Entropy Analysis of Synthetic Time Series

Mario Silber and Susana Vanlesberg

Abstract: One of the characteristics of hydroclimatic variables is its variability, understanding this as a measure of the dispersion of the samples. The most universally measures used to analyze the variability of hydroclimatic variables are the variance, standard deviation, range and, with certain limitations, the coefficient of variability (undefined when the average is close to zero).

The aim of this paper is to attempt to exploit the entropy (as has been defined by Shannon in 1948 as a measure of disorder) of the probability distribution as a measure of variability, associating with other measures of the distribution characteristics of variables (such as skewness) and various indicators of non-stationary.

Using Monte Carlo methods discrete time series have been generated, both stationary and with different types of non-stationary (white noise with hidden fluctuations and trends), drawn from normal and non-normal populations.

The variations of the entropy of all samples were analyzed, in order to relate them with different statistics (variability, skewness) and with different parameters used in the generation and analysis (linear trends wavelengths, lengths class intervals).

Once found the sought relationships, they are used as patterns in the entropy analysis of real time series of hydroclimatic variables such as precipitation with different levels of aggregation. The joint entropy of random variables can also be analyzed, as well as areal distribution.

Keywords: Entropy, simulation, precipitation.

Mario SILBER
Susana VANLESBERG

Facultad de Ingenieria y Ciencias Hidricas
Universidad Nacional del Litoral
Ciuciad Universitaria
CC 217 - Ruta Nacional 168 - Km 472,4
(3000) - Santa Fe - Rep. Argentina
Tel: (54) (342) 457 5243/44 - Int. 160
Fax: (54) (342) 457 5224

Mario SILBER
silber.mario@gmail.com

Susana VANLESBERG
susvan@gmail.com

Received: 19 october 2011 / Accepted: 15 december 2011
Published online: 30 june 2012

Riassunto: Una delle caratteristiche delle variabili idroclimatiche è la loro variabilità, intendendo ciò come la misura della dispersione dei dati. I metodi matematici più utilizzati per analizzare la variabilità delle variabili idroclimatiche sono la varianza, la deviazione standard, l'analisi della gamma e, con alcune limitazioni, il coefficiente di variabilità (non definito quando la media è prossima allo zero).

Scopo di questo lavoro è di tentare di scoprire l'entropia (come è stata definita da Shannon nel 1948 come misura del disordine) della distribuzione della probabilità come misura della variabilità, in associazione con altre misure della distribuzione caratteristica delle variabili (intesa come asimmetria) e con l'uso di vari indicatori di instabilità. Attraverso l'uso di metodi Monte Carlo sono state generate serie discretizzate sul tempo sia su dati stazionari che su dati non-stazionari (rumore bianco con tendenze e fluttuazioni nascoste), tracciate per popolazioni normali e non. È stata analizzata la variazione dell'entropia per tutti i campioni al fine di porli in relazione con differenti statistiche (variabilità, asimmetria) e con differenti parametri utilizzati nella generazione e nell'analisi (tendenze lineari sulle lunghezze d'onda, classi di intervalli di lunghezze).

Una volta trovate le relazioni cercate, queste sono state utilizzate come modelli nell'analisi dell'entropia di serie in tempo reale delle variabili idroclimatiche come le precipitazioni con differenti livelli di aggregazione. Può essere anche analizzata l'entropia congiunta delle variabili casuali, così come la loro distribuzione areale.

Introduction

One of the characteristics of hydroclimatic variables is its variability, understanding this as a measure of the dispersion of the samples. (Castañeda E. and Barros V., 1994), (Penalba O. and Vargas W., 2001).

If what is analyzed is a hydroclimatic time series with samples from a given site, the variability is associated somehow with the stationary and level of aggregation, whereas if what is observed is a random variable distributed in the space, its variability is related to other parameters such as topography, the type of events that generate the variable (in the case of precipitation, for example, it may be convective phenomena, frontal, orographic), the patterns of the general atmospheric, the seasonality of the variable and, among other things, the level of aggregation.

If what is studied is the variability of time series of spatially distributed variables, then a description of the variability in time(in sample points) is needed and then all estimators will be associated in a spatial, geographic interpretation, based on geostatistical procedures, allowing areal definition, globalizing the statistical analysis.

The most universally measures used to analyze the variability of hydroclimatic variables are the variance, standard deviation, range and, with certain limitations, the coefficient of variability (undefined when the average is close to zero).

Hydroclimatic systems are highly complex, both spatially and temporally. They are grounded in an interconnected dynamic network,

DOI 10.4409/Am-038-12-0038

strongly interdependent, and may be characterized as the "output" of chaotic systems with nonlinear dynamics. The variability provides valuable information.

Entropy as a concept was developed by Clausius (1850) in the nineteenth century, when trying to understand the machines behavior He defined the entropy as a measure of energy that can be used to perform work on a given system and postulated that this will always increase over time in an isolated system.

