Calidad de datos en series de tiempo

Un enfoque metodológico aplicado en una red meteorológica

Rovere, Santiago Pecina, Lucas Estrin, Hernán Aprea, Mariano

***Abstract*—El tratamiento de series de tiempo suele representar problemas particulares debido a la correlación latente en los datos respecto a la variable temporal. Este desafío no pasa desapercibido al momento de evaluar la calidad de los datos, en particular en redes de estaciones meteorológicas, donde emergen innumerables situaciones que pueden afectar a la correcta recolección y almacenamiento de la información.**

En este trabajo se plantea la problemática y se propone un enfoque metodológico para el procesamiento de series de tiempo, esto es la detección de anomalías y la imputación de datos faltantes. Por último se presenta, a efectos prácticos de disponer de datos de calidad, una segmentación de estaciones georeferenciadas a partir de sus observaciones climáticas.

***Index Terms*—calidad de datos, series de tiempo, sensor de temperatura, red meteorológica, STL, anomalías, imputación, clustering, regionalización**

I. Introducción

A. La calidad en datos meteorológicos

La observación geoespacial sostiene un crecimiento acelerado en los últimos años, producto de la constante innovación en materia de adquisición, transmisión y almacenamiento de grandes volúmenes de datos. Esto ha permitido la incesante proliferación de estaciones meteorológicas, aquellas responsables de la recolección de información relevante del entorno, lo cual representa un desafío interesante si se desea obtener un conjunto de datos consistentes, en particular, si se tratase de infraestructuras heterogéneas distribuidas en extensos territorios.

La fiabilidad de las mediciones se transforma, por lo tanto, en una necesidad crucial para lograr el desarrollo de modelos exitosos que permitan no sólo dar respuestas a análisis descriptivos, sino que también provean herramientas predictivas para dar apoyo al proceso de toma de decisión, en lo que respecta a problemas ecológicos y ambientales. En consecuencia, la calidad de los datos surge como una propiedad relevante, o más bien determinante, y debe ser dirigida en etapas tempranas de todo proyecto que persiga proporcionar cierta interpretación sobre los datos recopilados [1].

En búsqueda de una clasificación generalizada, Woodall et al. [2] ofrecen una amplia variedad de motivos que derivan en la degradación de la calidad de los datos e introducen respectivos métodos para su tratamiento apropiado. En particular, las observaciones meteorológicas pueden desembocar en mediciones imprecisas en diferentes fases de su ciclo de vida [3], abarcando desde la instalación y mantenimiento de sensores hasta el almacenamiento final de las mediciones obtenidas.

No obstante, en base al dominio de estudio, el cual es también analizado por Gitzel et al. [4], y sujeto al alcance de este trabajo, se observan tres problemáticas predominantes: la ausencia de datos, errores de carga y la inconsistencia en las mediciones. Comprendiendo que las causas asociadas son fruto de diversos factores, entre los cuales se encuadran errores en la transmisión, calibración inadecuada de sensores, variabilidad en la precisión según el fabricante o errores de carga [5], se excluye de esta labor el análisis extendido sobre dichas fuentes para orientar la investigación en la detección y resolución de los problemas mencionados que tienen efecto sobre la calidad de los datos.

Considerando que las mediciones provistas por este sistema mantienen una estrecha relación con el instante en el cual fueron obtenidas, se conforman así diversas series de tiempo sujetas al sondeo propio de cada sensor en cada una de las estaciones, sugiriendo entonces una distinción especial frente a conjuntos de datos tradicionales en la medida que se debe contemplar la variable temporal asociada. Esta apreciación no resulta trivial al momento de trabajar sobre la calidad de los datos mencionada, y se añade el concepto espacial al servirse de información geolocalizada, aumentando en cierto grado la complejidad del preprocesamiento adecuado al tratarse de un problema espacio-temporal.

Sin embargo, la problemática planteada merece ser atendida, en un primer término, mediante la validación supervisada basada en el conocimiento experto del dominio, cuestión clave en el aporte de calidad de los datos. Moraru et al. [6] exponen en este marco a la auditoría humana como punto de partida para la generación de lo que denominan un *dataset* aumentado, aunque habitualmente esta actividad resulta inviable acorde a los recursos requeridos, por lo cual, en una segunda instancia, los datos son sometidos a tareas respaldadas en modelos estadísticos y aprendizaje no supervisado, siendo estas últimas el foco del desarrollo realizado aquí.

Siguiendo un enfoque más ambicioso, Rahman et al. [7] proponen un *framework* destinado a evaluar la calidad de datos mediante la construcción de un clasificador ensamblado, logrando un control completamente automatizado. El avance en este tipo de modelos es esperanzador y posiblemente se asienten como firmes candidatos a reemplazar métodos manuales en un futuro cercano. Aún así, es importante ampararse en técnicas y modelos que puedan ofrecer una mejor interpretabilidad en cómo se determinan dichos controles.

B. Trabajo a desarrollar

El presente estudio tiene como principal objetivo ofrecer una metodología de trabajo que permita, contemplando el contexto espacio-temporal, tratar las problemáticas puntuales de imputación de datos faltantes y la detección de anomalías en series de tiempo univariadas, sean producto de errores de carga o de sesgos de calibración. Para el primer caso, se explora una forma de interpolación incorporando correlación geoespacial [8]. En el segundo caso, algunos autores proponen técnicas multivariadas para la detección de anomalías, por ejemplo, basadas en el cálculo de la distancia de Mahalanobis y la transformación log-ratio de las series [9], o bien, bajo la observación de múltiples fuentes y métodos geoestadísticos [10].

