Universidad de Cantabria

Facultad
de
Ciencias

TEORÍA DE LA INFORMACIÓN APLICADA A SISTEMAS COMPLEJOS: UNA APLICACIÓN PARA EL ESTUDIO DE PLASMAS
(Information Theory Applied to Complex Systems: An Application for the Study of Plasmas)

Trabajo de Fin de Grado
para acceder al

GRADO EN FÍSICA

Autor: Iker León Rivas

Director: Juan Manuel López Martín

Codirector: José Ángel Mier Maza

Junio - 2024

Resumen

Lo primero de todo cabe destacar que todos los resultados tanto teóricos como computacionales de este trabajo son originales salvo la Sección 2.2 que se usa como base para construir sobre ella los resultados teóricos adaptados al trabajo. El procedimiento que se ha seguido es el de demostrar todos los resultados teóricos originales del trabajo y no demostrar ninguno de los no originales, indicando por supuesto en qué referencia se puede ver el resultado.

El objetivo de este trabajo es el desarrollo de una técnica basada en la teoría de la información para que un conjunto de series temporales o procesos estocásticos puedan ser ordenadas en función de su grado de correlación, con el foco puesto en las series temporales provenientes de un sistema complejo como un plasma. En [1] ya se intentó realizar esto pero debido a limitaciones el método no era capaz de detectar correlaciones a medio o largo plazo, como pueden ser las de una señal de plasma. En este trabajo se resuelve este problema usando una vía alternativa.

Esto se ha llevado a cabo en 3 fases. Primero, en el Capítulo 2 se han introducido las bases teóricas y a través de una estructura matemático-deductiva se ha concluido cómo debe de ser el método desde un punto de vista teórico. En un primer instante, en la Sección 2.2 se ha descrito la teoría clásica existente en la literatura, que es la concerniente a la entropía de bloque. En las siguientes secciones, no obstante, se han deducido los problemas que tiene esta teoría al llevarla a la práctica, a partir de lo cual se ha decidido usar un método alternativo no usado en la literatura en la forma que se usa en este texto, que es la entropía de permutación. Por ello se ha tenido que adaptar el marco teórico existente de la entropía de bloque a la entropía de permutación, lo cual se ha llevado a cabo en la Sección 2.5 . Una vez hecho esto se ha podido deducir la forma teórica óptima del método que se ha dado en la Sección 2.6.

Por último, en los Capítulos 3 y 4 se ha realizado la prueba del método. En primer lugar, en el Capítulo 3 se ha probado con datos simulados provenientes de la aplicación logística y después, en el Capítulo 4 se ha probado con datos reales provenientes de las señales temporales del plasma de la Santander Linear Plasma Machine (SLPM). En ambos casos los resultados han sido compatibles con el comportamiento esperado dado por la literatura. Los resultados del texto se sintetizan en las Figuras 3.5 y 4.6, que son la representación de la complejidad para las señales del logístico y del plasma respectivamente.

Palabras clave : Entropía, permutación, estocástico, complejidad, correlación, plasma.

Abstract

First of all it is noted that all results both theoretical and computational are original except for the Section 2.2 which has been included as a basis to construct the required theoretical foundations of the work over it. The procedure that has been followed is to prove all original results and none of the non original, giving of course the reference where the proof may be found.

The main goal of this work is the development of a technique based on information theory for classifying time series or stochastic processes according to their correlation, focusing on time series generated by plasmas. In [1] it was attempted to carry this out but due to limitations of the method used there, it wasn't capable of detecting correlations at medium and high temporal scales, as are the ones that usually appear in plasmas. In this work we solve these limitations by following an alternative way.

This has been carried out in 3 steps. First, in Chapter 2 the theoretical basis has been introduced and through a logical structure it has been concluded how the method must be from a theoretical point of view. Initially, in Section 2.2 the classical theoretical basis has been described, which is the one constructed for block entropy. In the next sections, however, the problems that arise when trying to apply this block entropy to practical situations have been shown, and as a consequence an alternative technique from information theory has been proposed, the permutation entropy. Consequently, we have had to adapt the theory thought for block entropy for it to be valid also for permutation entropy, which has been carried out in Section 2.5. Once done this, in Section 2.6, the optimal shape of the method has been deduced from a theoretical point of view.

