Experiencias en la aplicación operativa de un método multivariado de imputación de datos meteorológicos

Experiences in the Application of a Multivariate Method for Imputation of Meteorological Data

José Luis Araya-López[1]

Fecha de recepción: 7 de febrero del 2014
Fecha de aprobación: 25 de mayo del 2014

Araya-López, J. Experiencias en la aplicación operativa de un método multivariado de imputación de datos meteorológicos. *Tecnología en Marcha*. Vol. 27, N° 3, Julio-Setiembre 2014. Pág 70-79.

1 Meteorólogo, Instituto Meteorológico Nacional, Costa Rica. Teléfono: (506) 22225616 ext. 209. Correo electrónico: jlaraya@imn.ac.cr

Palabras clave

Imputación; datos faltantes; componentes principales; meteorología; climatología.

Resumen

Se han propuesto diferentes métodos y técnicas estadísticas para poder lidiar con el problema de los datos ausentes. En este estudio se analiza la aplicación de un método de imputación de datos usando componentes principales. El fin de este trabajo es discutir las posibilidades de este método para poder completar datos en resolución horaria generados por la red de estaciones del Instituto Meteorológico Nacional. Para lograr este fin se realizaron algunos experimentos con datos de prueba, a los que aleatoriamente se les eliminaron datos para su posterior estimación. Los resultados muestran que este método podría predecir la información que falta con un error absoluto medio de alrededor de 1 °C en la mayoría de los casos.

Key words

Imputation; missing data; principal components; meteorology; climatology.

Abstract

Different statistical methods and techniques have been proposed for dealing with missing data. This study discusses the application of the principal components approach for filling hourly meteorological data. In order to test the possibilities that this approach offers, preliminary tests were conducted by random removal of real data in time series. Missing data were predicted using a principal-components algorithm. The results show that this method could predict the missing information with an mean absolute error that is around 1°C in most of the cases.

Introducción

Es un hecho conocido por los profesionales que lidian con datos en servicios meteorológicos/hidrológicos que los sistemas de medición en ocasiones presentan problemas técnicos, los cuales pueden comprometer la integridad de los datos generados. Los datos faltantes son problemáticos, debido a que la mayoría de las métodos estadísticos no pueden ser aplicados de forma directa a datos incompletos (Josse y Hudson, 2012). En muchos casos, el mantenimiento adecuado y un protocolo detallado para el aseguramiento de la calidad pueden reducir al mínimo este tipo de problemas. Sin embargo, estas condiciones no siempre pueden garantizarse. En algunas redes de datos se puede tener que hacer frente a valores ausentes de forma sistemática. En tales circunstancias, puede que sea necesario tomar decisiones con respecto a los métodos que han de aplicarse para la estimación de los datos perdidos.

Diversos métodos han sido probados con anterioridad para resolver el problema de información faltante en datos meteorológicos. Alfaro y Pacheco (2000) presentaron un estudio en el que describen y prueban un compendio de métodos de relleno de datos generados por la red de estaciones del Instituto Meteorológico Nacional (IMN). En dicho estudio, los autores aplican métodos tales como el de regresión, de la razón, de la razón ajustada y de la razón-normal a datos anuales de precipitación (provenientes de estaciones meteorológicas convencionales). Posteriormente, Alfaro y Soley (2009) proponen, implementan y validan la metodología de imputación por componentes principales sobre un amplio conjunto de datos mensuales. El lector puede consultar Tabony (1983) para más detalles sobre el formalismo matemático que admite el método, así como el trabajo de Alfaro y Soley (2009).

La técnica de componentes principales (ACP) es ampliamente aplicada en climatología (aunque no exclusivamente al problema de imputación de datos), ya que es un método no paramétrico que ofrece amplias posibilidades en la interpretación de datos multivariantes. En el contexto de una red

de estaciones establecida como la del IMN, esta metodología resulta atractiva por el hecho de que se pueden generar datos estimados sobre un amplio conjunto de series de datos en una sola ejecución del método (en tanto estas series cumplan ciertos requisitos), de modo que si se cuenta con un conjunto de estaciones meteorológicas climáticamente similares es posible generar estimados de forma conjunta sobre los datos de las estaciones meteorológicas analizadas.

