

MÉTODOS DE ESTIMACIÓN DE PERCENTILES

María de los Milagros Skansi^(a), Guillermo P. Podestá^(b)

Liliana N. Núñez^(a), Silvia E. Núñez^(a)

^(a)Servicio Meteorológico Nacional, Argentina
mms@meteofa.mil.ar

^(b)University of Miami, Rosenstiel School of Marine and Atmospheric Science
gpodesta@rsmas.miami.edu

RESUMEN

La demanda de información sobre perspectiva climática aumentó considerablemente y con ello surgieron diferentes requerimientos por parte de los usuarios. Como los pronósticos climáticos son presentados asignando la probabilidad de ocurrencia de terciles, una consulta muy asidua es: ¿Cuales son los límites de los mismos? Si bien éstos ya se han calculado empíricamente y en algunos casos utilizando distribuciones teóricas, en este trabajo se exploran las diferencias que pueden surgir al utilizar distintas funciones, incluyendo el calculo de intervalos de confianza. Se presentan los resultados para cuatro trimestres, observándose que en precipitación las diferencias entre los percentiles empíricos y los calculados con las metodologías propuestas fueron predominantemente inferiores al 10%, siendo enero-marzo y octubre-diciembre los períodos con mayores apartamientos, en tanto que los intervalos de confianza mostraron mayor amplitud en julio-septiembre. En temperatura las diferencias no excedieron 0.2°C con los mayores intervalos de confianza en el abril-junio.

The increased demand for predictions of seasonal climate conditions has caused many questions or requests from users of this information. One question asked most often involves the values of boundaries between the lower, middle or upper thirds (or terciles) of the historical distribution of a climate variable, for a given location and period. These values are needed to interpret a common format of seasonal forecasts, which lists the probabilities of a climate variable falling in each tercile. Although boundaries between terciles have been estimated previously for the study region (northern and central Argentina), this work presents a comparison of percentile values estimated using a variety of methods. Additionally, we estimate confidence intervals, a metric which so far has not been presented for the boundaries. We present the results obtained for the four quarters of the year. In relation to precipitation, the differences between the empiric percentiles and those obtained by means of the proposed methodologies were generally below 10%, being maximum in January-March and October-December. As to the confidence intervals, they present the greatest amplitude in July-September. In regard to temperature, no differences greater than 0.2°C were found, with larger confidence intervals in April-June.

INTRODUCCIÓN

En los últimos años la demanda de información acerca de las perspectivas climáticas a 3 y 6 meses por parte de diferentes usuarios, en especial de los sectores agropecuario, hidroeléctrico y turístico, aumentó considerablemente y con ello surgieron diferentes dudas y requerimientos por parte de los mismos. En la actualidad, existen organismos internacionales, regionales y nacionales que producen pronósticos climáticos estacionales en forma regular y operativa.

En muchos casos, el formato elegido para los pronósticos climáticos estacionales es probabilístico, indicando las probabilidades de que el valor de una variable climática determinada (por ejemplo, precipitación en octubre-diciembre) caiga en los tercios inferior, medio, o superior de la distribución histórica de valores. Cada uno de esos tercios o terciles está asociado, respectivamente, con condiciones “por debajo de lo normal”, “cercanas a lo normal”, y “por encima de lo normal.” En ausencia de un pronóstico, entonces, la probabilidad de que el valor de una variable caiga en un tercil es aproximadamente 33.3%. Cuando existen motivos para suponer que un parámetro tenga mayor probabilidad de tomar determinados valores, entonces las probabilidades asignadas a cada tercil cambian. Por ejemplo, si un pronóstico indica que la probabilidad de observar lluvias en el tercil inferior de la distribución (o sea valores menores que lo normal) es de 45%, esto indica que condiciones secas (bajas precipitaciones) son más probables que otro tipo de condiciones. El formato de probabilidad por terciles está bastante difundido y ha sido adoptado, por ejemplo, para los pronósticos climáticos producidos rutinariamente por el International Research Institute for Climate Prediction (IRI). En el sudeste de Sudamérica, desde diciembre de 1997 se han realizado foros regionales de perspectivas climáticas, en los cuales expertos de diferentes países producen un pronóstico consensuado que utiliza ese mismo formato (Buizer y otros, 2000).

