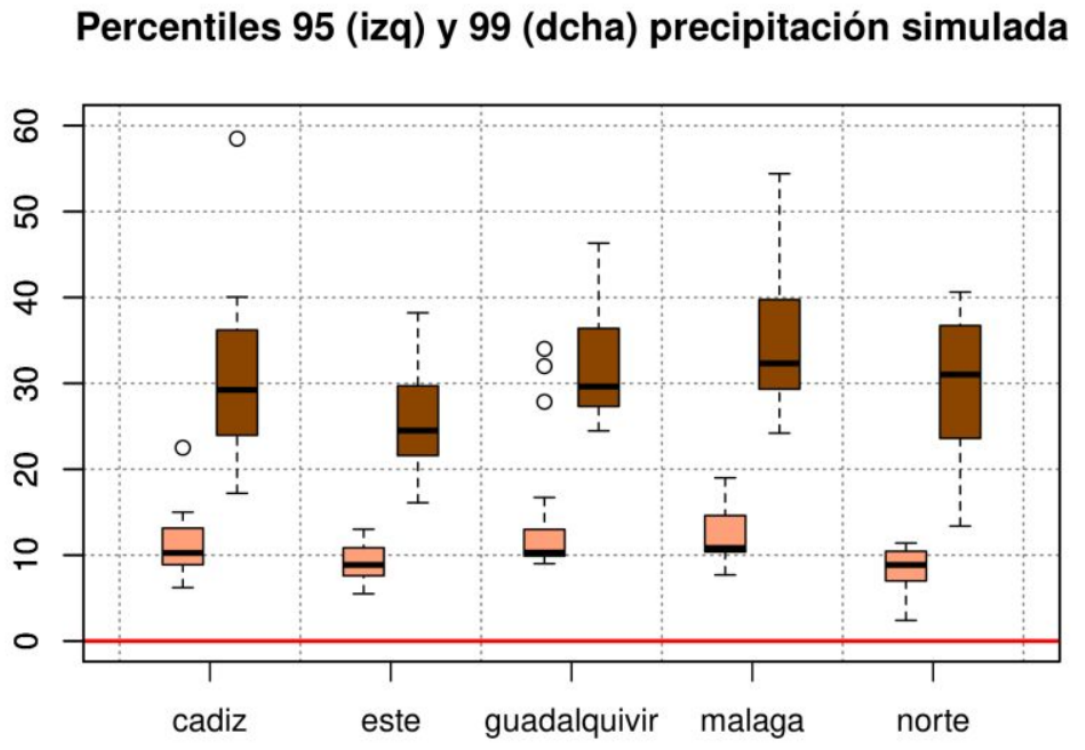


Figura 3.2.1.6: Percentiles 95 y 99 de los datos rellenos de precipitación empleando el método 2 por zonas.