Although this definition is widely used in physics and chemistry, there is another definition, proposed by Boltzmann (Boltzmann L.1872) also in the nineteenth century, which gives a more intuitive concept. The usual presentation of entropy as "disorder" comes from the Boltzmann formulation. This definition said that entropy is the number of microscopic states compatible with a macroscopic state. As the entropy of Clausius, Boltzmann's also increases with time.

There is yet another definition of entropy does not seem to be related to physical entropy, but it is. It's called Shannon entropy (Shannon C.E., 1948). The Shannon entropy is a mathematical concept; it is a measure of information in a system and, therefore, is measured in bits.

This definition may seem so far from physics, but in fact the concept of physical entropy is very close to the information entropy. For example, if a gas is contained in a large volume, randomness, and therefore, uncertainty in the position of each individual molecule is much larger. So, to specify the position of each requires much more information than the one required for a smaller volume.

The relationship between entropy and information is a crucial step in the process of analyzing the variability (or disorder) of random variables.

The aim of this paper is to attempt to exploit the entropy (as has been defined by Shannon in 1948 as a measure of disorder) of the probability distribution as a measure of variability, associating with other measures of the distribution characteristics of variables (such as skewness) and various indicators of non-stationary.

Method

For the development of this study open source software, free and freely available (Gnumeric) was used. In this way, the designed experiments can be reproduced in any usable computer equipment.

Using Monte Carlo methods discrete time series have been generated, both stationary and with different types of non-stationary (white noise with hidden fluctuations and trends), drawn from normal and non-normal populations.

The variations of the entropy of all samples were analyzed, in order to relate them with different statistics (variability, skewness) and with different parameters used in the generation and analysis (linear trends wavelengths, lengths class intervals or bins).

Analyzed cases:

Case 1: Series drawn from a uniformly distributed population

In this case different histograms generated descriptive of the sample have been prepared by varying the binning. In this way as many histograms as proposed bins were obtained.

The histogram is representative of the probability distribution and the entropy associated with it was estimated directly from each histogram.

A sample of 10000 values uniformly distributed in the interval [0, 10000] was generated. Then it was described with 50 histograms each with a different number of intervals (1 to 50). This is shown in Fig. 1.

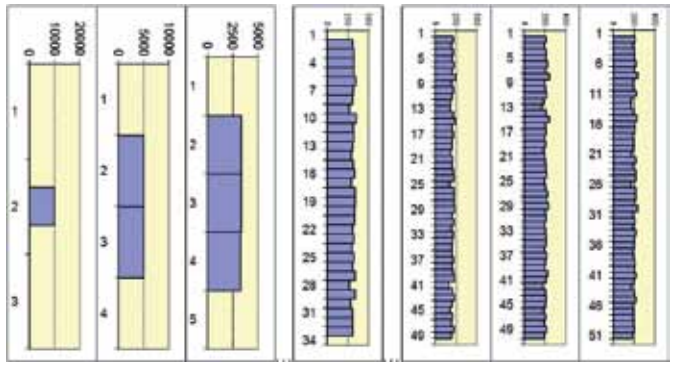

Fig. 1 – *Generated sample histograms*

For each histogram interval, entropy calculated with the expression given by equation (1).

$$H_i = -p_i \ln_2 p_i \quad (1)$$

the sample entropy, expressed in *bits* was estimated by aggregation of the equation (1):

$$H = -\sum_{i=1}^{n} p_i \ln_2 p_i \quad (2)$$

H: sample entropy
p_i = Probability of each interval, estimated thru the relative frequency;
$\ln_2$ = binary logarithm

This can be seen partially in Tab. 1:

From the above process graph of the relationship between the entropy of the sample and the number of intervals was constructed.

Thus a relationship between the bin or class width (or number of classes used) and entropy of the analyzed sample for this variable could be achieved.