Una publicación reciente [11], enfocada en la detección automática de anomalías en la infraestructura en la nube, introduce una estrategia interesante que combina la descomposición de señales en sus componentes de tendencia, estacionalidad y ruido restante [12], con el uso de indicadores estadísticos robustos, colocándola como un buen candidato para su evaluación en la aplicación a redes de sensores.

En un análisis complementario, y con la intención de poner de manifiesto la utilidad de contar con datos de calidad, se plantea la regionalización de los dispositivos de la red a partir de sus mediciones. Esto es, determinar, mediante la utilización de métodos de agrupación no supervisada, si es posible agrupar estaciones meteorológicas con un patrón que presente cierta cohesión en el espacio [13].

El documento se presenta en tres secciones principales que describen la metodología propuesta para cada uno de los objetivos planteados, siendo la primera dedicada a la detección de anomalías, seguida de la imputación de datos faltantes y concluyendo con la regionalización de las estaciones en base a series de temperaturas máximas. Por último, se exponen reflexiones relacionadas al proceso completo y los resultados obtenidos.

C. Área de estudio y materiales

El caso de estudio actual se desprende de un análisis sobre registros de temperatura máxima provistos por estaciones meteorológicas no automáticas dentro de la red del Servicio Meteorológico Nacional de la República Argentina [1], comprendiendo la cobertura completa del territorio nacional, los cuales se encuentran publicados desde el año 1961. Con el propósito de acotar el período a trabajar y así partir de observaciones estables, se han seleccionado para las tareas propuestas en este trabajo un conjunto de estaciones representativas en lo que respecta a la extensión territorial, abarcando las mediciones obtenidas entre los años 2001 y 2010.

El desarrollo en su totalidad se implementa a través del lenguaje R e incluyendo diversas librerías, entre las cuales se destacan STL[2] que permite la descomposición de series de forma robusta y TSclust[3] para el cómputo de agrupaciones. En relación a la visualización interactiva donde se presentan resultados (Anexo A[4]), la herramienta utilizada es Shiny Applications Online[5].

[1] API SMN: https://api.crc-sas.org/ws-api/

[2] Paquete R STLplus - https://cran.r-project.org/package=stlplus

[3] R Package for Time Series Clustering - http://www.jstatsoft.org/v62/i01/

II. Detección de anomalías

La detección de anomalías es el proceso por el cual se logran identificar ciertas observaciones que se desvían del comportamiento normal del conjunto de datos del cual provienen, o bien aquellas que no se ajustan a un patrón esperado. En línea con lo introducido en la sección previa, la importancia de poder reconocerlos radica en el hecho de que, frecuentemente, la causa de dicho dato atípico es un error en su proceso de generación, lo cual puede resultar en grandes pérdidas, tanto financieras como de reputación. A través de los años se han desarrollado múltiples técnicas en distintas áreas del conocimiento para atacar este problema, entre las cuales se encuentran la econometría, procesamiento de señales y estadística.

Detectar anomalías en series temporales no resulta un problema trivial. La utilización de métodos clásicos no suelen funcionar de manera óptima para este tipo de datos debido a las propiedades que habitualmente expresan las series en la vida real. Considerando el ejemplo de la regla de 3 sigmas (σ), la cual identifica como atípicos los valores ubicados a distancias mayores a 3 veces el desvío estándar desde la media, se identifica un problema recurrente: estas técnicas convencionales sólo tienen el potencial de descubrir outliers globales pero no aquellos estacionales. En la Fig. 1 se observa la aplicación de la regla de 3 sigmas en una serie temporal correspondiente a la temperatura máxima diaria capturada entre los años 2016 y 2019 en la estación del Aeroparque de Buenos Aires, a la que se le agregaron 3 anomalías de forma artificial. Las líneas horizontales azules demarcan los 3 desvíos absolutos, de manera que aquellos puntos que las superen, serán clasificados como outliers. Tal como se observa,

[4] Anexo A: Visualización interactiva

[5] Shiny Applications Online - https://www.shinyapps.io/

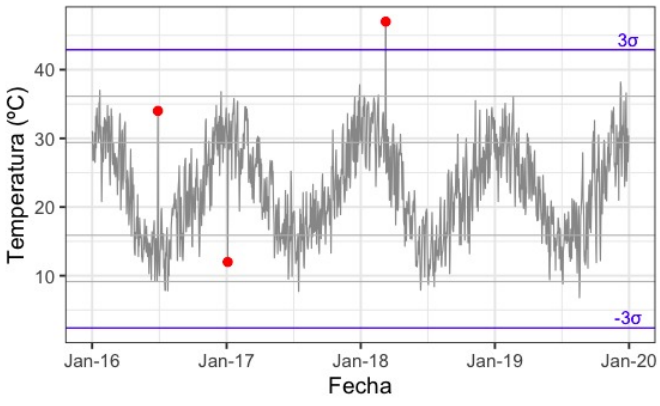

Fig. 1: Temperatura máxima diaria comprendida entre los años 2016 y 2019 en la Ciudad de Buenos Aires, Argentina, con anomalías agregadas manualmente (rojo) y límites de 3 desvíos (azul).

sólo la anomalía global es detectada, mientras las dos restantes locales no llegan a ser identificadas.

En el presente trabajo, se introducirá una técnica llamada Seasonal ESD (S-ESD) y una variante robusta conocida como Seasonal Hybrid ESD (S-H-ESD). Dichos métodos, desarrollados por Twitter, fueron construidos para lidiar con ciertas características propias de las series temporales, como la tendencia y la estacionalidad. Ambas estrategias realizan una descomposición STL (modificada) para extraer el componente residual de la serie y luego aplican un procedimiento basado en tests estadísticos llamado ESD para detectar las anomalías.