A nivel nacional, la técnica de análisis por componentes principales aplicada a la imputación de datos meteorológicos ha sido propuesta por Alfaro y Soley (2009). Aún cuando estos autores desarrollaron una serie de pruebas con datos reales usando datos mensuales, la intención de este análisis es evaluar que tan útil resulta la aproximación de imputación de datos por componentes principales en un contexto operacional, donde se cuenta con una cantidad considerable de estaciones meteorológicas que están generando datos continuamente, esto en la resolución de los datos generados actualmente por la red de estaciones meteorológicas automáticas del IMN. También resulta útil tener una estimación del error asociado a la imputación usando este método en diversas resoluciones de datos. Esto puede ser útil sobre todo si se piensa en las posibilidades de la metodología para la estimación de datos ausentes en el contexto de la automatización, en el cual los datos son recolectados en tiempo real. Además, permite ganar experiencia sobre su aplicabilidad y la prueba e implementación de herramientas computacionales adecuadas para poder lidiar con estos problemas.

Este trabajo se divide en tres partes. La primera etapa (la cual no se discute aquí pero resulta fundamental para su elaboración) incluyó un control preliminar con procedimientos similares a los utilizados por Araya (2007), de modo que los valores sospechosos pudieran ser detectados, analizados y validados. Esto es importante debido al efecto de los valores atípicos que pueden tener en la matriz de correlación cruzada (Peña, 2002). En la segunda parte se procede a efectuar pruebas sobre diversas regiones climáticas de Costa Rica, explicando en detalle el caso particular de un conjunto de datos de la región Pacífico Norte, con el fin de efectuar pruebas de imputación de datos y garantizar que no existieran datos faltantes en la muestra utilizada, lo cual permite definir una métrica del error de estimación y a su vez caracterizar la eficacia del método en esta resolución particular de datos. Esto mismo se realizó sobre diversos conjuntos de datos provenientes de diferentes regiones del país, lo cual se muestra en la parte final de este artículo.

Datos y métodos

Los datos utilizados fueron selecciones de un registro de información de los valores de temperatura por hora, que fueron generados por la red de datos del INM desde sus inicios. Con el fin de evaluar el desempeño del método, se seleccionaron periodos arbitrarios que presentan datos completos para todas las estaciones de cada grupo analizado, para poder efectuar una eliminación al azar de los datos existentes para su posterior imputación. Debido a la irregularidad en la distribución de las brechas en las series, fue difícil encontrar un periodo común para todo el conjunto de datos analizados. En el caso particular del conjunto de datos que se discute en este artículo, ese período corresponde a diciembre de 2007, siendo este uno de múltiples casos que fueron analizados en diversas regiones climáticas de Costa Rica. La figura 1 muestra un mapa de las distribuciones de las estaciones meteorológicas automáticas (EMA) del IMN, tanto abiertas como cerradas, y de las cuales se seleccionó el grupo indicado en el cuadro 1. El mapa muestra la variabilidad en términos de localización y altitud que muestra la red de estaciones del IMN, así como el número y distribución a lo largo de Costa Rica.

Con el fin de ilustrar la aplicación del método, se muestra en este artículo el caso de los datos de la región del Pacífico Norte de Costa Rica. Nótese que se han agrupado convenientemente para la realización de las pruebas de rellenado. Durante la selección de los diversos conjuntos de datos a los que se les aplicó este método de imputación, se consideró una serie de criterios que puedan garantizar cierta correspondencia climática entre los datos de las diversas series:

1. El conjunto de estaciones meteorológicas debe encontrarse en la misma región climática y/o tener cierta cercanía geográfica entre ellas.
2. Las estaciones deben tener una elevación similar.
3. Correlaciones altas entre los datos.