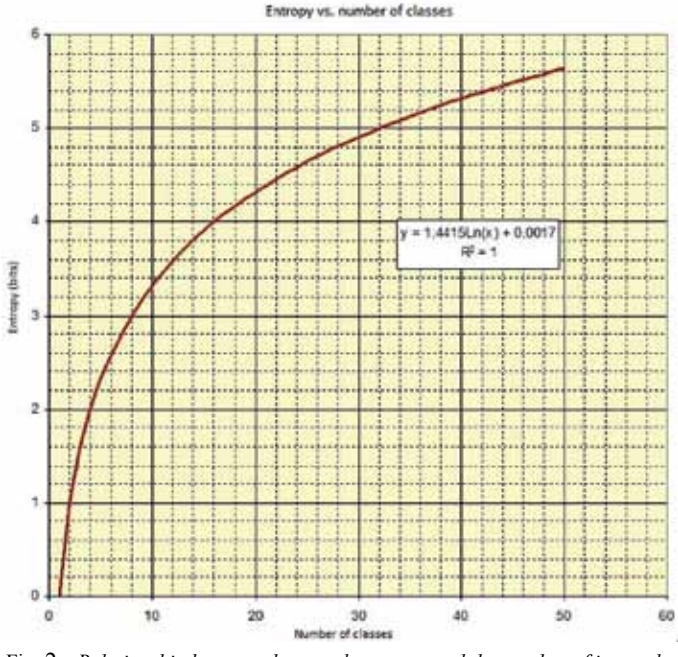

Fig. 2 - *Relationship between the sample entropy and the number of intervals*

Tab. 1: *Estimation of the histogram entropy*

	Number of intervals of the histogram								
	1	2	3	4	5	…	48	49	50
Entropy of each interval Hi	0,0000	0,5007	0,5288	0,4996	0,4696	…	0,4080	0,3770	0,3556
		0,4992	0,5284	0,4994	0,4579	…	0,3983	0,3718	0,3552
			0,5275	0,5035	0,4686	…	0,3952	0,3790	0,3518
				0,4971	0,4658	…	0,4078	0,3693	0,3460
					…	…	…	…	…
						…	…	…	…
						…	…	…	…
						…	0,1094	0,1199	0,1107
						…		0,1077	0,1145
						…			0,1073
Entropy (H)	0,0000	0,9999	1,5848	1,9998	2,3213	…	5,5803	5,6112	5,6396

Case 2: Series drawn from normal populations with different coefficients of variation and described with the same number of classes.

Keeping constant mean and varying standard deviation, 50 samples from normal populations with a wide range of coefficients of variability (between 1 and 50) were generated, but keeping the same number of class intervals (20 intervals). It should be noted that in this case what remains constant is the number of classes and not its width.

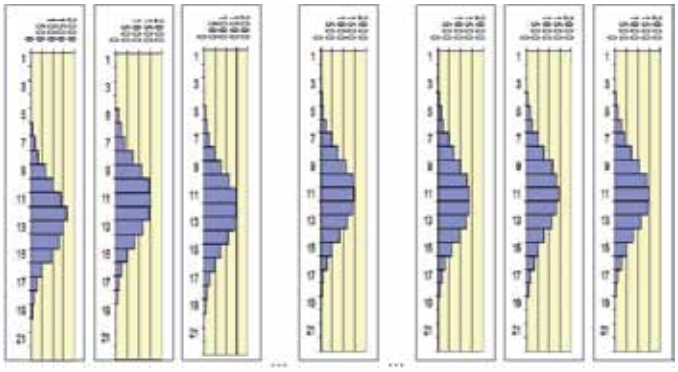

Fig. 3 - *Histograms of the generated sample.*

For each interval of each and every one of the entropy histograms according to Equation (1) was estimated, and was added according to Equation (2) to finally obtain the entropy of the sample.

This can be seen partially in Table 2:

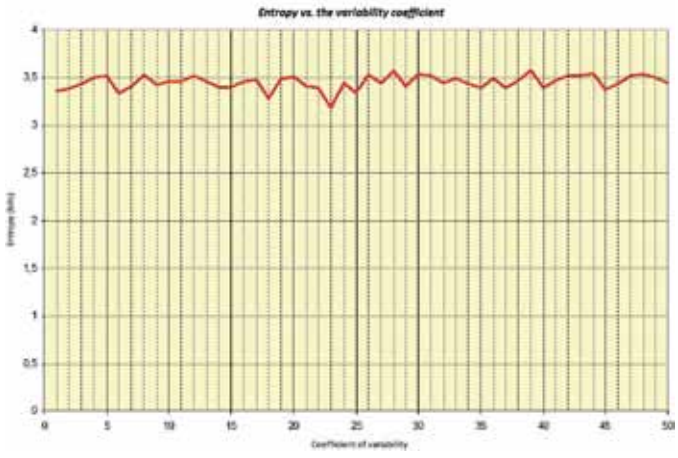

Fig. 4 - *Relationship between the entropy of the sample, the number of intervals and the variability.*

This leads to the conclusion that the entropy is not a function of the variability of the sample (and, therefore, of the variance) if the number of class intervals that attempt to describe the samples *remains constant*. That is, if the samples are analyzed independently, without maintaining the characteristics of the analysis, the variations of entropy *will not be found*.

Tab. 2

	Coefficient of Variation								
	1	2	3	4	5	…	48	49	50
Entropy of each interval Hi	0,0024	0,0035	0,0013	0,0045	0,0073	…	0,0082	0,0064	0,0073
	0,0024	0,0082	0,0064	0,0132	0,0148	…	0,0132	0,0099	0,0099
	0,0091	0,0237	0,0108	0,0258	0,0331	…	0,0272	0,0237	0,0344
	0,0186	0,0643	0,0344	0,0611	0,0674	…	0,0590	0,0518	0,0690
						…	…	…	…
						…	0,0054	0,0054	0,0024
Entropy (H)	3,3605	3,3831	3,4347	3,4994	3,5228	…	3,5366	3,5040	3,4434

DOI 10.4409/Am-038-12-0038

Case 3: Series drawn from normal populations with different coefficients of variability, described by uniform class widths

51 ad-hoc generated samples were used, with coefficients of variation ranging between 1 and 6. Being the mean equal to 10, standard deviations ranged between 10 and 60.

Then, each one of the synthetic samples was ranked through a histogram that kept the same width class in all cases.