La detección se realiza mediante un procedimiento llamado Generalized Extreme Studentized Deviate (ESD), que es una extensión del test estadístico de Grubb. El test de Grubb es usado para encontrar un único outlier en un conjunto de datos univariado, que se asume proviene de una distribución normal. Su idea es simple: en primer lugar ubica el dato cuyo desvío absoluto sea máximo (entre los valores y la media) y luego decide si dicho valor es un outlier al comparar su valor estadístico, G (*z-score*) con un valor crítico que depende de la distribución t student. Si G supera al valor crítico, ese elemento es calificado como anomalía.

En muchos casos, puede existir más de una anomalía en una serie temporal. ESD generaliza la idea del test de Grubb para permitir la detección de numerosos outliers mediante la aplicación de múltiples tests de Grubb de forma sucesiva, eliminando el dato con el máximo desvío absoluto en cada iteración y volviendo a computar el test. Esta técnica requiere un límite superior para la cantidad de iteraciones a realizar. Como se mencionó anteriormente, el test asume la normalidad de los datos, lo cual no siempre ocurre en series temporales. Para solucionar este problema, el S-ESD realiza una descomposición, la cual elimina los componentes de estacionalidad y de tendencia de la serie, dejando solamente el componente residual. Este componente puede ser tratado mediante el ESD por tratarse de una distribución unimodal. La descomposición se realiza mediante una variante de la técnica conocida como STL (Seasonal and Trend decomposition using Loess). Se dará una explicación más detallada del STL en la sección III de este trabajo. Sin embargo, es necesario aclarar que la aplicación directa de esta técnica puede producir anomalías espurias (por algunos efectos de los outliers en la tendencia), por lo que suele reemplazarse la componente de tendencia por un valor ”estable” como la mediana de la serie.

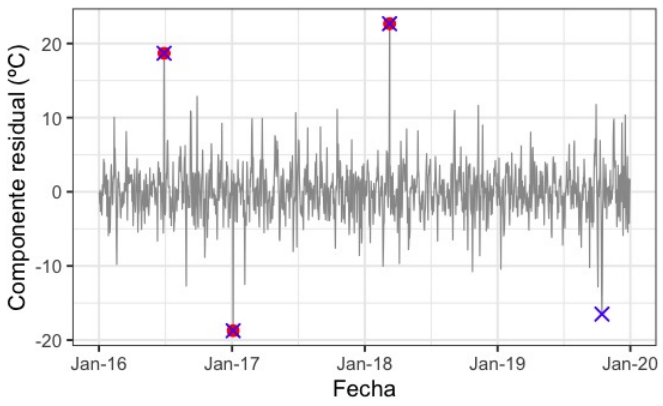

Fig. 2: Anomalías de temperatura máxima detectadas mediante algoritmo S-ESD sobre la componente residual (α = 0.05).

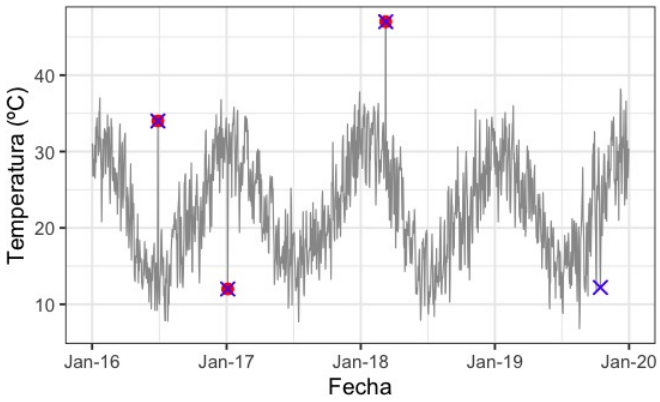

Fig. 3: Anomalías de temperatura máxima detectadas mediante algoritmo S-ESD sobre la serie original (α = 0.05).

Se procede a aplicar el algoritmo S-ESD en la serie temporal de temperatura máxima con anomalías creadas artificialmente (puntos rojos). Esta vez, el método descubre tanto la anomalía global como las anomalías locales. Se grafica el componente residual luego de la variante de la descomposición STL, presentado en la Fig. 2, y se marcan las anomalías detectadas por el ESD con un valor de α (para el test estadístico) de 0.05. Por último, en la Fig. 3, se grafica la serie temporal original con los respectivos outliers.

Si bien el S-ESD puede detectar anomalías tanto globales como locales en series temporales, también tiene sus limitaciones. Debido a que los datos atípicos pueden distorsionar la media y el desvío estándar, resulta un problema para el algoritmo cuando la serie contiene muchas anomalías, ya que puede ”corromper” los residuos. Para solucionar este problema, se suele utilizar la variante robusta S-H-ESD, la cual reemplaza la media y el desvío estándar por medidas estadísticas robustas como la mediana y el MAD (desvío mediano absoluto) para computar el *z-score* en el ESD.

Por último, es necesario mencionar que las técnicas analizadas hasta el momento solo detectan datos anómalos que resultan extremos en la componente residual de la serie. Sin embargo, no es posible identificar ciertos datos anómalos que no cumplan con dichas características. Una situación anómala indetectable es aquella en la que, por ejemplo, el medidor de temperatura reporta por varios días consecutivos el mismo valor, por alguna falla en su funcionamiento. Como se aprecia en la Fig. 3, estos algoritmos no logran hallar dichas situaciones. Una potencial manera de detectar esta falla es usando reglas fijas o realizando una autocorrelación.