Figura 1. Distribución de las estaciones meteorológicas automáticas en Costa Rica. Los círculos de menor diámetro indican las estaciones de gran altitud, en tanto que los círculos de mayor diámetro muestran las estaciones en altitudes menores.

El problema de que los datos estén autocorrelacionados en resolución horaria y el efecto que esto tenga en sobreestimar los valores de la matriz de correlación cruzada (Soley y Alfaro, 2009) no se consideró a la hora de efectuar estos ensayos, ya que para este estudio había interés en explorar las características generales del método para estimar datos faltantes. La resolución de ese problema requiere un estudio enfocado en ello que se sale de los objetivos de este trabajo.

Se tuvo que realizar un trabajo de programación de la aplicación para efectuar los experimentos, en el cual se incorporó la función de relleno por componentes principales aportada por Alfaro y Soley (2009). Posteriormente, se seleccionaron diferentes periodos de datos para algunas ubicaciones. Finalmente, se llevó a cabo una serie de ensayos con los datos agrupados según se muestra en el cuadro 1. Estos ensayos requirieron una extracción aleatoria de los valores en los conjuntos de datos, usando una función generadora de números aleatorios que opera sobre los índices de filas y columnas de los datos. Una vez removidos, los valores faltantes artificiales se estimaron utilizando ACP. La cuantificación del error asociado en estas estimaciones es el Error Medio Absoluto (MAE, por sus siglas en inglés). Se ha preferido el uso de MAE sobre el error medio cuadrático debido a que este último tiene algunas características que pueden hacer su interpretación problemática, tales como que tiende a ser incrementalmente mayor que MAE conforme la distribución de la magnitud de los errores se hace más variable, aparte de que tiende a incrementarse más que MAE con $n^{1/2}$, ya que su límite inferior está fijado en MAE y su límite superior se incrementa con $n^{1/2}$ (Willmott y Matsuura, 2005). El error medio absoluto está definido por la relación

$$MAE = \left[n^{-1} \sum_{i=1}^{n} |e_i| \right] \qquad (1)$$

Si no se toma en cuenta el valor absoluto en (1), el error medio absoluto llega a ser lo que se conoce como Error de Sesgo Promedio (MBE, por sus siglas en inglés)

$$MBE = n^{-1} \sum_{i=1}^{n} e_i = \overline{p} - \overline{o} \qquad (2)$$

donde $\overline{p}$ corresponde al valor estimado y $\overline{p}$ es el valor observado, y es un indicador del sesgo

Cuadro 1. Localización de grupos de estaciones meteorológicas automáticas utilizadas en las pruebas de imputación.

Nombre de estación	Número de estación	Latitud	Longitud
La Linda, Perez Zeledón	98079	9,35	-83,63
Pindeco	98087	9,13	-83,33
Golfito	100003	8,63	-83,16
Los Patos	100625	8,60	-83,42
Coto 47	100631	8,60	-82,98
San José, Pinilla	72149	10,25	-85,83
Finca La Ceiba	72157	10,10	-85,32
Paquera	72159	9,82	-84,93
Aeropuerto Liberia	74051	10,58	-85,53
Santa Cruz	74053	10,28	-85,58
Hacienda Mojica	76055	10,67	-85,07
Museo Nacional	84000	9,93	-84,08
Cigefi	84139	9,93	-84,03
IMN, Aranjuez	84141	9,93	-84,07
Aeropuerto Pavas	84195	9,97	-84,13
Santa Bárbara	84197	10,03	-84,15
Canta Gallo	71015	10,48	-83,67
Aeropuerto Limón	81005	9,95	-83,02
Manzanillo	85023	9,63	-82,65
Hitoy Cerere	85021	9,67	-83,02
Sixaola	87013	9,52	-82,63
Comando Los Chiles	69633	11,02	-84,70
Finca Brasilia	69647	10,97	-85,33
Ciudad Quesada	69661	10,30	-84,42
Laguna, Caño Negro	69677	10,90	-84,78
Upala	69679	10,87	-85,07
La Rebusca	69681	10,48	-84,02

promedio del modelo utilizado. Luego se generaron algunas estadísticas antes y después del proceso de imputación de datos.