Una desventaja del formato de probabilidades por terciles es que los usuarios potenciales de los pronósticos generalmente desconocen los límites entre estas categorías, que por supuesto varían por región. Por esta razón, es difícil para los usuarios asociar un pronóstico climático con la probabilidad de ocurrencia de ciertos valores específicos de la variable de interés en un lugar determinado.

El propósito de este trabajo, entonces, es estimar en forma consistente los valores que definen los límites entre terciles de dos variables climáticas para las cuales se emiten pronósticos estacionales: precipitación total y temperatura media. Aunque se han hecho algunas estimaciones empíricas de estos límites, en este trabajo utilizamos varias formas alternativas para estimar los valores de interés, incluyendo dos estimaciones de densidad empírica poco utilizadas hasta ahora en aplicaciones climatológicas. Otro aspecto importante es la estimación de intervalos de confianza para los percentiles estimados, pues como señalaran Skansi y Hordij (1999), cuando las probabilidades asignadas a cada tercil por los pronósticos climáticos no difieren significativamente de la climatológica, la variación de los percentiles resulta pequeña y puede ser que quede comprendida dentro del mismo intervalo de confianza.

METODOLOGIA Y MATERIAL

Datos Utilizados

Para este estudio se utilizaron datos mensuales de precipitación y temperatura media de 42 estaciones meteorológicas de la red que opera el Servicio Meteorológico Nacional, ubicadas

entre 54° y 64° de latitud Sur, 23° y 41° de longitud Oeste. El período considerado fue 1961-1990, para ser consistente con el período de referencia adoptado en los foros regionales de perspectivas climáticas. En los casos particulares de las estaciones meteorológicas de Coronel Suárez, Iguazú, Tandil, Tres Arroyos y Viedma el período es 1971-1990, en tanto que en Resistencia, Formosa y Bolívar hay entre 2 a 5 años sin registros.

Estimación de Percentiles

El propósito de este trabajo es estimar, de varias formas alternativas, los valores de los límites entre las categorías o terciles utilizados en pronósticos climáticos estacionales. El límite entre el tercil inferior y el tercil medio de una distribución histórica de valores, corresponde al percentil 33% (de aquí en más, p-33). En forma similar, el límite entre el tercil medio y el superior corresponde al percentil 66% (o p-66). Estimamos aquí los valores de p-33 y p-66 para cada una de las estaciones meteorológicas consideradas, y para cuatro trimestres: enero-marzo, abril-junio, julio-septiembre, y octubre-diciembre. Para estimar los percentiles de precipitación utilizamos varios métodos que describimos brevemente en los párrafos siguientes.

1. Estimación empírica de percentiles.

Este es el método más simple para estimar p-33 y p-66, y es el que se ha utilizado hasta el presente en la mayoría de los casos. Por ejemplo, supongamos que estamos estimando los percentiles para precipitación total en el periodo octubre-diciembre en Junín, Provincia de Buenos Aires. Utilizando el período de referencia 1961-1990, hay 30 valores históricos en la distribución. Se ordenan en forma ascendente los 30 valores de precipitación. Los 10 valores más bajos corresponden al tercil inferior, los 10 intermedios al tercil medio, y los restantes más altos corresponden al tercil superior. Los valores de p-33 y p-66 son aquellos que separan, respectivamente los 10 valores más bajos de los 10 intermedios, y éstos de los los 10 valores más altos.

Además de estimar los percentiles para las distribuciones históricas, realizamos en cada caso un ejercicio de remuestreo, de modo de poder obtener estimaciones menos sensibles a valores muy altos o bajos en la serie histórica. Por otro lado, el remuestreo nos permitió obtener intervalos de confianza para los percentiles, cantidades que generalmente no son reportadas (Dunn, 2001). Para cada variable climática, estación meteorológica y trimestre, se tomaron 1000 muestras con reposición de tamaño igual al de la serie histórica. Al tomar las muestras con reposición, un valor en la serie histórica podía no estar incluido en una muestra, o aparecer más de una vez. Para cada muestra se estimaron los valores de p-33 y p-66, y luego se promediaron para obtener las estimaciones por remuestreo. Con las 1000 muestras, se estimaron intervalos de confianza del 95% para cada uno de los percentiles, utilizando el método de “tilting” (Hesterberg, 1999), que permite obtener intervalos más confiables con menor cantidad de muestras.