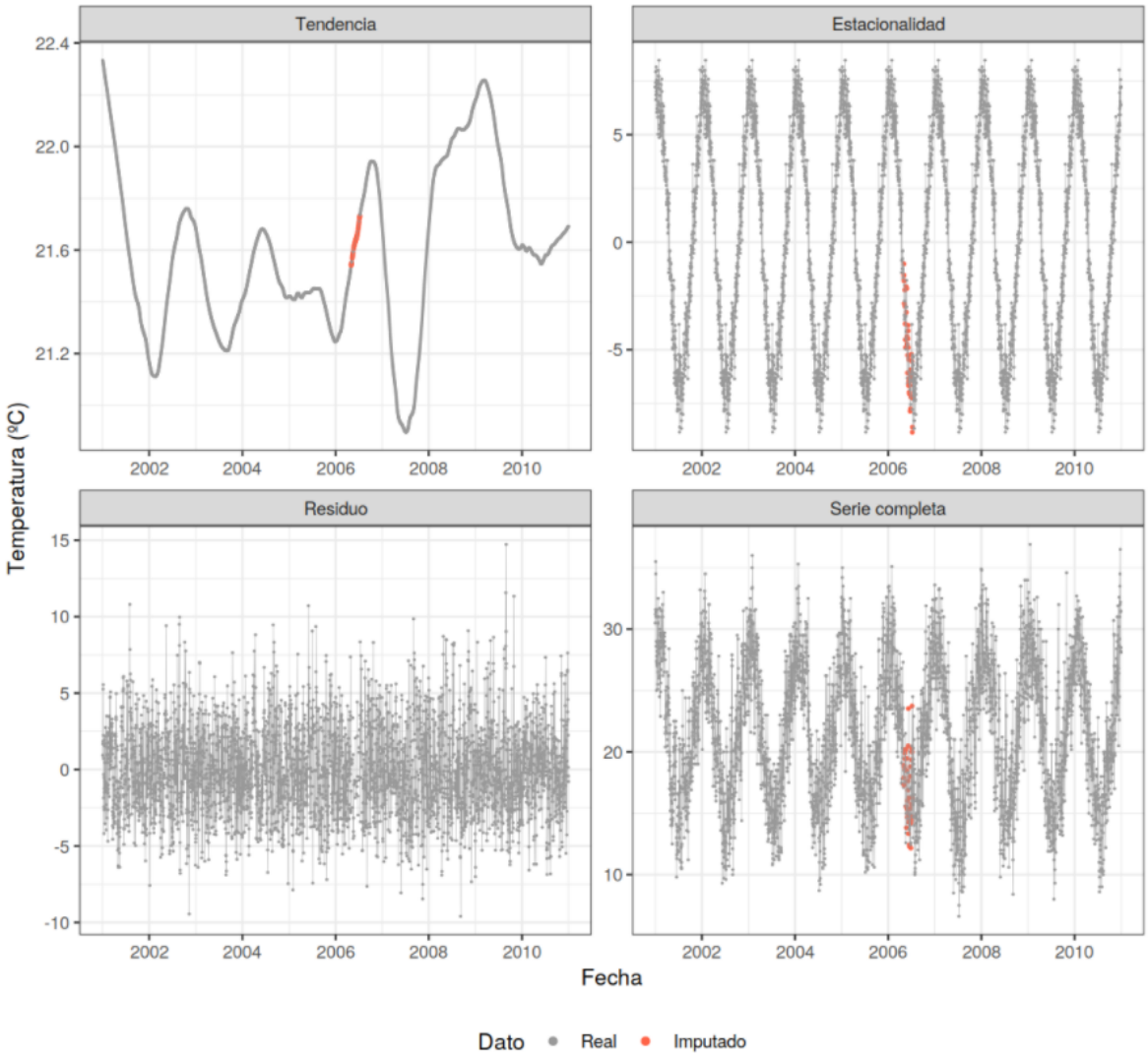

Fig. 4: Descomposición STL aditiva de la temperatura máxima en el Aeroparque Buenos Aires (Ciudad de Buenos Aires, Argentina).

III. Imputación de datos faltantes

El proceso de detección de datos anómalos puede desembocar en la aparición de observaciones faltantes, en particular si decidimos corregir los valores desviados del patrón de la serie. Por lo tanto, una alternativa es eliminarlos para luego adecuarlos al conjunto de datos. En consecuencia, tanto estos datos eliminados como aquellos que se encontraban omitidos de las mediciones originales, representan un problema cuando su proporción dentro del conjunto no resulta despreciable.

Para abordar esta problemática se suele completar estos datos faltantes mediante algún proceso de imputación. En la literatura existen diversos mecanismos de imputación aplicables a series temporales [14]–[17]. En este caso particular, en el que se sabe a priori que la serie temporal tiene ciertas características de estacionalidad y continuidad en el espacio, se aplican dos métodos de imputación de forma complementaria.

Las series temporales como la de temperatura máxima (y mínima) pueden ser descompuestas en una componente estacional (con periodicidad anual), en otra de tendencia (sin periodicidad en general) y en una restante de alta frecuencia denominada residuo. Esta descomposición será realizada mediante el algoritmo STL [12] aditivo, el cual permite definir a una serie temporal como una suma de una tendencia, una componente estacional y un residuo, presentados en la Fig. 4. En este caso, se utiliza el paquete STLplus[2] de R que admite la existencia de valores faltantes en la serie temporal de entrada.