Discusión y resultados

El cuadro 1 muestra uno de los conjuntos de datos que se seleccionaron para los ensayos preliminares de este estudio, correspondiente al Pacífico Norte del país. La figura 2 muestra la matriz de correlación cruzada de los datos para el conjunto de ellos. Obsérvese que en general las entradas de esta matriz de correlación son siempre superiores a 0,8, llegando a 0,9 en algunos casos. Una buena correlación entre los datos de las estaciones es un requisito fundamental para la aplicación de este método (Alfaro y Soley, 2009).

Con el fin de visualizar qué tan cercanos resultan los valores estimados a los valores reales, la figura 3 muestra ambos conjuntos de datos contrastados en diagramas de dispersión estadística. La recta de mejor ajuste trazada es el resultado de una regresión

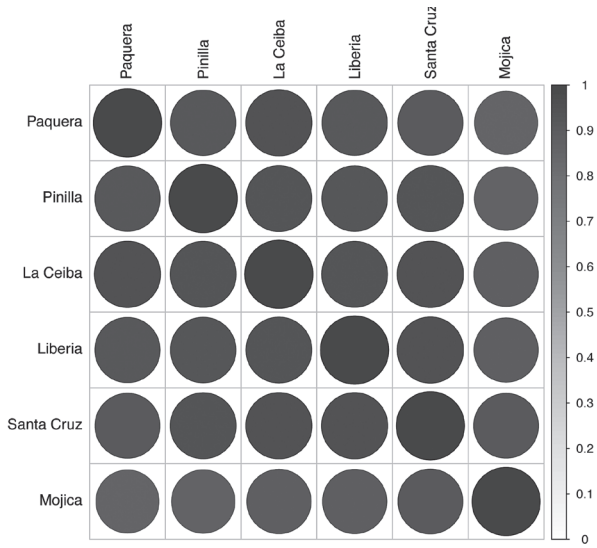

Figura 2. Visualización de la matriz de correlación cruzada de los datos analizados para diciembre de 2007 en las estaciones del Pacífico Norte.