2. Ajuste de función gamma

El segundo método utilizado para estimar los percentiles (sólo para precipitación) se basó en ajustar una distribución gamma de dos parámetros (Thom, 1958, Barros y Rivero, 1981) a la distribución histórica para cada estación meteorológica y trimestre. Para la estimación de los parámetros de forma y de escala de la distribución, se utilizó un método robusto (Marazzi y Ruffieux, 1996) en lugar del método de máxima verosimilitud. La bondad del ajuste de la

función gamma a cada serie histórica fue verificada a través del test de Kolgomorov-Smirnov. Una vez obtenidos los parámetros de la función, se utilizó la función de probabilidad acumulada de la gamma para obtener los valores de p-33 y p-66.

En este caso también se utilizaron técnicas de remuestreo para obtener valores más confiables de los parámetros de la distribución gamma. Se tomaron 500 muestras con reposición de tamaño igual al de la distribución original. Para cada muestra se estimaron los parámetros de la función gamma ajustada, y se calcularon los valores de p-33 y p-66. El promedio de los percentiles estimados para cada muestra se utilizó como valor final estimado por remuestreo. El remuestreo permitió también obtener intervalos de confianza para las estimaciones.

3. Ajuste de densidad empírica por “kernel”

En este caso, no se asume una forma funcional determinada (por ej., una gamma) para la distribución de los valores históricos de precipitación, sino que se ajusta una función en forma empírica. La estimación no-paramétrica de funciones de densidad de probabilidad es un tema ampliamente tratado en la literatura estadística y puede ser interpretada como un suavizado de histogramas (Venables y Ripley, 1997, Matyasovszky, 1997, Bowman y Azzalini, 1999). Brevemente, el método está basado en estimar una función de densidad de probabilidad $\hat{f}_K(x)$ para una muestra x_1, \dots, x_n , utilizando un “kernel” (núcleo) fijo $K()$ (una función de densidad de probabilidad, por ejemplo la distribución normal) y un ancho de banda b :

$$\hat{f}_K(x) = \frac{1}{nb} \sum_{j=1}^n K\left(\frac{x-x_j}{b}\right)$$

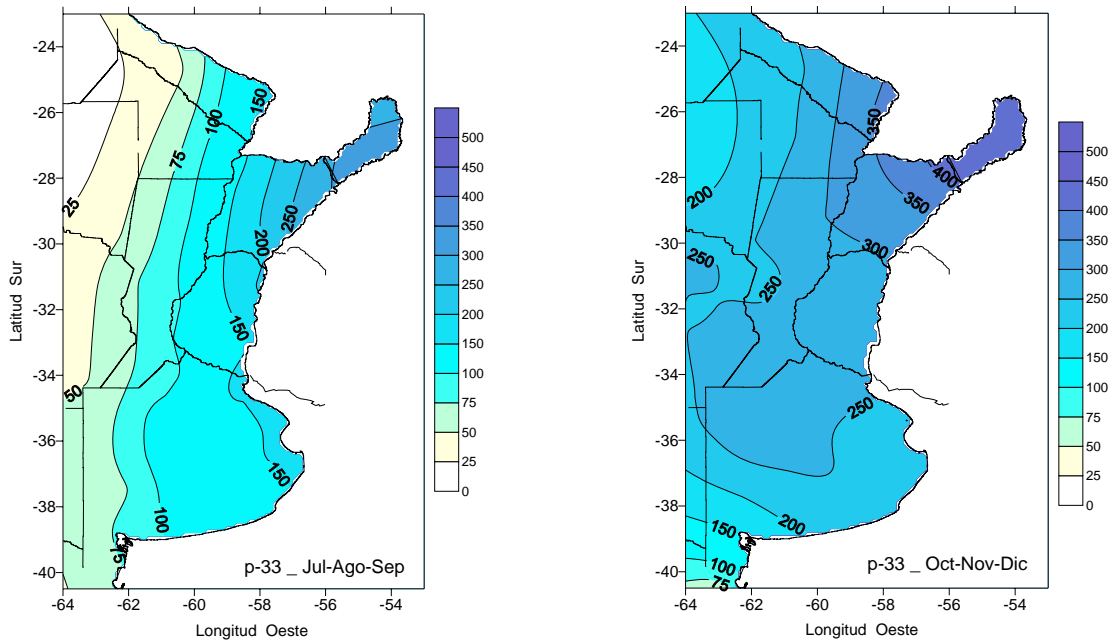
La elección del ancho de banda es un compromiso con el grado de suavizado, de forma tal de no suavizar excesivamente picos reales. Con un ancho de banda pequeño se obtiene una curva de densidad muy rugosa en tanto que al aumentar el tamaño de la ventana la curva es más suave. Lo que se busca con la elección de “b” es tratar de que el error cuadrático medio no aumente considerablemente. En este caso, el ancho de banda utilizado fue obtenido mediante el método de Sheather y Jones (1991). Además para evitar en la estimación valores negativos se aplicó una transformación logarítmica a los datos antes del ajuste de densidad. Este método de estimación de densidad ha sido implementado en la librería de programas “sm” (Bowman y Azzalini, 1999), disponible para el software estadístico S-Plus o R.