Tanto la tendencia como la estacionalidad son componentes de baja frecuencia, por lo que pueden ser aproximadas por curvas suaves, debido a que no sufren cambios abruptos. Por lo tanto, a pesar de la existencia de datos faltantes en la serie de entrada, el algoritmo automáticamente devuelve las series completas de tendencia y estacionalidad. De esta forma, parte del proceso de imputación es realizado utilizando solamente la naturaleza temporal de los datos y teniendo presente que los dos componentes previos representan las frecuencias más bajas de la señal.

Sin embargo, la componente residual es de alta frecuencia y, por lo tanto, imposible de imputar mediante la metodología previa. Es aquí donde se busca explotar la naturaleza espacial de la señal. Los campos escalares de temperatura son funciones que tienen continuidad espacial y que, en general, no tienen cambios abruptos, particularmente, en ubicaciones geográficamente cercanas. Esto es, existe una alta correlación cruzada entre las series de residuos para puntos cercanos. Por tal motivo, se procede a imputar los residuos faltantes mediante un proceso de interpolación espacial denominado *kriging*.

El proceso de *kriging* es un procedimiento geoestadístico que permite inferir valores de una variable en lugares no muestreados [18], [19]. Este método es el mejor estimador lineal insesgado de varianza mínima. La aplicación de *kriging* para realizar interpolación espacial de temperatura ya ha sido utilizado anteriormente en otros trabajos, por ejemplo en [20]. Existen distintas técnicas de *kriging* cuya explicación va más allá del alcance de este trabajo, pero pueden encontrarse en [21].

Para interpolar los residuos de temperatura, se utiliza *kriging ordinario*. Esta técnica requiere que el campo a interpolar cumpla los siguientes supuestos:

a La media se mantiene constante en el espacio.
b La covarianza o correlación entre los valores del campo entre dos puntos depende sólo de la distancia entre ellos.

Debido a que el proceso de descomposición STL elimina la tendencia, el primer supuesto (a) se debería cumplir automáticamente, dejando los residuos de todas las muestras centrados en el valor cero.

El segundo supuesto (b), en realidad, no es posible asegurarlo formalmente, pero, como se expuso anteriormente, las series de temperatura y, en particular, sus residuos son funciones que deberían estar fuertemente correlacionadas para puntos muy cercanos. Asimismo, esa correlación debería disminuir conforme se aumenta la distancia. Esa presunción puede ser puesta a prueba mediante el ajuste de un modelo llamado variograma, el cual permite encontrar una expresión analítica para definir la covarianza de los residuos entre dos puntos como función de su distancia. El ajuste del variograma es parte del proceso de *kriging* y, si éste puede ser correctamente modelado, entonces no solamente se acepta el supuesto (b) sino que además se tienen todos los elementos necesarios para proceder con la interpolación.

En la Fig. 5 se observa el variograma ajustado y el campo interpolado de residuos para una fecha en particular. A partir de este campo interpolado es posible calcular el valor del residuo en ubicaciones donde se detecten valores faltantes. Finalmente, sumando los valores de estacionalidad y tendencia previamente calculados a partir del algoritmo STL, se obtiene un valor de temperatura que respeta tanto la estacionalidad y tendencia de la señal completa. Adicionalmente, la imputación realizada guarda correlación espacial con las señales de ubicaciones vecinas, tal como se puede observar en el panel de la serie completa en la Fig. 4.

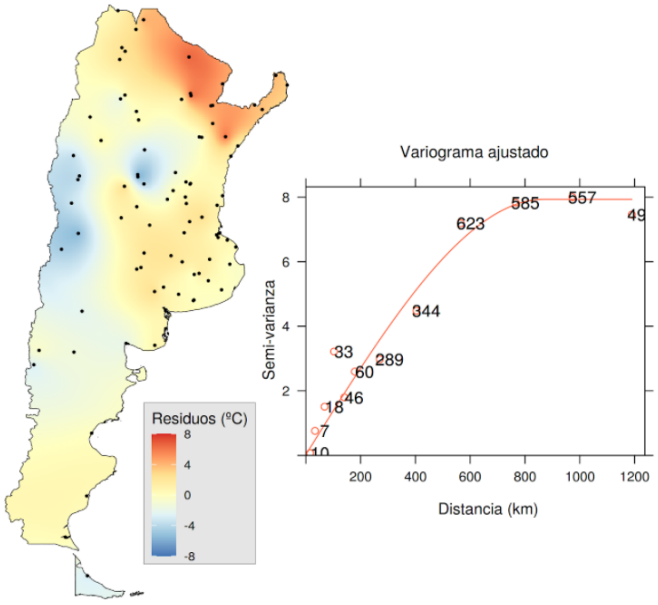

Fig. 5: Interpolación de residuos de la temperatura máxima observada el 16 de Junio de 2006 en el Aeroparque Buenos Aires (Ciudad de Buenos Aires, Argentina).

Haciendo uso de esta metodología se pueden obtener series temporales de temperatura completas, las cuales pueden formar parte de la entrada de procesos posteriores, como lo es el proceso de clustering de las estaciones que se presenta a continuación.

IV. Regionalización Climática

Como se anticipó, se realiza a continuación un clustering de las series de temperatura de las estaciones bajo análisis, de modo de poner de manifiesto en un caso práctico el valor de disponer de datos de calidad. Al trabajar con un fenómeno geográfico y más específicamente climático, dicha agrupación no supervisada de las observaciones se enmarca el campo de la denominada *regionalización climática*. Siguiendo a Badr et al. [22], en cuyo trabajo se presentan herramientas específicas para tal tarea, "la regionalización es el proceso de dividir un área en regiones más pequeñas que resulten homogéneas con respecto a una determinada característica climática; es un ejercicio fundamental en los estudios climáticos porque permite distinguir entre los mecanismos responsables de la variabilidad espacio-temporal específica de cada región".