Finally, in Chapters 3 and 4 the method has been tested. First, in Chapter 3 the method has been tested on simulated data generated by the logistic map and then, in Chapter 4 the method has been tested on real data coming from flux measurements of the plasma generated in the Santander Linear Plasma Machine (SLPM). In both cases the results have been coherent with the behaviour expected based on literature. The results of the text are synthesized in Figures 3.5 and 4.6, which show the complexity of the different time series coming from the logistic map and the plasma.

Keywords : Entropy, permutation, stochastic, complexity, correlation, plasma.

Agradecimientos

En primer lugar, quisiera agradecer a Juanma y José Ángel por haberme propuesto este tema, así como por su guía e indicaciones que tan necesarias han sido para que este trabajo tenga el contenido y la forma que tiene.
Por otro lado, quiero expresar mi gratitud a mi familia y amigos por haberme acompañado durante estos años. También agradecer a todos los profesores que han permitido mi desarrollo tanto académico como personal.
Por último, me gustaría destacar que la presente memoria ha sido realizada fruto del trabajo llevado a cabo en el Departamento de Física Aplicada gracias a la concesión de la Beca de Colaboración por parte del Ministerio de Educación y Formación Profesional, por la que me siento tremendamente agradecido.

Índice general

Capítulo 1

Introducción

1.1. La fusión nuclear: Una energía limpia e ilimitada

Hoy en día, en los países occidentales donde se puede afirmar que se ha erradicado el hambre, dos de los principales problemas o preocupaciones son, por un lado, el cambio climático y por otro, los costes de la energía. Esto último sobre todo a partir de la invasión de Ucrania por parte de Rusia, lo que conllevó una disminución notable del gas natural y combustibles fósiles exportados por este país a los países europeos, lo que causó un aumento considerable de los costes de la electricidad, calefacción y gasolina que afectó directamente a los consumidores. Cualquier persona vería la resolución de alguno de estos problemas como un gran éxito. La fusión nuclear puede resolver ambos problemas [2].

La fusión nuclear es la forma a través de la cual se produce energía en el sol. Debido a las grandes presiones reinantes dentro de nuestro astro, la repulsión electrostática entre dos átomos de hidrógeno cargados positivamente se puede superar y así se puede dar la fusión de ellos, liberando una gran cantidad de energía en el proceso, ya que dos átomos de hidrógeno por separado tienen una masa mayor que un átomo de helio. De hecho, usando la ecuación de Einstein $E = mc^2$ se puede calcular que con 1 kg de hidrógeno se podrían generar 7.5×10^{14} Joules de energía, lo que bastaría para suplir el actual consumo global total de energía durante un periodo de 10000 años [2].

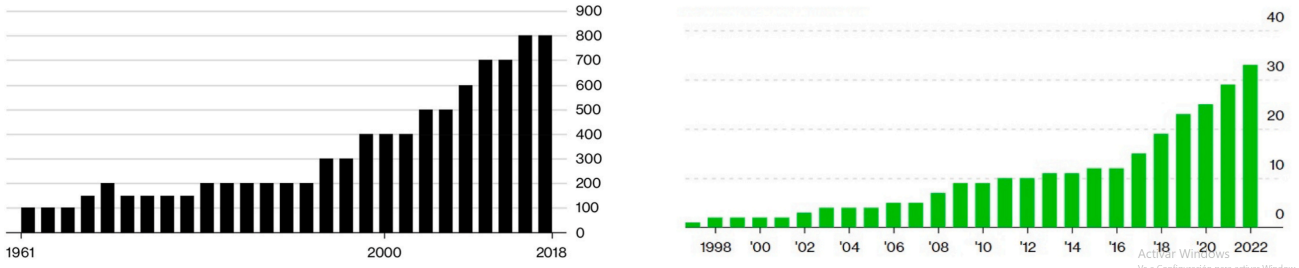
Ya desde que en la década de 1920 el físico inglés Eddington conjeturara que la fusión de átomos de hidrógeno era la forma en la que el sol obtenía energía, surgió el interés en el mundo científico de algún día poder conseguir llevar a cabo esta fusión nuclear en la Tierra [2].