4. Ajuste de densidad empírica por “logsplines”

Como en el método anterior, no se asume aquí una forma funcional determinada para la distribución de los valores históricos de precipitación. Por el contrario, siguiendo el método de Kooperberg y Stone (1991) se modela una función de log-densidad como un “spline” cúbico:

$$\hat{f}_{LS}(x, \theta) = \exp\left\{\sum_{i=1}^p \theta_i B_i(x) - c(\theta)\right\}, \quad x \in \mathfrak{R},$$

donde $c(\theta)$ es un factor de normalización y las funciones $\{B_i(x)\}$ denotan los B-splines cúbicos standard. Los coeficientes θ son estimados por máxima verosimilitud. Este método de



Figuras 1 c-d: p-33 de precipitación (mm) correspondiente a jul-sep y oct-dic.

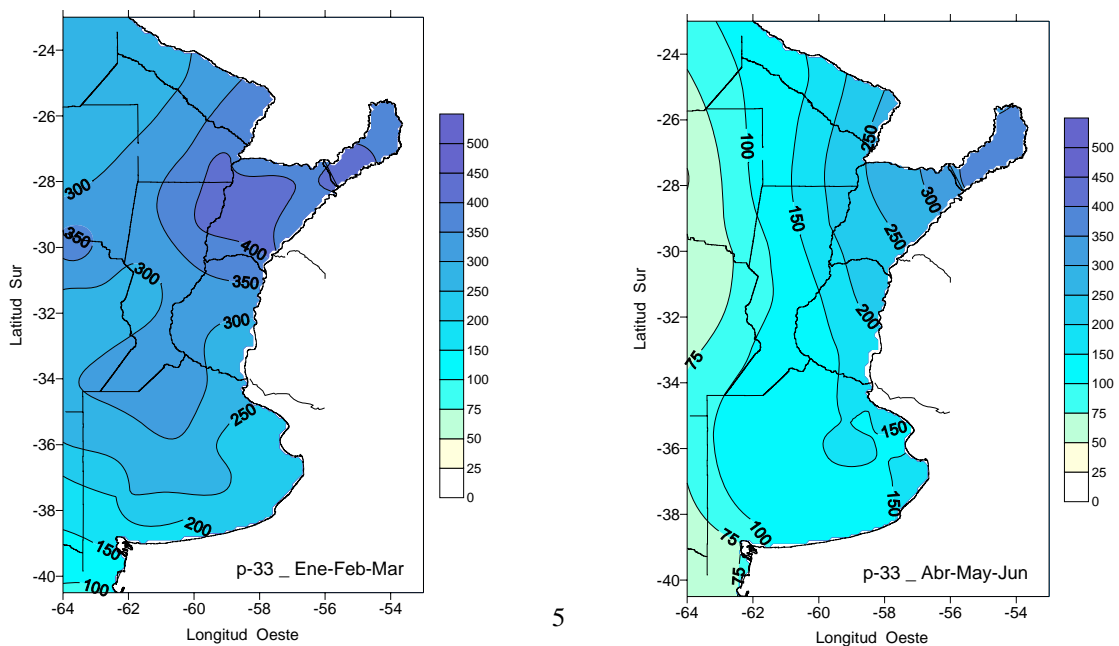
estimación de densidad ha sido implementado en la librería de programas “best”, disponible para el software estadístico S-Plus o R.