Para llevar adelante tal agrupación, la primera y más relevante cuestión resulta determinar cuál es el criterio para evaluar que un conjunto de series temporales resultan (di) similares. Tal concepto resulta particularmente complejo al trabajarse con datos en el dominio tiempo, ya que los criterios utilizados a tales fines en conjuntos de datos "estáticos" no suelen funcionar adecuadamente en la medida en que ignoran la interdependencia entre los valores. En ese contexto, una gran cantidad de enfoques han sido propuestos en la literatura, cuya discusión excede el presente trabajo y cuyo análisis detallado puede encontrarse en el trabajo de Liao [23]. En

el presente trabajo se procede a utilizar el enfoque detallado por Chouakria y Nagabhushan [24], implementado en la librería TSclust [13], que combina la intuición de los enfoques no paramétricos más ampliamente difundidos basados en la proximidad de los valores de las series con otro basado en el comportamiento de las mismas (1), denominada correlación temporal de primer orden (CORT) y definida en (2), esto es, si (de)crecen simultáneamente y a qué velocidad. La importancia de estos efectos queda luego modulada por una función de ajuste (3); en este caso se utilizó el valor por defecto $k = 2$.

$$d_{CORT} = \Phi_k(CORT) * dEUC_{XY} \tag{1}$$

$$CORT = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}} \tag{2}$$

$$\Phi(k, x) = \frac{2}{(1 + e^{kx})}, k \geq 0 \tag{3}$$

El segundo paso consiste en identificar el algoritmo de clustering a utilizar. Cada uno de ellos presenta ventajas y desventajas y no existe un consenso en la literatura respecto a cuál es el mejor [25]. En esta ocasión, se optó por trabajar con un algoritmo de cluster jerárquico aglomerativo (en su variante del método de Ward, ampliamente difundido en análisis de clima ([26]–[28], entre otros)), principalmente porque permite observar en forma clara los sucesivos agrupamientos, y así seguir la evolución de los grupos si se los examina en el espacio. A este respecto, cabe destacar que no se cuenta con una verdad de campo externa para validar los clusters, con lo cual se utiliza el criterio sugerido en [29], según el cual la agrupación "resulta satisfactoria cuando las regiones son homogéneas y geográficamente contiguas".

La primera prueba para evaluar la posibilidad de formar agrupamientos relevantes en las series temporales en cuestión fue obtener el estadístico de Hopkins, índice que mide la tendencia a la agrupabilidad de un conjunto de datos [30]. El valor obtenido, aproximadamente 0.8, da cuenta de la existencia de una estructura subyacente en los datos que permite generar grupos entre las series de temperaturas analizadas.

En la misma dirección, el dendrograma presentado en la Fig. 6, resultante de llevar adelante el agrupamiento con las técnicas antes descritas, evidencia la presencia de clusters, en la medida en que diferentes cortes permiten obtener grupos compactos, que se van a su vez uniendo cada vez a mayor distancia. Se ilustran aquí con líneas punteadas 3 puntos de cortes plausibles, en distancia 400, 700 y 1100, lo que arroja 8, 4 y 3 clusters respectivamente. Como se destacó antes, estas particiones no son mutuamente excluyentes, si no que por el contrario, la capacidad de ir agrupándolas jerárquicamente que ofrece la técnica resulta particularmente valiosa, en la medida en que puede analizarse el espacio en forma recursiva, comenzando por regiones amplias hasta llegar a otras subregiones de menor extensión geográfica.

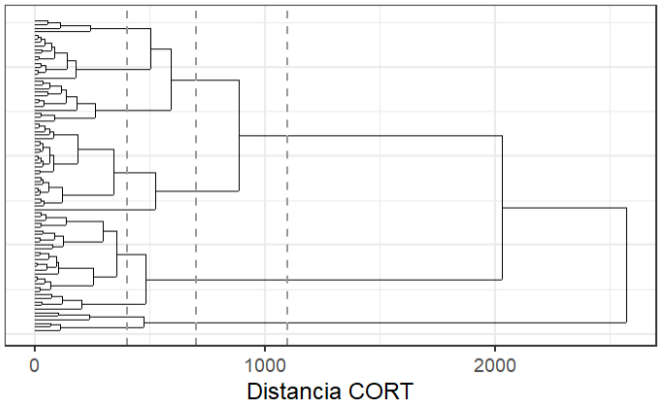

Fig. 6: Dendograma de agrupación de estaciones incluyendo diferentes puntos de corte.

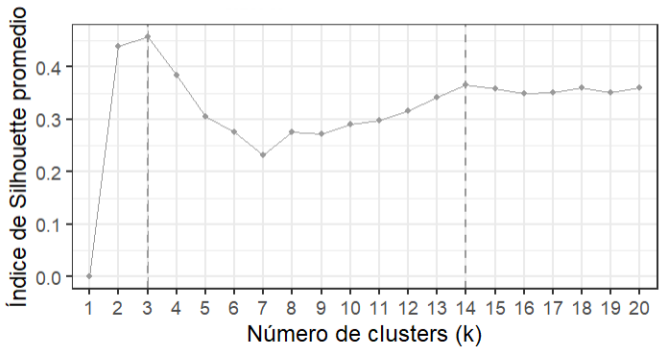

Fig. 7: Cálculo de k óptimo con coeficiente de Silhouette promedio vs cantidad de clusters.

Hecha esta aclaración, aún resulta de interés determinar analíticamente la partición más compacta, para lo cual se utilizó el criterio del coeficiente de Silhouette promedio. El resultado, presentado en la Fig. 7 arroja que el número óptimo de clusters es 3, pero que sin embargo existe un máximo local en 14, posibilitado por una tendencia creciente que comienza en 8 clusters y se estabiliza tras superar estos 14 grupos.