The aim was to establish a relationship between entropy and the coefficient of variation, varying the number of classes so that the width class remains constant for each histogram.

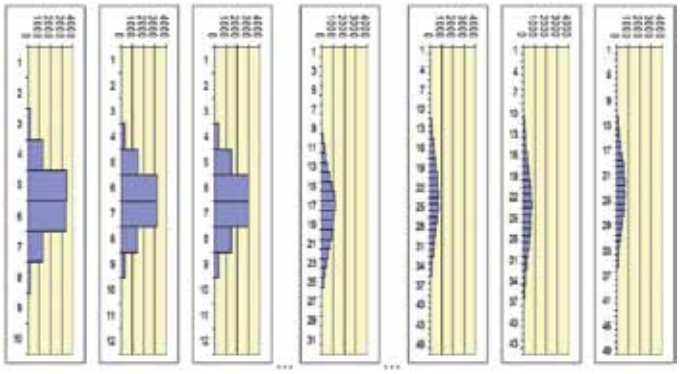

Fig. 5 - *Histograms of the generated sample*

For each interval of each and every one of the histogram entropy was estimated by Eq. (1), and summed by Eq. (2) to finally obtain the entropy of the sample.

This leads to the conclusion that the entropy itself is a function of the sample variability (and thus of the standard deviation) if the class width used to describe the samples *remains constant*. That is, if the samples are analyzed keeping constant the characteristics of the analysis (in this case, the class width), *variations of entropy are not only able to detect, but are a direct function of the coefficient of variability*, possibly with a pattern behavior and therefore a criterion or comparison of samples based on their entropy.

This can be seen in Figure 5, which describes the behavior of the entropy of the samples generated according to the standard deviation.

Case 4: Series drawn from Gamma distributed population, with different coefficients of skewness

The coefficient of skewness of the two parameters Gamma probability distribution is related to the shape parameter of the distribution, making it possible to extract samples from populations with different skewness Gamma-distributed variables simply by varying the shape parameter.

The Gamma density probability function is:

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)} \qquad (3)$$

The skewness is a function of the shape parameter $\gamma_1 = \frac{2}{\sqrt{\alpha}}$.

Then, setting the skewness (for the analyzed example, between 6 and 0.5) is possible to obtain different values of the shape parameter.

Once samples have been generated, and using the same width for each class and the same standard deviation, a relationship between entropy and the skewness of the samples was derived.

In Fig. 6 some histograms of the generated variables histograms are shown, in which, of course, can be seen that the distribution is close to normal as the skewness decreases.

From each of these samples entropy was estimated and associated with skewness

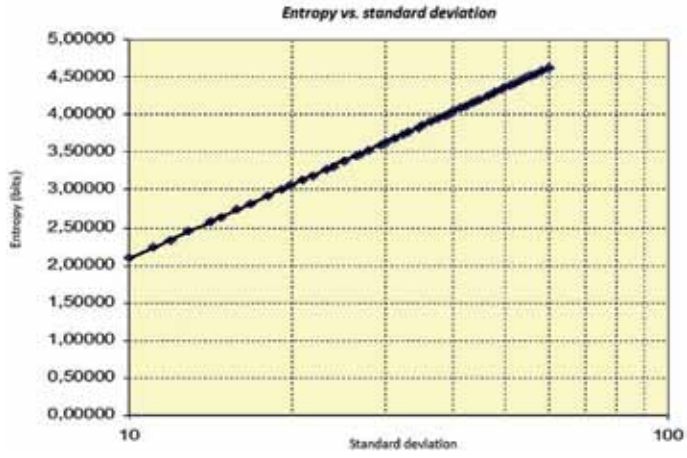

Fig. 6 - *Relationship between entropy and sample standard deviation for class width constant.*