RESULTADOS

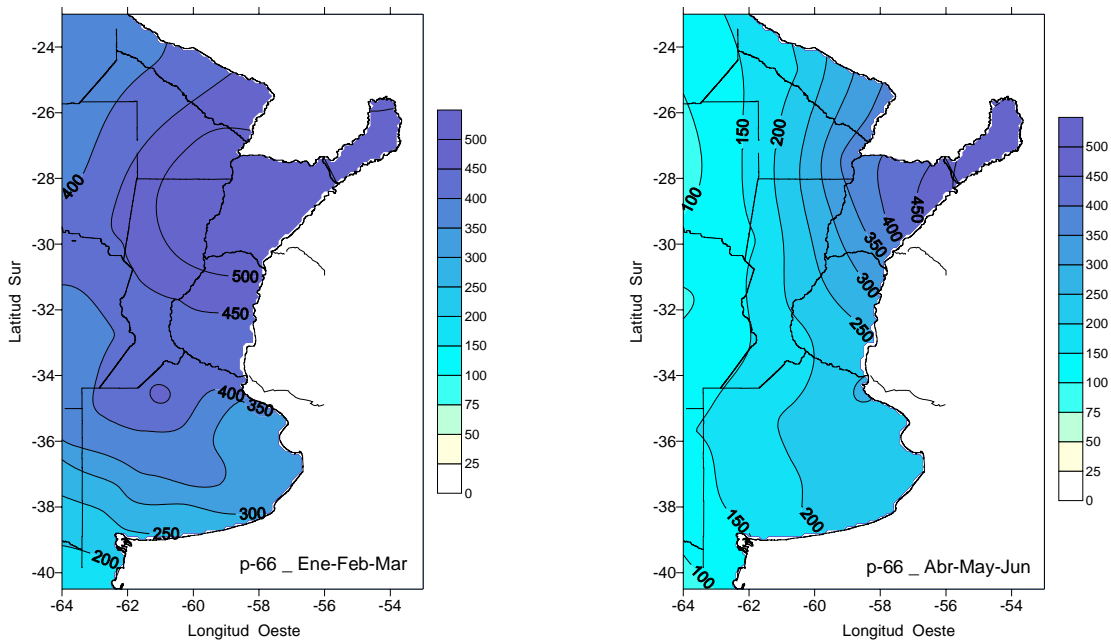
Precipitación

A continuación se presentan, para los trimestres analizados, los mapas de los p-33 y p-66 obtenidos empíricamente luego de aplicar a la serie un remuestreo. Se eligieron estos mapas para representarlos, ya que como se discutirá posteriormente, no mostraron diferencias significativas con respecto a los percentiles empíricos, los cuales serán luego la referencia para hacer las comparaciones con las diferentes metodologías.

En las figuras 1 a-d se presentan los valores del p-33, para los cuatro trimestres. En las mismas se observan los mayores valores en período de ene-mar con más de 200 mm a excepción del sudoeste de la provincia de Buenos Aires. Los valores mayores corresponden al noreste del



Figuras 1 a-b: p-33 de precipitación (mm) correspondiente a ene-mar y abr-jun



Figuras 2 a-b: p-66 de precipitación (mm) correspondiente a ene-mar y abr-jun.

país donde superan los 350 mm. En abr-jun y jul-set los valores de este percentil son los más bajos con una clara disminución hacia el sector oeste de la región de análisis, indicando el período con menores precipitaciones. En jul-sep los valores del p-33 son inferiores a 100 mm excepto en el Litoral fluvial y este de Buenos Aires.

Con respecto al p-66, en las figuras 2 a-d se presentan los valores para cada uno de los períodos considerados. Concordante con el comportamiento del p-33, se observan los mayores valores en el trimestre ene-mar, con un percentil superior a 350 mm, salvo en el sur de la provincia de Buenos Aires donde oscila entre 150 mm y 350 mm. En jul-sep se registran los valores más bajos, excepto en Misiones donde no hay casi diferencia con el trimestre abr-jun. En ambos trimestres se observa también una marcada disminución hacia el oeste donde los percentiles no superan los 150 mm.