Finalmente, al no contar con una verdad de campo, la inspección visual en el espacio de los agrupamientos resulta el factor de validación más relevante. De la Fig. 8 se desprende rápidamente que los agrupamientos o regiones resultan contiguos en el espacio ya desde la primera imagen, usando 3 clusters (a), quedando delimitada una región de Sur, Centro y Norte del país. Al utilizar un mayor número de clusters, 8 y 14 respectivamente en (b) y (c), se observa la subdivisión en regiones cada vez menos extensas que siempre conservan su contigüidad geográfica o, dicho de otro modo, que guardan relación con sus coordenadas de latitud y longitud. Tal es así, que (con contadas excepciones) es posible unir los puntos de las estaciones que pertenecen al mismo cluster sin incluir ninguno de una región distinta.

Estos signos cualitativamente inequívocos de éxito en la regionalización, dan cuenta de su utilidad para el análisis y resolución de problemáticas que tengan a la dimensión

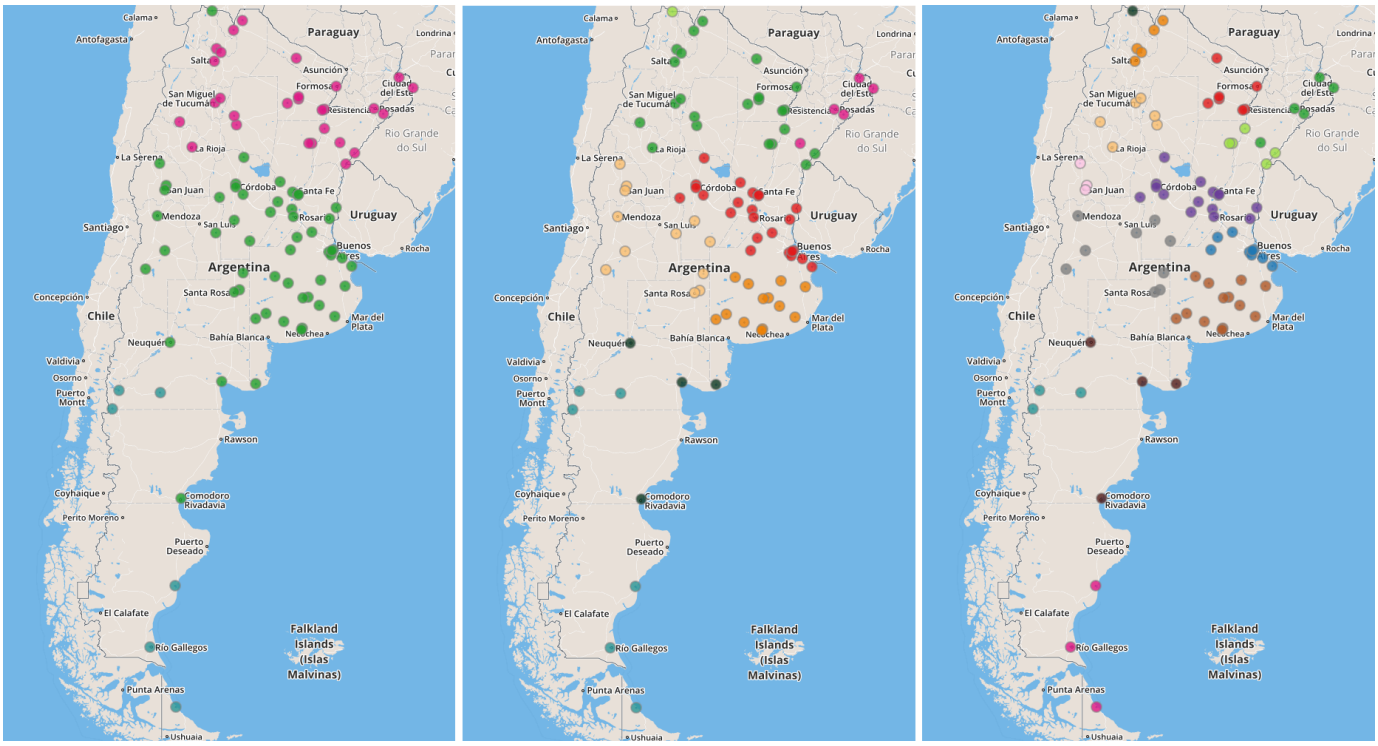

Fig. 8: Agrupamientos de estaciones según sus registros de temperatura para diferentes selecciones de cantidad de clusters (k). El primer mapa (a) presenta 3 grupos ($k = 3$), el segundo (b) 8 ($k = 8$) y el tercero (c) 14 ($k = 14$).

geoespacial y temporal como dos de sus características más relevantes, sobre todo en la medida en que se extienda al campo multivariado, esto es, que incorpore un mayor número de variables permitiendo precisar así sectores del mapa aún más específicos.

V. Conclusiones

El proceso metodológico propuesto en este trabajo propone abordar el problema de la calidad de datos en series de tiempo a través de la aplicación de técnicas sencillas. Sustentado en la detección de datos anómalos y en la posterior imputación de datos faltantes, es posible obtener un conjunto de datos completo y consistente con los patrones latentes de las series de tiempo, incorporando inclusive el concepto geoespacial para una interpolación más adecuada. En ese aspecto, la decisión de utilizar la técnica de *kriging* para la imputación manifiesta resultados muy satisfactorios.