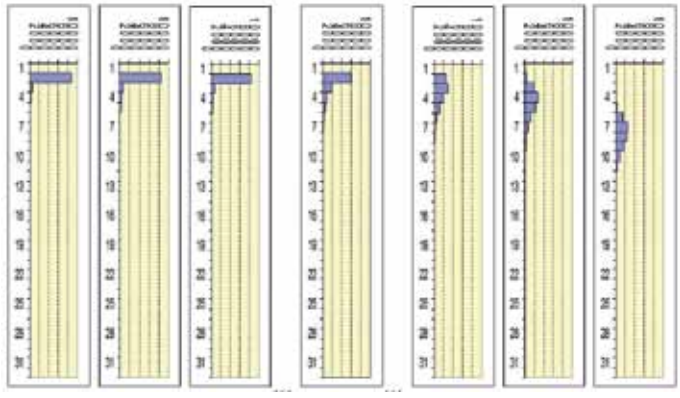

Fig. 7 – *Histogram of the generated gamma samples*

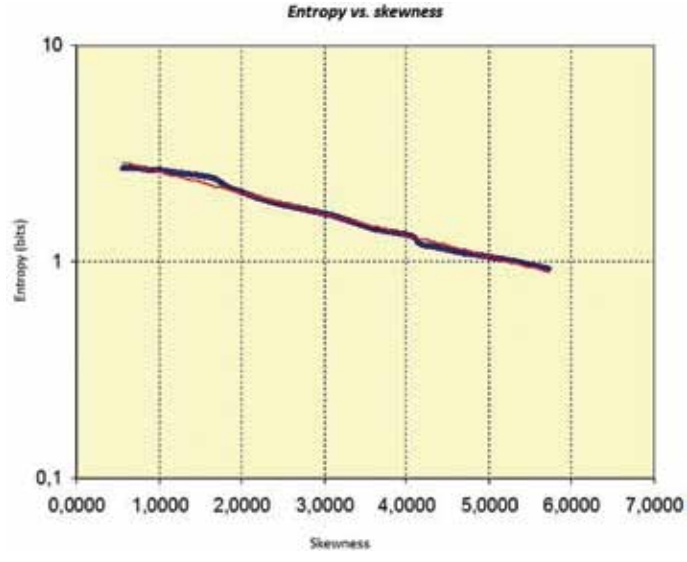

Fig. 8 – *Relationship between skewness and entropy*

Case 5: Series drawn from normal populations, autocorrelated, with different coefficients of first order autocorrelation

Keeping the variability of each sample, and by varying he coefficient of first order autocorrelation, different time series using the methodology of Box & Jenkins was simulated. The time series generating equation is

$$\phi(B)\%_t = a_t \qquad (5)$$

where:

$\phi(B)$ = first order autoregressive operator = $1 - \phi_1 B$

B = backward operator, $Bz_t = z_{t-1}$

$\%_t = z_t - \bar{z}$

a_t = white noise : $N(0, \sigma_a)$

$\phi_1 = \rho_1$ = first order autocorrelation coefficient

To compose the different time series, a unique series of "white noise" was generated, and the autocorrelation coefficient fluctuate between -0.99 and +0.99, thereby covering all the possible range of variation.

The following chart shows the first 100 values of the first three and last three series generated with autocorrelations of 0.99, 0.95, 0.90,-0.90, -0.95, -0.99, plus one with an intermediate value (ρ_1=0.5)

Then a histogram (Fig.10), descriptive of each sample, was prepared, and the entropy of each of them was estimated at constant width classes. This will attempt to establish a relationship between entropy and auto-dependence-or "memory" – of the time series.

In Figure 9 some histograms of the samples generated with the above characteristics can be observed.