Con respecto a las diferencias entre los métodos, se compara cada uno con respecto a los percentiles empíricos:

1. Empírica con remuestreo

Como ya se ha señalado anteriormente, en general, no se presentaron diferencias superiores al 5%. Dentro del año las mayores diferencias correspondieron al p-33, en los trimestres de ene-mar y oct-dic. Con respecto a los intervalos de confianza en ene-mar y oct-dic la amplitud fue menor que en los otros períodos analizados, observándose los mayores intervalos de confianza en el período jul-sep, en el sector oeste del área de análisis.

2. Gamma

Los percentiles calculados utilizando la distribución gamma presentaron mayores diferencias que en el caso anterior, especialmente en el p-33. Asimismo, dentro de los períodos analizados, dichas diferencias fueron menores en los trimestres de abr-jun y jul-sep. En cuanto a la bondad de los ajustes, en la mayor parte de las estaciones fue aceptado con nivel de significancia del 5%. Dentro de las excepciones se destaca el ajuste de la serie de Paso de los Libres, en el trimestre oct-dic, que fue rechazado, observándose en el mismo el mayor apartamiento en el valor de p-33 (28 %). En ene-mar se observó que la distribución tendió a desplazar ambos percentiles en el mismo sentido, mientras que en los otros períodos no siempre se observó esta tendencia. También se vio que las mayores diferencias, no necesariamente eran concordantes con una menor bondad en el ajuste. En cuanto a los resultados obtenidos a través del remuestreo, se observó predominantemente que las diferencias se mantenían con el mismo signo pero con valores más pequeños. Con respecto a los intervalos de confianza, se observó un comportamiento similar al obtenido en el punto anterior, es decir mayores intervalos en el trimestre jul-sep, aunque en este caso los intervalos de confianza resultaron casi siempre más amplios.

3. Kernel

Los resultados obtenidos utilizando esta distribución mostraron, en la mayor parte de las estaciones diferencias inferiores al 10%. En particular en ene-mar se observó una tendencia a valores más bajos del p-33, tendencia que predominó en todos los trimestres. En cuanto al p-66, en general, las diferencias fueron menores al 2% excepto en sectores aislados, sin observarse una tendencia hacia valores mayores o menores que los empíricos. Se destaca que contrariamente a los casos analizados anteriormente, en los períodos de abr-jun y jul-sep, si bien se observaron predominantemente diferencias inferiores al 5%, puntualmente las mismas fueron superiores a dicho valor, especialmente en el p-33, quedando reflejado que en las estaciones con alta frecuencia de valores inferiores a 100 mm (Santiago del Estero, Pilar) el ajuste no es adecuado. También en estos períodos el p-66 tendió a ser predominante superior al empírico.

4. Logspline

Las diferencias obtenidas en general fueron inferiores al 8%, siendo mayores en ene-mar y oct-dic, al igual que en los casos anteriores. En particular, en ene-mar se observó una tendencia a valores mayores del p-33 y con respecto al p-66, a diferencia de los casos anteriores, hay varios sectores con diferencias superiores al 5% y contrariamente a lo ocurrido con la distribución gamma, en general no se observa la misma tendencia en ambos percentiles, es decir tiende a subir uno y bajar el otro. La mayor diferencia se observó en la estación meteorológica de Iguazú en el trimestre oct-dic con +17% en el p-33.