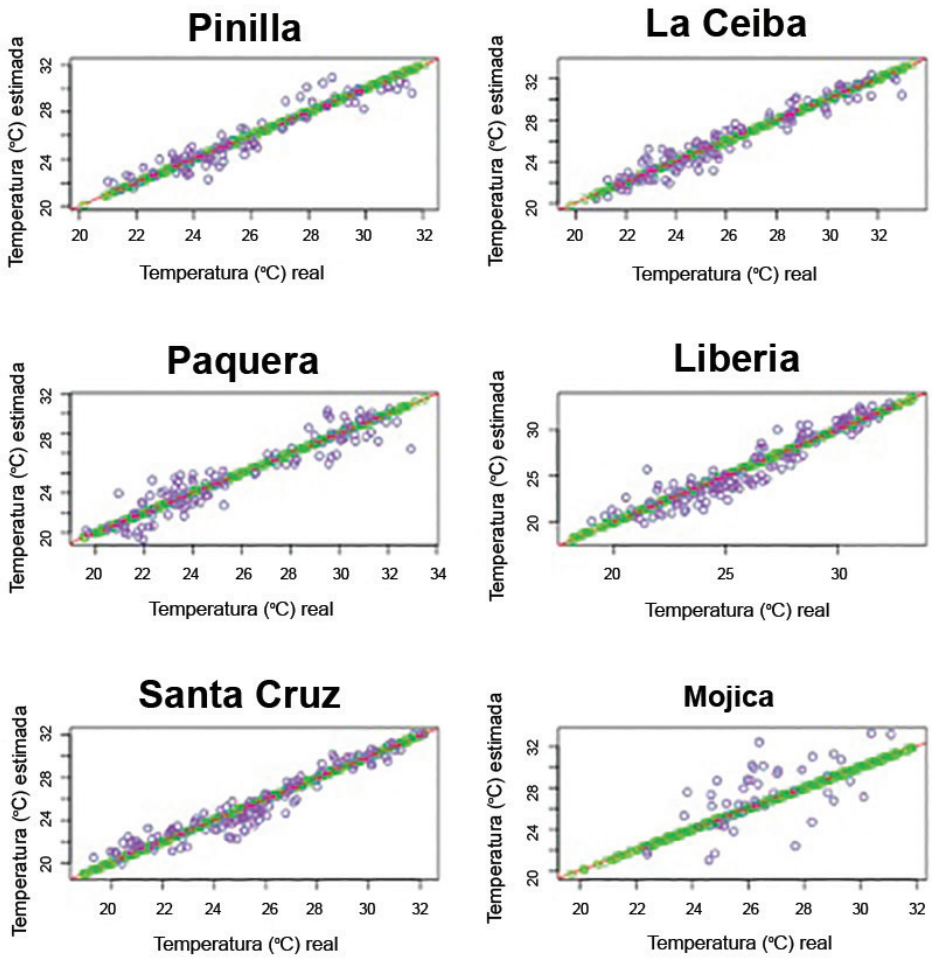

Figura 3. Diagramas de dispersión que muestran los datos estimados en función de los datos reales para diciembre de 2007 en las estaciones del Pacifico Norte.

entre los datos originales y los datos con datos sustituidos y rellenados. Los datos que no varían en ambas muestras se representan con círculos verdes, esto puede suceder porque los datos no fueron removidos o porque el método de relleno los estimó iguales. Los datos que difieren entre sí debido al proceso de imputación se muestran como círculos morados. Como puede notarse, esta representación tiene la ventaja de mostrar no solo los datos que no fueron aleatoriamente reemplazados por dato faltante, sino que también da una idea de cuánto se desvían los datos imputados con respecto al dato real. De las estaciones mostradas, puede observarse que Hacienda Mojica es la que tiende a mostrar mayor diferencia en las estimaciones.

Otra propiedad que debe tener un método de imputación adecuado es que los estadísticos de la serie se preserven a pesar de la estimación de los datos faltantes. En este sentido, la figura 4 muestra gráficos de cajas y bigotes (Wilks, 1995) para el conjunto de los datos de las seis estaciones, tanto antes como después del proceso de imputación. Puede observarse que, en general, la comparación entre medias, intervalos intercuartiles y valores extremos usando en los datos usando este criterio son muy similares durante los ensayos de sustitución y su posterior rellenado.

El procedimiento de evaluación descrito aquí se realizó sobre un conjunto de datos de estaciones meteorológicas ubicadas en diversas regiones del país. El cuadro 3 muestra los resultados para otros ensayos similares sobre otras regiones, así como los periodos comunes de los registros con datos completos utilizados. Las estaciones analizadas para cada región corresponden a las presentadas en el cuadro 1. Nótese que de cinco ensayos realizados a cada uno de los conjuntos de datos, la magnitud del error se mantiene inferior a 1 °C en la mayoría de los casos, a excepción de uno de los ensayos realizados en la región Caribe.

La figura 5 muestra la diferencia en grados Celsius entre los valores estimados y los valores reales. Puede observarse que, en general, estas diferencias tienden a centrarse en torno a diferencias cercanas a cero. Se observa también que existen casos en los cuales la diferencia puede ser de hasta 4 °C, aunque la frecuencia de estos valores es menor por encontrarse en los extremos de los histogramas. Dichos

