Título: Control de calidad de la Base de Datos del Centro Regional del Clima para el Sur de América del Sur: Eficiencia del control en Pehuajó

Autores: Natalia Herrera, Hernán Veiga, María de los Milagros Skansi, Guillermo Podestá

Gerencia: Investigación, Desarrollo y Capacitación

Departamento: Climatología

Congreso: CONGREMET XII

Lugar: Mar del Plata - Buenos Aires

Fecha: Mayo 2015

Tipo de documento: Póster

Número interno del documento: 0016CL2015

Control de calidad de la Base de Datos del Centro Regional del Clima para el Sur de América del Sur: Eficiencia del control en Pehuajó

N. Herrera[1,2](nherrera@smn.gov.ar), H. Veiga[1,2], M. Skansi[1,2], G. Podestá[2,3]

1. Servicio Meteorológico Nacional - Argentina; 2. Centro Regional del Clima para el Sur de América del Sur;
3. Universidad de Miami, Escuela Rosenstiel de Ciencias Marinas y Atmosféricas - Estados Unidos

Resumen

- Una base de datos extensa y confiable es indispensable para desarrollar productos y servicios climáticos. Para ello es necesario contar con un esquema de control de calidad a través del cual se puedan identificar posibles datos erróneos.
- Se analizó la eficiencia de los tests de control de calidad aplicados a Pehuajó (provincia de Buenos Aires, Argentina). Estos resultados ayudan a configurar el control de calidad operativo para el resto de las estaciones que conforman la Base de Datos del Centro Regional del Clima para el Sur de América del Sur (CRC-SAS).

Base de Datos

Estaciones: Los datos provienen de estaciones convencionales pertenecientes a los servicios meteorológicos e hidrológicos (SMHNs) de:
- Argentina (121 estaciones),
- Bolivia (33 estaciones),
- Brasil (78 estaciones),
- Chile (32 estaciones),
- Paraguay (23 estaciones),
- Uruguay (15 estaciones),
- y al Instituto Nacional de Tecnología Agropecuaria (INTA) de Argentina (44 estaciones) (Fig. 1).

Variables meteorológicas: Durante la primera etapa, los países participantes contribuyeron datos diarios de precipitación, y temperatura máxima y mínima. Posteriormente se incluirán otras variables necesarias para el cálculo de otros productos relevantes para los tomadores de decisión.

Cobertura Temporal: 1961 al presente.

Metadatos: Por ahora, sólo son incluidos los metadatos básicos (ubicación de la estación, elevación, frecuencia de observación).

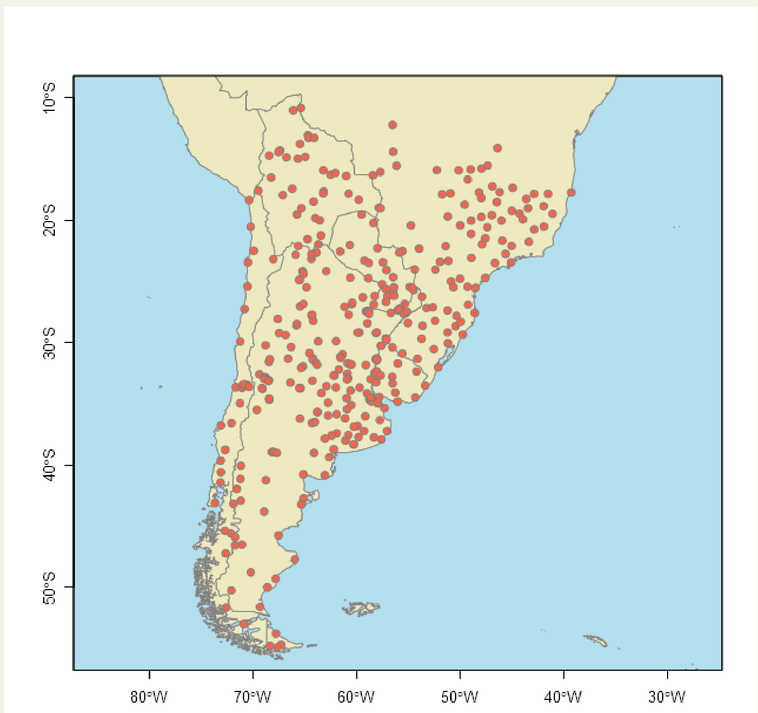

Figura 1: Estaciones meteorológicas incluidas en la Base de Datos del Centro Regional del Clima para el Sur de América del Sur.