Uno de los grandes avances hacia este objetivo se dio en el año 1956 cuando los científicos soviéticos Andrei Sakharov e Igor Tamm inventaron el Tokamak permitiendo así confinar partículas a altas temperaturas mediante campos magnéticos. No obstante, desde ese año hasta la fecha no se ha conseguido el objetivo de producir más energía de la que se emplea [2]. Las principal razón de ello es la imposibilidad de alcanzar un valor suficientemente alto del triple producto $nT\tau$, donde T es la temperatura del plasma, n es su densidad y τ su tiempo de confinamiento, es decir, el tiempo que el plasma pasa confinado a través del campo magnético.

Una forma de facilitar esto es fusionando átomos de deuterio y tritio en vez de átomos de hidrógeno. Además, como el deuterio es fácilmente extraíble del agua por lo que se dispone de una fuente casi ilimitada y el tritio se podría generar en la propia central de fusión nuclear, no habría problemas de desabastecimiento. En este caso la reacción entre deuterio y tritio genera un núcleo de helio y un neutrón que no es afectado por el campo magnético y así su energía cinética puede ser colectada en las paredes del recipiente donde se confina el plasma transformándola en energía

térmica. Como los neutrones llevan la mayor parte de la energía cinética, la eficiencia de estas reacciones podría ser tan alta como del 90 % [2].

No obstante, pese a todos los beneficios potenciales de la fusión en la sociedad actual hay un sentimiento no muy positivo en torno a ella [2] y es que se considera a la fusión algo como que siempre se espera que se consiga en un periodo de 30 años, pero luego nunca se consigue, y es que desde que por primera vez en la década de 1960 ya se viera la fusión cerca, debido a dificultades técnicas se ha ido posponiendo y han pasado casi 70 años y se sigue sin conseguir. No obstante, parece que el paradigma está cambiando y es que en los últimos años la fusión nuclear ha tomado un especial interés en el ámbito científico y empresarial, para ello basta ver las dos siguientes figuras.



(a) (b)

Figura 1.1: Representación del número de documentos publicados sobre la fusión por año en la subfigura (a) y número de *startups* de fusión que en el periodo 1998-2022 han recaudado 4700 millones de $ en la subfigura (b) [2].

De hecho, en el año 2022 se consiguió por primera vez en la historia un hito, por primera vez se consiguió producir más energía de la suministrada. Fue en el *Federal Livermore Laboratory* (NLC) de California donde suministrando 2.1 MJ de energía, se produjeron 2.5 MJ de energía. Y el año pasado, 2023, consiguieron superar su propia marca y produjeron 3.1 MJ de energía a partir de 2 MJ de energía. Se establece el límite para poder construir reactores de fusión comerciales de manera sistemática el alcanzar la barrera de producir 10 veces la energía suministrada. Sin embargo, con las perspectivas de que el experimento ITER se pueda poner en marcha en los próximos años se espera poder conseguir reactores de fusión nuclear para finales de la década de 2040 o principios de la década de 2050 [2].

La construcción de reactores de fusión nuclear supondría una autentica revolución y es que por un lado eliminaría los problemas de los combustibles fósiles que es que hay una cantidad finita de ellos, mientras que hay reservas del combustible necesario (deuterio y tritio) para al menos 20000 años a un ritmo de producción energético actual y lo más importante, no se producirían emisiones contaminantes a la atmósfera, lo que contribuiría a frenar el cambio climático.