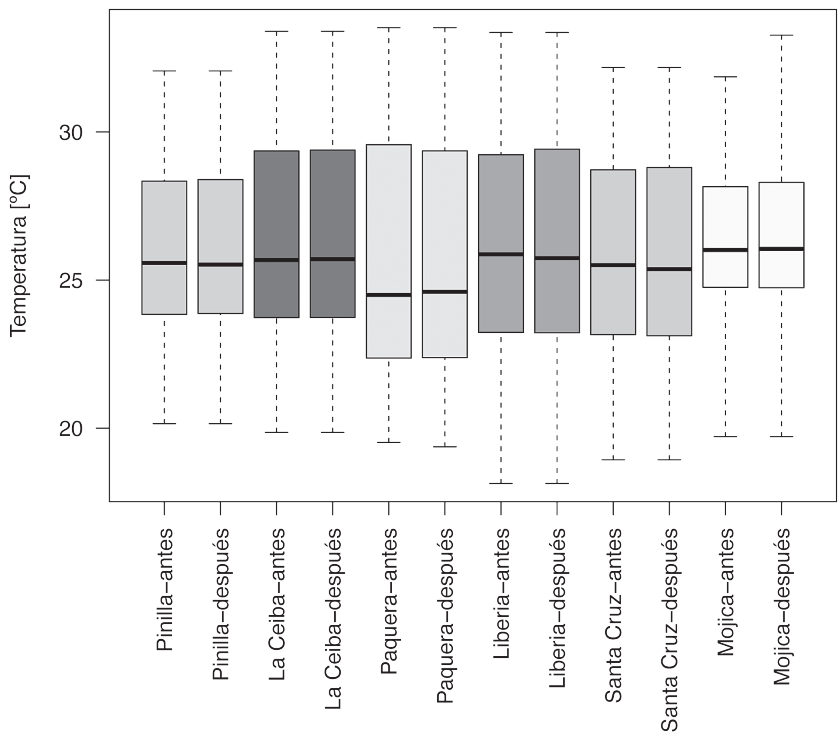

Figura 4. Conjunto de gráficos de cajas y bigotes para seis estaciones cercanas antes y después del proceso de imputación de datos.

Cuadro 3. Errores medios absolutos obtenidos para un conjunto de cinco ensayos de sustitución aleatoria sobre cinco sets de datos de estaciones ubicadas en diversas regiones climáticas de Costa Rica.

Región	Fechas		Errores Medios Absolutos (°C)				
Climática	Inicial	Final	1	2	3	4	5
Pacífico Sur	08/11/2007	31/12/2007	0,75	0,91	0,84	0,85	0,79
Pacífico Norte	02/12/2007	31/12/2007	0,88	0,89	0,82	0,91	0,88
Valle Central	14/06/2002	01/10/2002	0,48	0,86	0,49	0,53	0,52
Región Caribe	03/08/2007	14/08/2007	0,85	0,98	1,23	0,76	0,76
Zona Norte	08/11/2007	31/12/2007	0,74	0,92	0,65	0,71	0,72