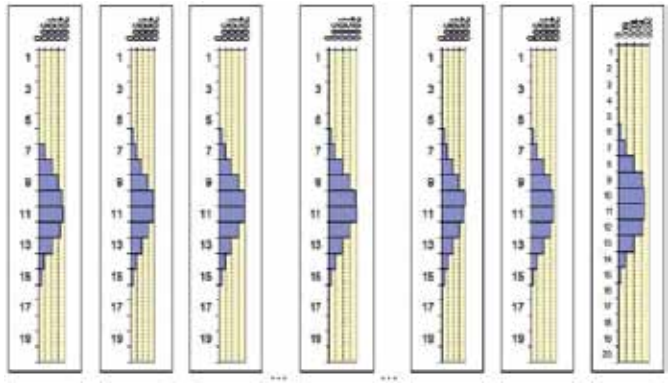

Fig. 10 – *Histograms of the synthetic autocorrelated samples)*

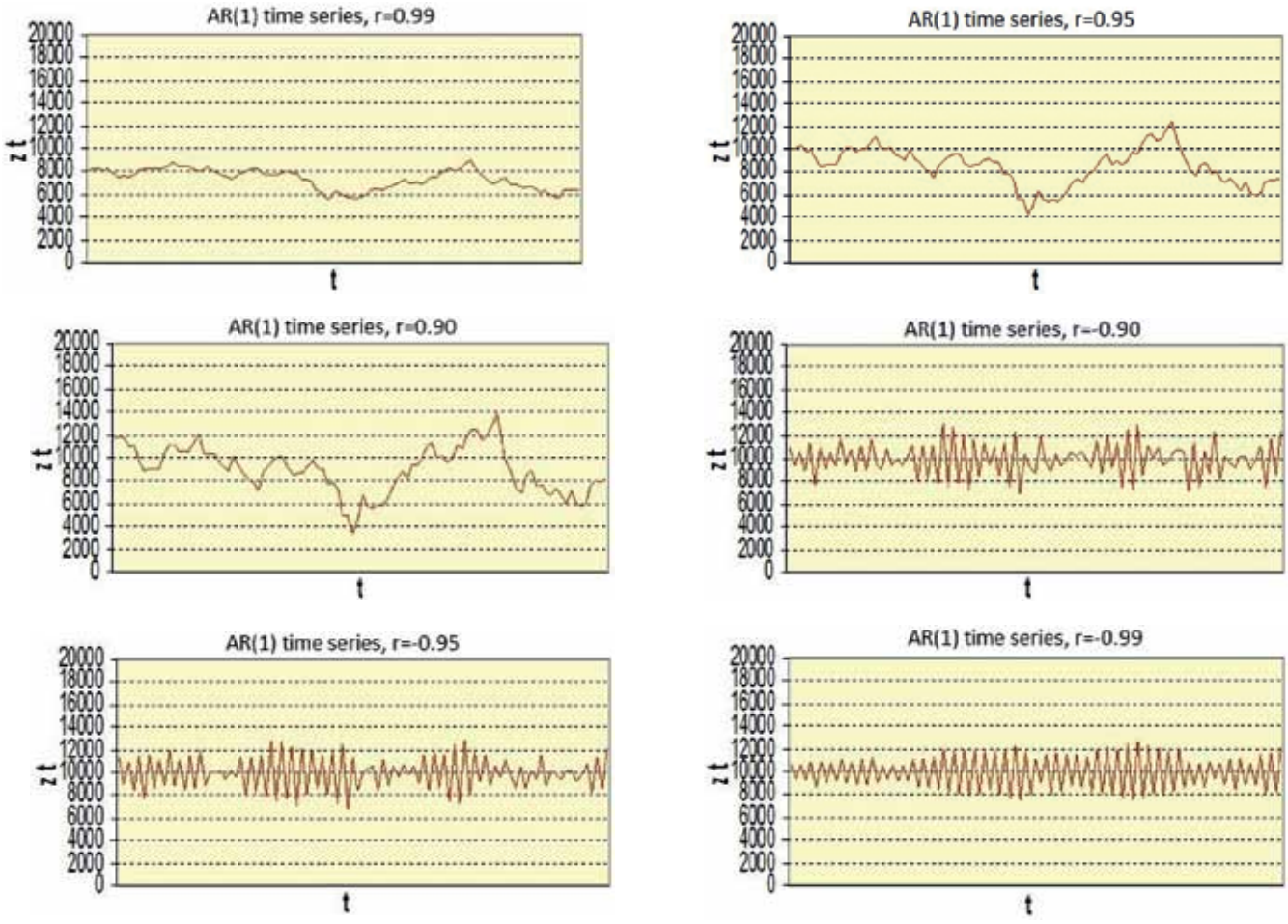

Fig. 9 – *Time series with different correlation coefficients*

DOI 10.4409/Am-038-12-0038

Once calculated the histograms, the entropy of each sample was estimated. This made it possible to relate the entropy with the auto-dependence. Evidence of this is the figure 11

As shown, the entropy is independent of the memory of the series (or its autocorrelation)

Fig. 11 – *Relationship between entropy and the autocorrelation coefficient)*

Case 6-a: Time series with two added components (a linear trend plus a normally distributed variable)

This is the first case of non-stationary analyzed. The samples are drawn from populations that show a linear non-stationary, plus a "white noise". This would imply that a deterministic linear process adds a homoscedastic stochastic process, which resulted in homoscedasticity non-stationary samples. In nature is not easy to find heteroscedastic variables, unless the case - for example - harmonic variables influenced by the phenomenon of resonance.

The entropy of samples from these populations was estimated. The entropy of each of these samples was attempted to estimate, where what remains constant is the "white noise" and the width of each histogram class, and what varies is the slope of the line that sets the trend.

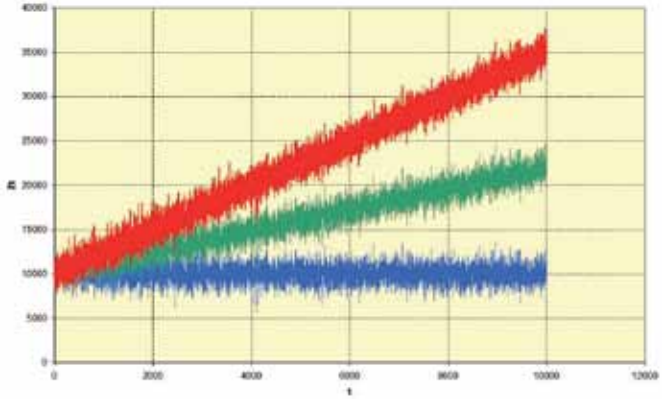

Fig. 12 – *Three time series generated with different trends*

Some of the histograms from which the entropy is estimated can be seen in the figure below.

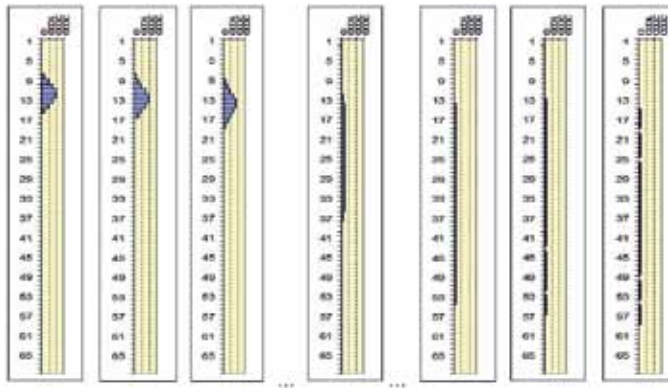

Fig. 13 – *Histograms of series generated with different trends*

It can be seen that the histograms range from a normal one, corresponding to a trend equal to 0, to a nearly uniform, representative of the steepest series. This is logical, since that increasing the trend, the noise are becoming less representative.