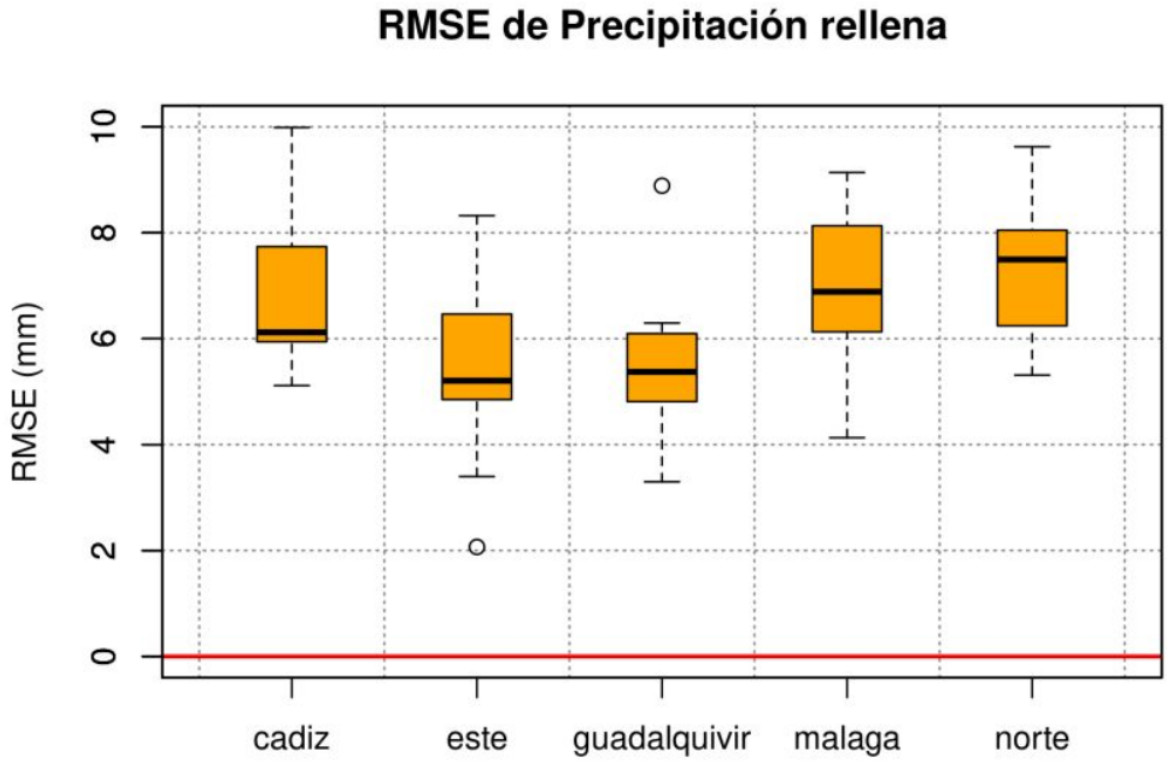


Figura 3.2.1.7: RMSE de los datos de precipitación rellenos empleando el método 2 por zonas.

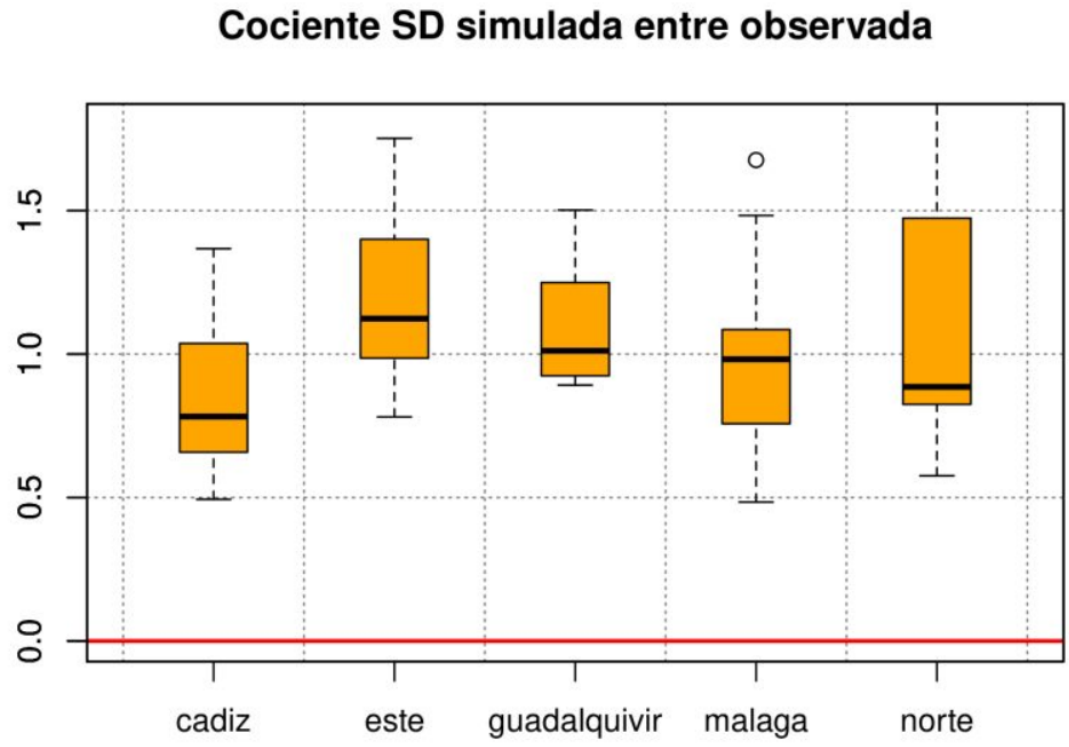


Figura 3.2.1.8: Cociente de la desviación estándar de los datos de precipitación rellenos entre los observados empleando el método 2 por zonas.