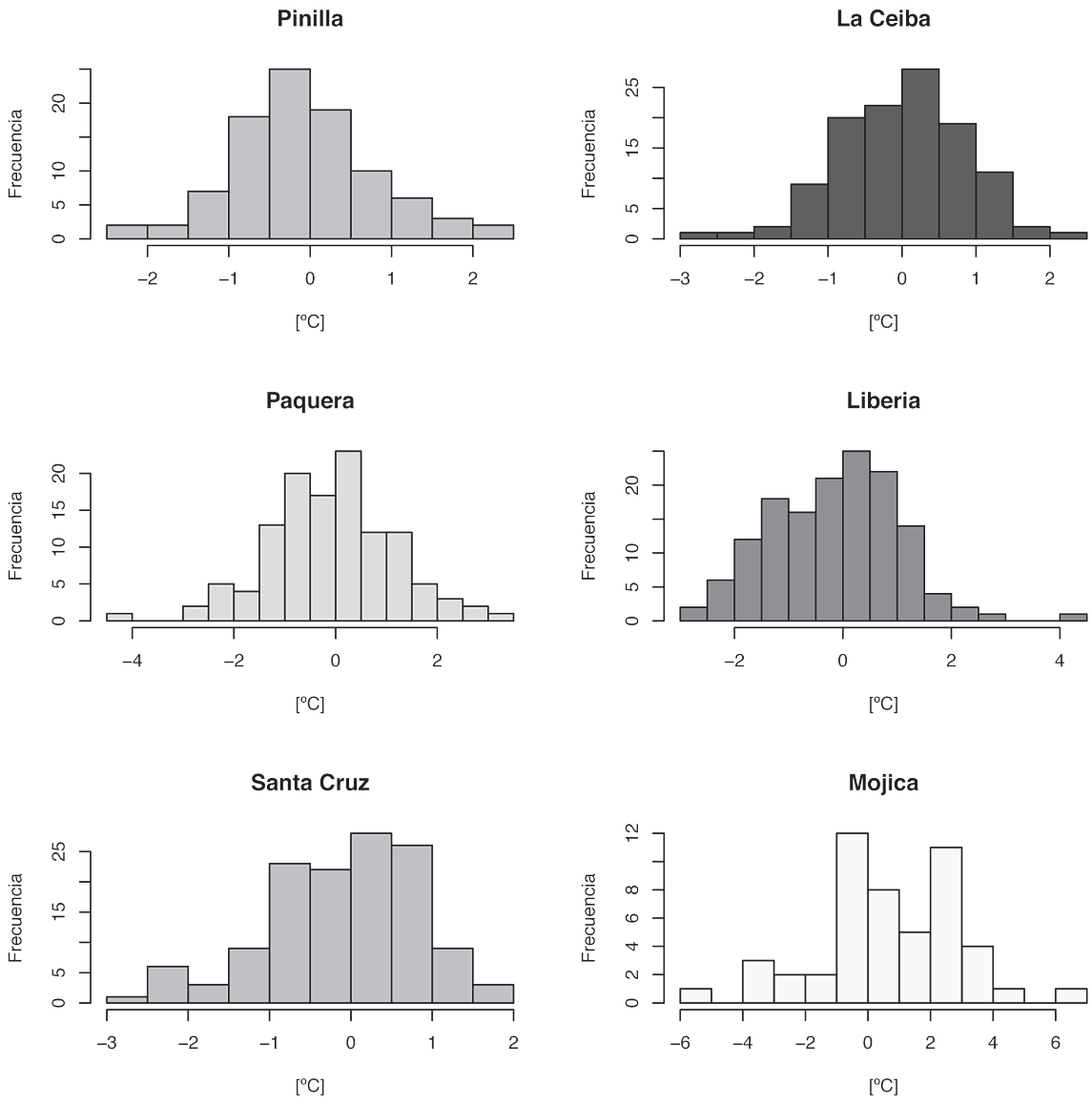

Figura 5. Diferencias entre valores reales y valores estimados.

puntos pueden deberse bien a un valor extremo, a la presencia de un valor atípico no detectado o a la incapacidad del método para describir ciertas condiciones particulares.

Conclusiones

La evaluación de los datos analizados indica que este método demuestra ser muy útil para completar la información cuando hay una buena densidad de EMA en la zona, que se reproducen en cierta medida las características climáticas de la región. Pueden ser útiles además consideraciones sobre la distribución y la representatividad de la EMA, así como la distribución de los datos atípicos en las series analizadas.

Una de las ventajas de ACP es que se conservan las propiedades estadísticas de los conjuntos de datos originales (Alfaro y Soley, 2009). Esto significa que se pueden generar reportes estadísticos basados en conjuntos de datos cuyos datos faltantes se han estimado utilizando CPA. Sin embargo, como explican los autores antes mencionados, corresponde a los usuarios definir si un enfoque basado en el ACP se puede implementar y tener éxito en un contexto operacional. En caso de un protocolo bien establecido para la aplicación de este método, resulta útil probar su funcionamiento para diferentes porcentajes de los datos ausentes y un número variable de componentes principales, aunque en los ensayos efectuados hasta el momento se notó que usando el criterio del gráfico Scree, dos componentes principales resultan suficientes para explicar el mayor porcentaje de la varianza. En particular, para el caso mostrado en este trabajo, al retener las primeras dos componentes principales, estas corresponden aproximadamente a una varianza acumulada del 85%.