Además, también eliminaría los problemas que tienen las energías verdes actuales como las energías solar o eólica entre otras, ya que estas centrales tienen un impacto medioambiental considerable en nuestros ríos, montes y bosques al ocupar un espacio que pertenece a la biodiversidad de la zona o que incluso podría usarse para la agricultura. Asimismo, la energía por fusión nuclear no tendría el problema de que es intermitente y depende de si es invierno o no, de si está nublado o no, de si hay viento o no, de si ha llovido o no..., es decir, es una fuente energética fiable [2].

Por último, también es preferible frente a las centrales de fisión nuclear debido a que con la fusión se genera más energía que con la fisión, pero sobre todo porque no genera residuos radioactivos de larga vida. De hecho, el tritio es el único elemento radioactivo que se generaría, pero su vida media es de aproximadamente 12 años, que no es ni comparable con los miles de años de vida media de los residuos de la fisión del U-235 [2].

1.2. Santander Lineal Plasma Machine (SLPM)

La forma de estudiar las propiedades del plasma en Santander es a través de la *Santander Linear Plasma Machine* (SLPM). Se trata de una máquina lineal de plasma frío no confinante. Se dice no confinante cuando no se forman superficies magnéticas, pero sí que existe un campo magnético para confinar las partículas axialmente. Maquinas confinantes por otro lado son por ejemplo los *Tokamaks*. La gran diferencia entre los no confinantes y los confinantes es que en los primeros no se puede dar la fusión nuclear.

Por ende se puede deducir que las propiedades macroscópicas de los plasmas generados en la máquina lineal de Santander son diferentes a las de los plasmas calientes generados en los mencionados *Tokamaks* o los *Stellarators* que a día de hoy son los dos dispositivos que mejores perspectivas dan para la consecución de la fusión nuclear [3]. No obstante, por un lado, las propiedades de densidad y temperatura son muy similares y recordemos que una condición necesaria para conseguir la fusión era obtener un valor suficiente del triple producto de fusión. Por otro lado, los fenómenos de turbulencia que tan importantes son como más adelante se explicará también son muy similares en este tipo de plasmas, lo que sumado a que la menor complejidad del dispositivo hace que sea más fácil obtener mejor estadística y reproducibilidad. Todo ello hace que el estudio del plasma en este tipo de dispositivos lineales sea interesante [3].

A continuación se muestra la disposición esquemática de la SLPM en la siguiente figura.

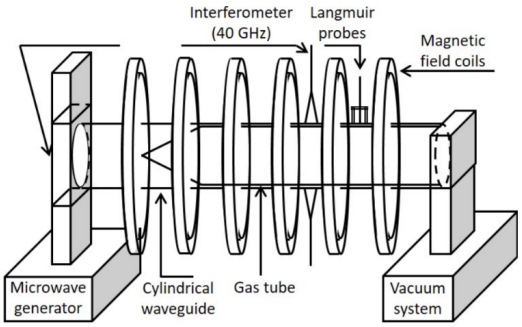


Figura 1.2: Representación esquemática del SLPM tomada de la referencia [4].

Lo primero de todo cabe resaltar que la SLPM es una máquina en la cual el plasma se genera mediante la inyección axial de microondas que trabaja en régimen continuo. La descarga de microondas se realiza durante 4 segundos y a los 2 segundos ya se adquieren los datos mediante una sonda de Langmuir (véase en la Figura 1.2 las *Langmuir probes*) [3], que permiten la medición de la densidad y el potencial eléctrico de un plasma, que permiten deducir los flujos de partículas, y por tanto detectar fenómenos de turbulencia, que se caracterizan por un flujo anormalmente alto en una región dada del espacio.