En un segundo aspecto, la clusterización aplicada sobre mediciones meteorológicas en estaciones georeferenciadas, permite detectar diversas regiones climáticas con gran precisión; el cómputo de disimilaridad entre series basado en la correlación temporal que estas presentan arroja así un buen rendimiento en la segmentación de las estaciones.

Las técnicas mencionadas en este trabajo forman parte de un corpus realmente amplio, donde conviven un sinfín de herramientas y algoritmos definidos y estudiados para problemáticas específicas. Una posible línea futura de investigación será comprobar la eficacia, por ejemplo, de distintas formas de cómputo de similaridad de series de tiempo, detección de anomalías o interpolación multivariada.

Referencias

[1] Mcnally, J. T., Price, J. U., Vaiser, B., & Allen, R. C. (2007). Weather data quality control and ranking method. U.S. Patent No. 7,228,234. Washington, DC: U.S. Patent and Trademark Office.

[2] Woodall, P., Oberhofer, M., & Borek, A. (2014). A classification of data quality assessment and improvement methods. International Journal of Information Quality 16, 3(4), 298-321.

[3] El Arass, M., & Souissi, N. (2018, October). Data lifecycle: from big data to SmartData. In 2018 IEEE 5th International Congress on Information Science and Technology (CiSt) (pp. 80-87). IEEE.

[4] Gitzel, R. (2016). Data Quality in Time Series Data: An Experience Report. In CBI (Industrial Track) (pp. 41-49).

[5] Estévez, J., Gavilán, P., & Giráldez, J. V. (2011). Guidelines on validation procedures for meteorological data from automatic weather stations. Journal of Hydrology, 402(1-2), 144-154.

[6] Moraru, A., Pesko, M., Porcius, M., Fortuna, C., & Mladenic, D. (2010). Using machine learning on sensor data. Journal of computing and information technology, 18(4), 341-347.

[7] Rahman, A., Smith, D. V., & Timms, G. (2013). A novel machine learning approach toward quality assessment of sensor data. IEEE Sensors Journal, 14(4), 1035-1047.

[8] Wu, T., & Li, Y. (2013). Spatial interpolation of temperature in the United States using residual kriging. Applied Geography, 44, 112-120.

[9] Filzmoser, P., & Hron, K. (2008). Outlier detection for compositional data using robust methods. Mathematical Geosciences, 40(3), 233-248.

[10] Zhang, Y., Hamm, N. A., Meratnia, N., Stein, A., Van De Voort, M., & Havinga, P. J. (2012). Statistics-based outlier detection for wireless sensor networks. International Journal of Geographical Information Science, 26(8), 1373-1392.

[11] Hochenbaum, J., Vallis, O. S., & Kejariwal, A. (2017). Automatic anomaly detection in the cloud via statistical learning. arXiv preprint arXiv:1704.07706.

[12] Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. Journal of official statistics, 6(1), 3-73.
[13] Montero, P., & Vilar, J. A. (2014). TSclust: An R package for time series clustering. Journal of Statistical Software, 62(1), 1-43.
[14] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
[15] Josse, J., & Husson, F. (2016). missMDA: a package for handling missing values in multivariate data analysis. Journal of Statistical Software, 70(1), 1-31.
[16] Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. Journal of statistical software, 45(7), 1-47.
[17] Groothuis-Oudshoorn, K., & Van Buuren, S. (2011). Mice: multivariate imputation by chained equations in R. J Stat Softw, 45(3), 1-67.
[18] Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy, 52(6), 119-139.
[19] Matheron, G. (1962). Traité de géostatistique appliquée (Editions Technip.).
[20] Monestiez, P., Courault, D., Allard, D., & Ruget, F. (2001). Spatial interpolation of air temperature using environmental context: application to a crop model. Environmental and Ecological Statistics, 8(4), 297-309.
[21] Moyeed, R. A., & Papritz, A. (2002). An empirical comparison of kriging methods for nonlinear spatial point prediction. Mathematical Geology, 34(4), 365-386.
[22] Badr, H. S., Zaitchik, B. F., & Dezfuli, A. K. (2015). A tool for hierarchical climate regionalization. Earth Science Informatics, 8(4), 949-958.
[23] Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.
[24] Chouakria, A. D., & Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. Advances in Data Analysis and Classification, 1(1), 5-21.
[25] Manning, C. D., Schütze, H., & Raghavan, P. (2008). Introduction to information retrieval. Cambridge university press.
[26] Regonda, S. K., Zaitchik, B. F., Badr, H. S., & Rodell, M. (2016). Using climate regionalization to understand climate forecast system version 2 (CFSv2) precipitation performance for the Conterminous United States (CONUS). Geophysical Research Letters, 43(12), 6485-6492.
[27] Ferrelli, F., Brendel, A. S., Aliaga, V. S., Piccolo, M. C., & Perillo, G. M. (2019). Climate regionalization and trends based on daily temperature and precipitation extremes in the south of the Pampas (Argentina). Cuadernos de Investigación Geográfica, 45(1), 393-416.
[28] Darand, M., & Daneshvar, M. R. M. (2014). Regionalization of precipitation regimes in Iran using principal component analysis and hierarchical clustering analysis. Environmental Processes, 1(4), 517-532.
[29] Dezfuli, A. K. (2011). Spatio-temporal variability of seasonal rainfall in western equatorial Africa. Theoretical and applied climatology, 104(1-2), 57-69.
[30] B. Hopkins, J. G. Skellam, A new method for determining the type of distribution of plant individuals, Annals of Botany 18 (70) (1954) 213–227. URL http://www.jstor.org/stable/42907238

ANEXOS

[A] Visualización interactiva publicada en Shiny Applicationes Online. https://srovere.shinyapps.io/SeriesTemporales/