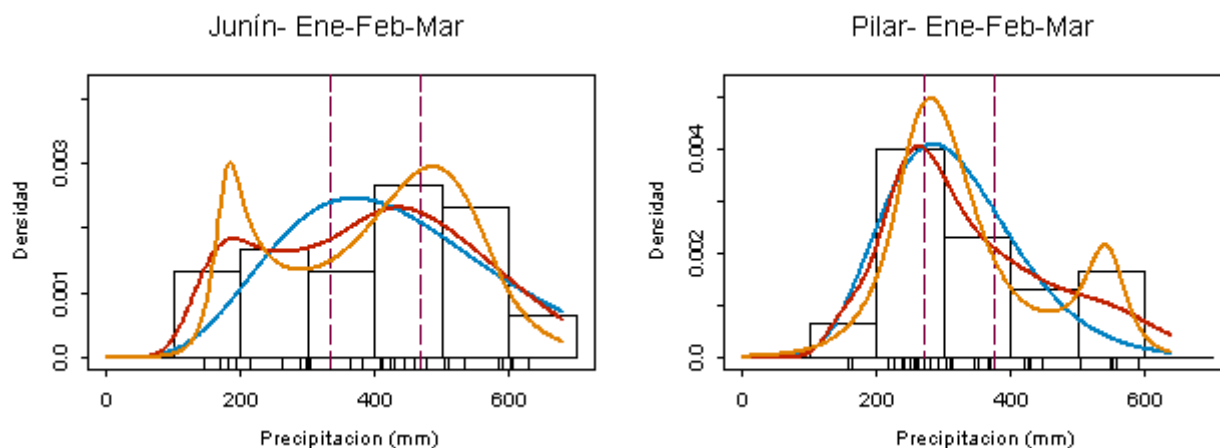
A continuación como ejemplo se anexa una tabla (Tabla 1) con los valores del p-33 calculado con las diferentes metodologías, para el trimestre ene-mar, en algunas estaciones y, dos gráficos mostrando el histograma con los ajustes (Figuras 3 y 4).

Tabla 1. Valores de p-33 (en mm) correspondientes al trimestre ene-mar según distintos ajustes (P_Emp: empírico, P_EmpR: empírico con remuestreo, P_Gam: gamma, P_GamR: gamma con remuestreo, P_K: kernel, P_LS: logspline).

Estación Meteorológica	P_Emp	P_EmpR	P_Gam	P_GamR	P_K	P_LS
Azul	293.3	290.7	293.3	298.3	281.2	304.3
Iguazú	378.1	375.5	376.6	377.3	347.1	387.8
Junín	309.2	333.6	348.2	341.5	310.4	326.3
Las Lomitas	269.5	287.6	296.4	298.3	273.8	270.1
Paraná	310.1	315.5	303.8	307.6	301.1	302.3
Pilar	269.8	271.2	266.6	269.2	266.6	277.4
Resistencia	438.9	422.5	405.7	422.3	403.7	428.6
Rosario	312.7	315.6	302.3	312.4	304.6	322.8
Santa Rosa	197.0	205.5	216.0	216.7	198.3	211.9

Temperatura

En lo que respecta al cálculo de los percentiles de temperatura las diferencias observadas entre los percentiles empíricos y los obtenidos, también empíricamente, pero luego de remuestrear la serie, en general son nulas o de una décima de grado. En cuanto a los intervalos de confianza, los mayores intervalos, correspondieron al p-33 en el período abr-jun, con valores entre 1.0°C y 1.5°C en el centro y norte del área de análisis y entre 0.6°C y 1.0°C en el sur de la misma. Los intervalos más pequeños se observaron en los trimestres de ene-mar y oct-dic con valores, en general, comprendidos entre 0.4°C y 0.8°C. Con respecto al intervalo de confianza del p-66, en el trimestre ene-mar se observaron los intervalos más chicos (0.4°C a 0.8°C) en tanto que en los períodos de abr-jun y jul-sep los mismos fueron superiores, predominando valores entre 0.6°C y 1.0°C. En la Tabla 2 se presentan los resultados del p-33 para ene-mar en algunas de las estacio-



Figuras 3 y 4: Histogramas de precipitación, ene-feb-mar, en Junín y Pilar con los ajustes de densidad empírica de kernel (línea bordo), logspline (línea naranja) y gamma (línea azul). Líneas rayadas verticales indican los p-33 y p-66 calculados empíricamente con remuestreo. Líneas pequeñas en la parte inferior señalan los datos reales.

nes consideradas en este estudio.