The entropy, however, suffers a reverse process: the larger trend, higher entropy. This relationship can be seen in the figure below:

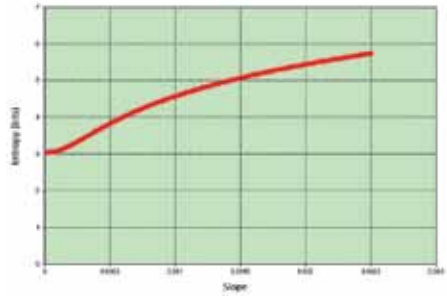

Fig. 14 – *Variation of entropy with a linear trend*

Case 6-b: Time series with two added components (a sine wave plus a normally distributed variable)

This is the second non-stationary case analyzed in this one the samples are drawn from populations composed of a deterministic component (sine or cosine wave) and a random one (white noise). This is a very common case where the variables being analyzed are hydroclimatic ones, most of which show seasonal variations that can be represented often by waves.

The equation from which the samples were generated was the following:

$$y_t = a_1 + a_2 \sin\frac{2\pi x}{\lambda} + a_t \qquad (6)$$

where: y_t = syntetic time series

$a_1 = cte = 10000$ = mean of the series

$a_2 = cte = 2000$ = amplitude

λ=series length

$\frac{2\pi x}{\lambda}$ = periodic frequency

a_t = white noise

The entropy of these samples has been evaluated with constant random component and varying the wave length of the harmonic component, starting from the Nyquist frequency to a reasonably high frequency in relation with the size of the sample. In all cases also remained constant the class width of the histograms from which the entropy was calculated.

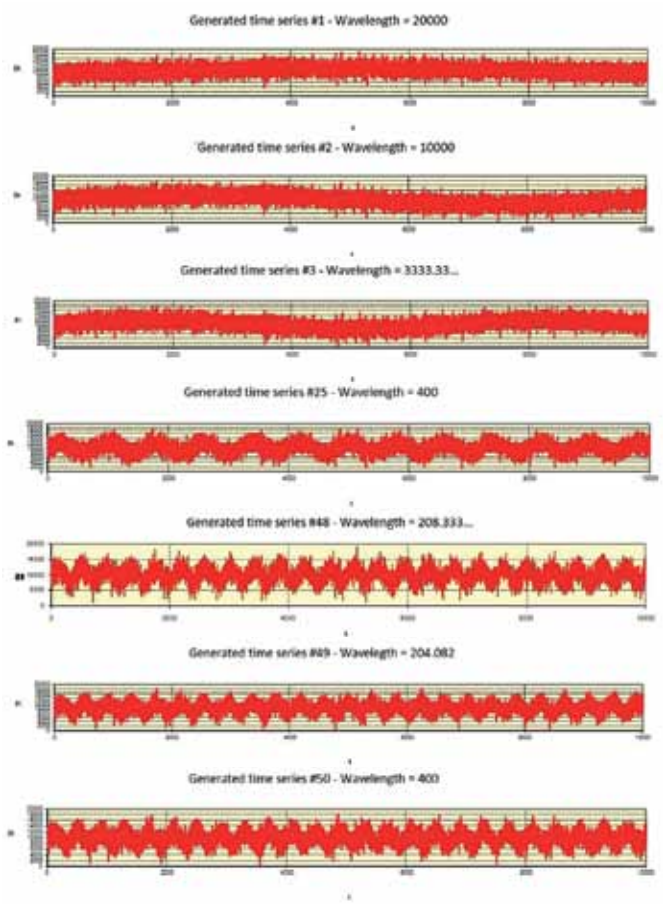

Fig. 15 – *Generated time series*

Then, from each sample, histograms were prepared and the entropy of each one was evaluated, being constant the class width.

The following graph shows some of the resulting histograms.

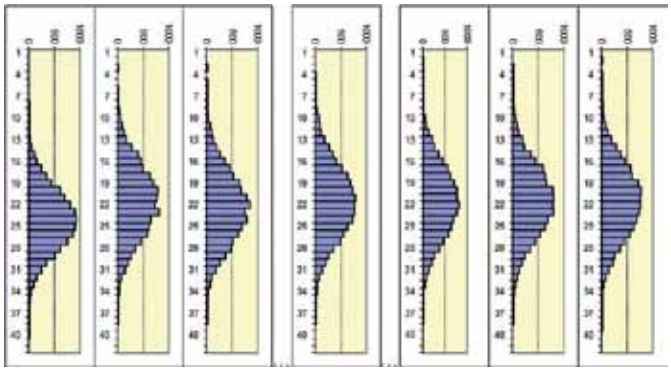

Fig. 16 – *Histograms representing the series*

The entropy of each partition proposed was obtained, and the relationship of each with the lengths of the waves of the generated variables was accomplished as well. This can be seen in the Figure 17.

As seen in this case, in which a deterministic component (sine wave) was added, the entropy remains unchanged. This relationship is one that can be seen in the Figure 17.