En la práctica este método continúa teniendo una serie de limitaciones que son ineludibles, debido a la distribución de los datos faltantes en las series. Por ejemplo, el método presenta restricciones cuando se tienen datos ausentes simultáneamente en todas las series, una situación que se presentó con conjuntos de datos horarios usados en este trabajo, y que no corresponde en sí a una deficiencia del método. También debe tomarse en cuenta que las series temporales presentan errores correlacionados, las cuales son típicas de observaciones que se toman en puntos discretos en el tiempo, lo que se conoce como correlacion serial, tema que no fue no se trata en este documento pero que ciertamente debe considerarse siempre que se lidie con series de tiempo.

Cuando se correlacionan los términos de error de diferentes periodos, se dice que el término de error está correlacionado serialmente. La correlación serial se produce en los estudios de series de tiempo en que los errores asociados con un período de tiempo dado se trasladan en períodos futuros. Si bien se han desarrollado muchos métodos para tomar en cuenta este tipo de error en series temporales (James et al., 2013) y por lo tanto este factor debería ser tomado en cuenta en el proceso de imputación para poder llegar a mejores conclusiones.

Otro aspecto que debe tenerse en cuenta es que, si bien en este estudio las sustituciones se hicieron de forma aleatoria, en el caso de datos obtenidos en el quehacer operativo de una red, dicha aleatoriedad en la distribución de los datos faltantes en las series reales puede ser cuestionable en muchos casos, además de que los periodos de prueba se han confinado solamente a periodos comunes sin información ausente. La ausencia de estaciones cercanas es sin duda la situación que más puede invalidar la aplicabilidad del método a un nivel operativo, de allí que es importante que en el planeamiento de una red se tome en cuenta que las estaciones meteorológicas tengan estaciones climáticamente comparables, para poder lidiar con el problema de pérdida de datos cuando se presente.

Resulta conveniente contar con estaciones ubicadas de tal manera que permitan reconstruir hasta cierto punto lo que ha ocurrido en alguna de ellas en caso de que exista pérdida de datos, además de promover prácticas que tiendan a evitar la pérdida de datos. El incremento del número de estaciones en la red de estaciones sin duda alguna favorece la existencia de datos adicionales que permiten extender la aplicabilidad de este método. La recolección en tiempo real puede contribuir a minimizar este inconveniente, eliminando el periodo de incluso meses en el que periódicamente deben ser extraídos de la EMA, y con ello la posibilidad de que en ese tiempo algún problema técnico ocasione una pérdida irremediable de la información.

Bibliografía

Alfaro, R. & Pacheco, R. (2000). Aplicación de algunos métodos de relleno a series anuales de lluvia de diferentes regiones de Costa Rica. *Tópicos Meteorológicos y Oceanográficos, 7*(1):1-20,2000.

Alfaro, E. J. & Soley, F. J. (2009). Descripción de dos métodos de rellenado de datos ausentes en series de tiempo meteorológicas. *Revista de Matemática: Teoría y Aplicaciones, 16*(1): 60-75.

Araya, J. L. (2007). *Algoritmos de Control de Calidad de Datos en Estaciones Meteorológicas Automáticas*. Tesis de Licenciatura. Escuela de Física, Universidad de Costa Rica. 172 pp.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to Statistical Learning*. Springer Texts in Statistics. 426 pp.

Josse, J. & Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal da la Société Française de Statistique, 153*(2),79-99.

Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw Hill. 539 pp.

Tabony, R. C. (1983). The Estimation of Missing Climatological Data. *Journal of Climatology, 3,* 297-314.

Wilks, D. S. (1995). *Statistical Methods in Atmospheric Sciences*. Academic Press, 466 pp.

Willmott, C .J. & Matsura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*, 79-82.