Tabla 2. Valores de p-33 (°C) y extremos de un intervalo de confianza de 95% correspondiente al trimestre ene-mar.(P_Emp: empírico, P_EmpR: empírico con remuestreo, IC_2.5: límite del 2.5%, IC_97.5: límite del 97.5%)

Estación Met.	P_Emp	P_EmpR	IC_2.5	IC_97.5
Azul	19.3	19.3	19.0	19.5
Iguazú	24.3	24.3	24.0	24.5
Junín	21.1	21.1	20.9	21.6
Las Lomitas	26.6	26.6	26.3	26.9
Paraná	23.1	23.1	22.7	23.5
Pilar	21.6	21.6	21.3	21.7
Resistencia	25.1	25.2	24.8	25.5
Rosario	22.4	22.5	22.2	22.6
Santa Rosa	21.3	21.4	21.1	21.5

CONCLUSIONES

Las funciones utilizadas para calcular los percentiles de precipitación, han mostrado predominantemente diferencias con respecto a los percentiles empíricos inferiores al 10%. Las excepciones corresponden a casos en los cuales no se logró un buen ajuste debido, por ejemplo, a series con muchos valores bajos, con asimetría negativa, bimodales, entre otras. De los resultados se desprende que utilizar la función gamma es una opción viable para la mayoría de las estaciones consideradas (en casi todos los casos el ajuste fue aceptado), pero al extender la región hacia la zona del NOA, esta opción tiene problemas en los casos de valores trimestrales nulos. Por lo cual las dos opciones sugeridas son el cálculo empírico a partir de la serie remuestreada, el cual se mostró en este escrito, o la densidad logspline. El primer caso, sencillo de aplicar, permite estimar intervalos de confianza y es menos sensible a la ocurrencia de valores de la serie muy altos o bajos. En cuanto a la segunda opción la misma arroja buenos resultados en los casos testeados y se puede utilizar en series con ceros. Además, al ser flexible representa también distribuciones bimodales u otras. Comparándola con la densidad de Kernel la ventaja es que no tiene problema en la elección del ancho de banda y tampoco es necesaria la transformación de los datos para evitar valores negativos. Si bien en general las mayores diferencias correspondieron a la distribución gamma, esto no implica que en todos los casos esta distribución haya sido la que presentó un mayor apartamiento. Cabe señalar que no se encontró un patrón definido en tal sentido. Con respecto a la temperatura la metodología propuesta resultó adecuada con la ventaja de que permite el cálculo de los intervalos de confianza.

REFERENCIAS

- Barros, V. y Rivero, M., 1981. Mapas de probabilidad de precipitación de la zona árida de Chu-but. *Meteorologica*, Vol XII, N°1, págs. 7-18.
- Bowman, A. W. y Adelchi, A., 1997. *Applied Smoothing Techniques for Data analysis*. Clarendon Press – Oxford, 191 p.
- Buizer, J.L., Foster J. y Lund, D., 2000. Global impacts and regional actions: preparing for the 1997–1998 El Niño. *Bulletin of the American Meteorological Society* 81: 2121–2139.
- Charles J. Stone, Hansen, M., Kooperberg, C., y Truong, Y.K., 1997. The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, 25, 1371–1470.
- Duna, P.K., 2001. Bootstrap confidence intervals for predicted rainfall quantiles. *International Journal of Climatology* 21: 89-94.
- Hesterberg, T.C., 1999. Bootstrap Tilting Confidence Intervals. Technical Report No. 84, Research Department, Insightful Corp., 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109, <http://www.insightful.com/Hesterberg>
- Kooperberg y Stone, 1991. A study of logspline density estimation. *Computacional Statistics & Data Analysis* 12, págs 327-347.
- Marazzi, A. y Ruffieux, C., 1996. Implementing M-estimators of the Gamma distribution. In: Rieder H. (Ed.), *Robust Statistics, Data Analysis, and Computer intensive Methods*, Springer Verlag.
- Matyasovszky, 1997. Estimating probability density functions by Kernel techniques. *Időjárás. Quartely Journal of Hungarian Meteorological Service*. Vol. 101, N°1, January-March 1997, págs 17-31.
- Skansi, M y Hordij, H., 2001. Estudio acerca de la presentación de los pronósticos de Tendencias Trimestrales de precipitación. 1999 *Anales Congreso de Agrometeorología*.
- Sheather, S.J. y Jones M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B* 53: 683-690.
- Thom, H.C.S, 1958. A Note on the Gamma Distribution. *Monthly Weather Review*, Vol 86, N°4, págs. 117-122
- Venables, W.N. And Ripley, B.D, 1996. *Modern Applied Statistics with S-Plus*, 462 páginas