Agradecimientos:

Banco Interamericano de Desarrollo

Inter-American Institute for Global Change Research

U.S. National Science Foundation

Control de Calidad

Se construyó un superset con los procedimientos de control de calidad utilizados por los países de la región, y se añadieron varios procedimientos publicados en la literatura científica. Todos los tests de control de calidad se implementaron en el lenguaje R.

Los controles de calidad se organizaron en seis "familias" que agrupan tests de características similares (Figura 2):

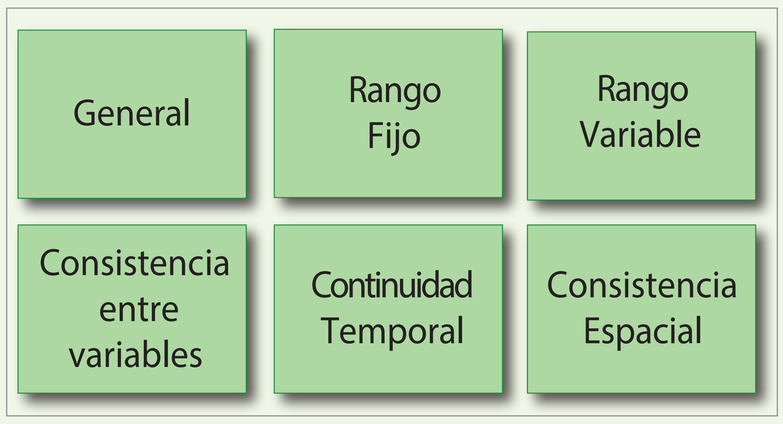

Figura 2: Familias de controles de calidad usadas para identificar valores sospechosos en variables climáticas diarias.

- **General:** verifican la integridad general de los datos. Por ejemplo, se controla que no haya fechas duplicadas o fuera de secuencia en las observaciones diarias.
- **Rango Fijo:** aseguran que no existan valores físicamente imposibles. Los límites propuestos son fijos para cada variable durante todo el período de datos y todas las estaciones meteorológicas.
- **Rango Variable:** los rangos o umbrales usados para identificar valores sospechosos varían con el tiempo, tomando valores específicos para cada día o mes del año.
- **Continuidad Temporal:** estudian las secuencias de valores en días consecutivos, buscando por ejemplo picos o saltos en valores diarios de una variable.
- **Consistencia entre variables:** evalúan la consistencia entre valores de pares o grupos de variables que deben guardar cierta coherencia.
- **Consistencia Espacial:** los valores de una variable para una estación determinada se comparan con los valores de esa variable registrados en estaciones geográficamente cercanas.

Eficiencia del Control de Calidad en Pehuajó

Un control de calidad eficiente es el que balancea el tiempo que demanda la verificación manual de los datos sospechosos y la cantidad de datos realmente erróneos no identificados. Para ello, se realizó la inspección manual de todos los datos de temperatura máxima de Pehuajó para 1961-2012 (18763 datos), de los cuales se encontraron 387 datos realmente erróneos (columna derecha de Tabla 1).

Para cada test analizado se utilizaron 4 configuraciones que van de menos estrictas (el test arroja mayor cantidad de datos sospechosos) a más estrictas (Fig. 3). En el eje x se puede ver la cantidad de datos sospechosos que identifica cada test, y en el eje y el porcentaje de errores encontrados.