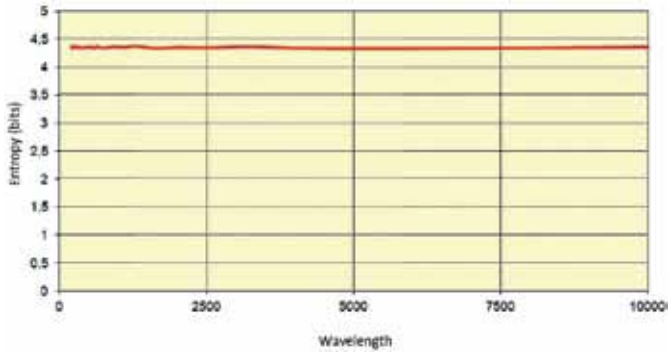

Fig. 16 – *Relation between the entropy of a series and its wavelength*

Results and Conclusions

The tables and graphs that are displayed throughout the work are shown how the entropy (a measure of the information or disorder) varies according to different characteristics of the samples.

It can be seen, for example, that the entropy is *strongly dependent* on the interval class width that produces a histogram representative of the sample, which - apparently -would introduce a subjective factor, unwanted, in its calculation. As a direct consequence, and in order to be able to have comparable results, is concluded that must set the *same class interval to analyze any variable geographically distributed, of whatever the rank, maxima and minima are*. For example, if the entropy of precipitation fields over a wide area is analyzed; all samples should be classified according to a common histogram.

It also can be observed that, contrary to what intuitively is thought, highly dependent samples generated by first order autoregressive schemes do not present lower entropy. That is, the entropy as a measure of disorder, in this case, is useful only if accompanied by another number, such as the autocorrelation coefficient of first order, or the covariance of the sample.

It can be seen that the entropy is highly dependent on the skewness of the samples. The greater is the skewness, the smaller is the entropy. This has been verified through Gamma-distributed variables, common in the universe of hydroclimatic variables.

The cases discussed in the non-stationary variables are striking. When introduced nonstationarity is a linear trend, disorder (entropy) increases as the series is seemingly more organized (as he slope is more significant), whereas when introduced nonstationarity is a periodic variable, the entropy is independent of frequency or frequency remains constant whatever the wavelength was.

Once the sought relationships established, they will be used as standards in the analysis of the entropy of real hydroclimatic time series hydroclimatic, such as precipitation with different levels of aggregation. The joint entropy of random variables and its areal distribution can be analyzed.

Throughout the work can be seen that the exploration of the entropy of synthetic series as indicative of the variability and its relationship and / or dependence on other features (such as skewness, nonstationarity, auto-dependence, type of distribution) opens pathways which can be very useful for characterizing real random variables.

DOI 10.4409/Am-038-12-0038

References

Ben-Naim Arieh (2007) Entropy demystified: the second law reduced to plain common sense, Ed. Hackensack: World Scientific.

Boltzmann Ludwig (1872). Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. Sitzungsberichte, der Akademie der Wissenschaften 66 (1872), 275–370. Translation: Further studies on the thermal equilibrium of gas molecules, in Kinetic Theory 2, 88–174, Ed. S.G. Brush, Pergamon, Oxford (1966).

Brillinger D. L. (2002) Second-order moments and mutual information in the analysis of time series, Recent Advances in Statistical Methods, Imperial College, London

Castañeda E. y Barros, V. (1994). Las tendencias de la precipitación en el Cono sur de América al este de los Andes, Meteorológica 19, 23–32.

Chapeau-Bolndeau F. (2007) Autocorrelation versus entropy-based autoinformation for measuring dependence in random signal, Physica A, 380.

Clausius R. (1850). *Annalen der Physik und Chemie* **79**: pp. 368-397, 500-524

Fazlollah M. Reza (1994) An Introduction to Information Theory, Dover Pub. Inc., New York.

Gray R. M. (2007) Entropy and Information Theory, Stanford Uni Press.

Kawachi, T., Maruyama, T., Singh, V. P. (2001) Rainfall entropy for delineation of water resources zones in Japan, J. Hydrol. 246, 36–44.

Kraskov A.; Stöghauser H. & Grassberger Estimating mutual information, Physical Review E 69: 066138

Lazo A. & Rathie, P.(1978) On the entropy of continuous probability distributions - Information Theory IEEE Transactions, 24(1)

Machta J. (1989) Entropy, information and computation, American Journal of Physics, 67.

Park Sung Y.; Bera, Anil K. (2011) Maximum entropy autoregressive conditional heteroskedasticity model, Journal of Econometrics (Elsevier)

Penalba O. y Vargas, W. (2001). Propiedades de precipitaciones extremas en zonas agropecuarias argentinas, Meteorol., 26, 39-55.

Shannon C. E. (1948) A Mathematical Theory of Communication Bell System, Technical Journal, 27.

Thomas M. & Thomas, U. J. (1978) Elements of Information Theory, Wiley, New York.

Zurek W. H. (1989) Algorithmic randomness and physical entropy, Phys. Rev. A40.