3.3. SELECCIÓN DEL MÉTODO

Comparando los resultados obtenidos por ambos métodos para la misma fuente de datos prueba, se obtiene que el método 1 obtiene mejores resultados en cuanto a MAE, BIAS, RMSE y tiende a reducir la desviación estándar con respecto de los valores originales; por su parte, el método 2 obtiene mejores resultados en cuanto a accuracy, F1 score, simulación de los percentiles 95 y 99 con respecto de la serie observada y no presenta una tendencia fija a reducir ni aumentar la desviación estándar con respecto de los valores originales sino que los resultados son diferentes para cada zona. Este resultado implica que empleando el método 1 se obtiene una simulación de los datos mejor en cuanto a magnitud, introduciendo un menor error, mientras que con el método 2 se obtiene un mayor acierto en cuanto a días de lluvia o no lluvia y en la representación de los valores extremos. Así, considerando los aspectos de la serie de precipitación diaria cuya representación se quiere priorizar al realizar el relleno de lagunas, se establece que el método 2 es el más adecuado.

A continuación, se ha calculado el porcentaje de huecos que hay en el periodo 1979-2020 tanto para los datos iniciales como para los datos rellenados por el método seleccionado:

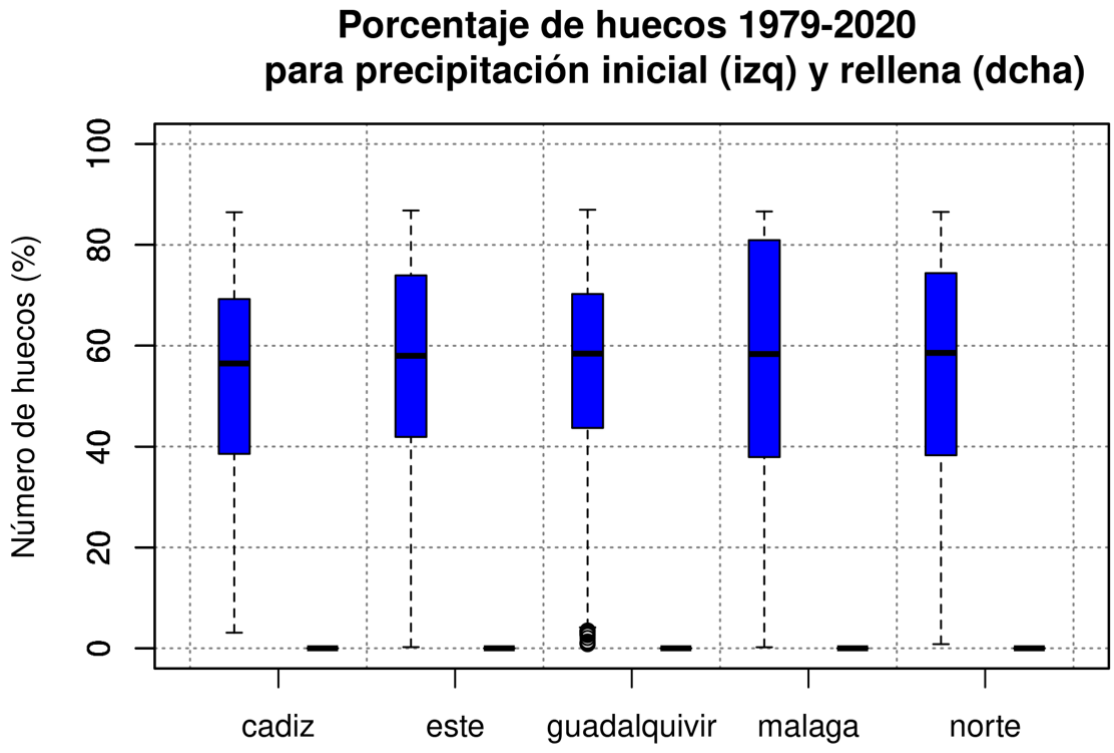


Figura 3.3.1: Porcentaje de huecos para la precipitación inicial (izquierda, azul) y rellena (derecha) para cada una de las cinco áreas climáticas estudiadas.

Inicialmente, se observa que la mediana del porcentaje de huecos se sitúa en torno al 58% en todas las zonas excepto en la zona de Cádiz, en la que la mediana es de 56.5%. Posteriormente, tras el relleno de huecos mediante el método 2 presentado anteriormente, se obtiene que el número de huecos es de 0 en todas las zonas y observatorios; por lo tanto, resulta un método muy exitoso en cuanto a relleno de lagunas.

REFERENCIAS

Serrano-Notivoli, R., de Luis, M., Beguería, S., An R package for daily precipitation climate series reconstruction, Environmental Modelling & Software (2016). http://dx.doi.org/10.1016/j.envsoft.2016.11.005