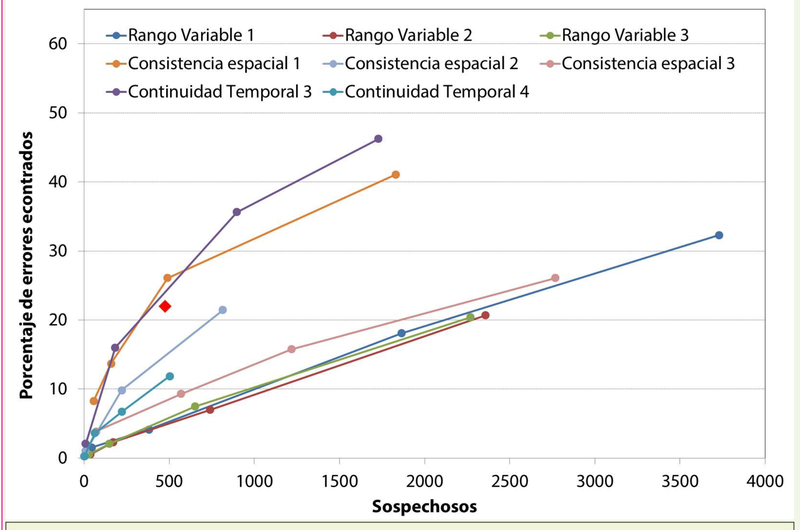

Figura 3. Eficiencia de los diferentes tests para la temperatura máxima en Pehuajó en el periodo 1961-2012. Cada punto representa una configuración distinta. El rombo rojo indica la configuración operativa.

La evaluación de la configuración operativa debería llevarse a cabo caracterizando los dos tipos de errores posibles en los controles (Durre et al., 2008):

(i) observaciones correctas identificadas como "sospechosas" o "errores de Tipo 1" (Tabla 1 celda roja); y

(ii) observaciones incorrectas que no son detectadas por los controles de calidad o "errores de Tipo 2" (Tabla 1 celda azul).

En general, al tratar de minimizar uno de estos tipos de error se aumenta la proporción de errores del otro tipo. El desempeño de este experimento se puede explorar con la matriz de contingencia (Tabla 1).

	Valores realmente válidos	Valores realmente erróneos	
Valores identificados por un test como válidos	17985 (95.85%)	302 (1.61%)	18287
Valores identificados por un test como sospechosos	391 (2.08%)	85 (0.45%)	476
	18376	387	

Tabla 1. Performance de la configuración operativa del control de calidad en Pehuajó en el periodo 1961-2012, luego de la verificación manual de datos.

Se identificaron 476 datos sospechosos, de los cuales 85 son realmente erróneos (22% del total de valores realmente erróneos), y llegaron a presentar diferencias absolutas con respecto al valor correcto de hasta 10.0 °C (Fig. 4), con una distribución semi uniforme entre todos los intervalos. Los errores no detectados fueron más pequeños, cercanos a 0.

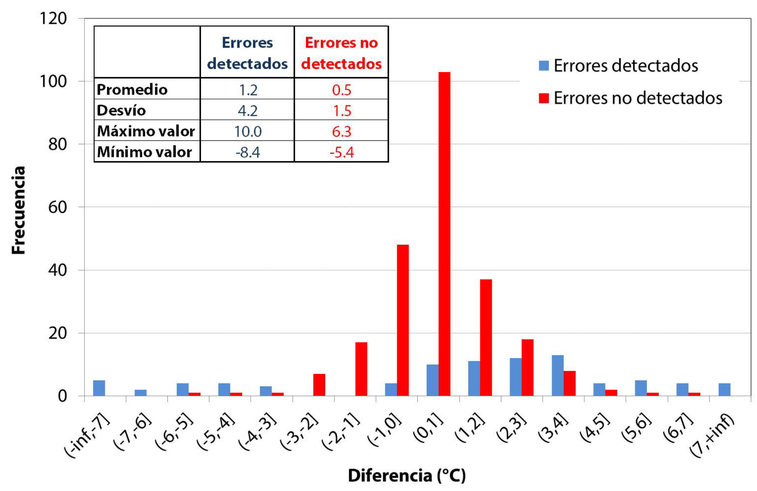

Figura 4. Histograma de las correcciones (diferencia entre el dato realmente erróneo y el correcto) para la configuración utilizada por el CRC-SAS.

Conclusiones

Se evaluó la eficiencia de los tests de control de calidad para detectar errores en la temperatura máxima de Pehuajó, con el fin de elegir una configuración operativa para el control de calidad de la base de datos del CRC-SAS.

- Consistencia espacial y continuidad temporal fueron las familias de tests que mostraron mayor eficiencia, mientras que los tests de rango variable fueron los de menor eficiencia.
- Los datos erróneos no detectados por la configuración operativa presentaron pequeñas diferencias absolutas con respecto al valor correcto. Los errores detectados presentaron diferencias absolutas de hasta 10.0 °C.

Referencias

Durre, I., Menne, M.J. y Vose, R.S., 2008. Strategies for Evaluating Quality Assurance Procedures. Journal of Applied Meteorology and Climatology, 47(6): 1785-1791.

goo.gl/SUw3Ko