Otra diferencia respecto a los dispositivos mencionados en la subsección previa es que la SLPM usa gas helio en vez de deuterio y tritio. Este plasma se produce en un recipiente cilíndrico de 7 cm de diámetro y 100 cm de longitud. El helio se inyecta mediante una válvula electromagnética y se trabaja con presiones en torno a 0.001-0.1 mbar[3]. Las microondas que generan el plasma son emitidas por un magnetrón que es un dispositivo que transforma la energía eléctrica en electromagnética en forma de microondas, cuya potencia está en el rango 0.06-6KW y su frecuencia es de 2.45 GHz sobre el gas que puede ser helio pero también neón o argón [3]. Este plasma generado está sometido a un campo magnético lineal o axial constante que está en el rango de 50-140 mT, el cual es generado por 6 bobinas (en la Figura 1.2 las *magnetic field coils*) con corriente máxima de 200 A. Seguido del

magnetrón que en la Figura 1.2 aparece en la izquierda del todo con el nombre *microwave generator* aparece un circulador o *cylindrical waveguide* que sirve para evitar que las microondas reboten de vuelta al magnetrón y así se impide que se dañe.

Finalmente de la Figura 1.2 queda por explicar la función del interferómetro. El interferómetro es un instrumento óptico que emplea las interferencias de las ondas de luz para medir la longitud de onda de la misma luz y en este caso, se utiliza para calcular la densidad media del plasma.

1.3. Teoría de la Información Aplicada a la Fusión

El principal problema para la consecución de la fusión nuclear en la Tierra es que los plasmas son muy impredecibles, lo que dificulta en gran medida la obtención de un valor del triple producto $nT\tau$ suficientemente elevado. No obstante, si se conociera con más detalle por qué el plasma se comporta como se comporta nos sería más fácil entender cómo obtener valores del triple producto de fusión altos, ya que es de sentido común afirmar que en cualquier ámbito de la vida si no entendemos realmente cuál es nuestro problema, nunca lo vamos a poder solucionar.

Aquí es donde entra en juego la teoría de la información. La teoría de la información permite describir mediante técnicas estadísticas el comportamiento de los plasmas. En [5] se afirma que aplicar la teoría de la información a plasmas es una técnica novedosa y muy efectiva para caracterizar estos procesos no lineales tanto desde un punto de vista fenomenológico como desde un punto de vista conceptual.

De hecho, en [5] se afirma que la información como tal es un concepto físico que toma gran relevancia en sistemas complejos como los plasmas donde hay un flujo constante tanto de energía como de información. Es por ello que un conocimiento de cuáles son las propiedades de estos transportes tanto de energía como de información, ayudarían en un futuro próximo a la construcción de *Tokamaks* o *Stellarators* adecuados, permitiendo así la deseada fusión nuclear de una vez por todas y así poder resolver muchos de los mayores problemas que tiene la sociedad actual.

Capítulo 2

Desarrollo teórico del método

En este capítulo se deduce cuál es el método óptimo de detección de correlaciones basado en el concepto de entropía de permutación. Primero, en la Sección 2.2 se introducen las bases teóricas existentes para la entropía de bloque para señales temporales de longitud infinita. En la Sección 2.3 se analiza el caso de series de longitud finita y en la Sección 2.4 se describen los problemas que surgen al llevar a la práctica el concepto de la entropía de bloque y se proporciona una definición alternativa: la entropía de permutación. En la Sección 2.5 se proporciona el marco teórico propio para la entropía de permutación. Finalmente, en la Sección 2.6 se deduce cuál debe ser la forma óptima del método, basados en lo deducido en la Sección 2.5.

2.1. Introducción a la Teoría de la Información

A día de hoy, el concepto de información es un concepto muy extendido en nuestra sociedad y cultura popular, no hay más que recordar que a la época en la que vivimos se le suele denominar como era de la información.

El nacimiento de esta era se puede datar a finales del siglo XIX, con la aparición de la prensa escrita y la comunicación por cable (telegráfo y teléfono) seguidos en el siglo XX por la aparición de medios de comunicación de masas como la radio y la televisión, y finalmente la informática y el internet, que propiciaron una sociedad basada en el conocimiento (y paralelamente, una economía del conocimiento).

Por tanto, el concepto de información es un concepto que está en el lenguaje común. No obstante, para que un concepto se pueda utilizar en el ámbito científico debe definirse de una manera rigurosa. Y no fue hasta finales de la década de los años 20 del siglo pasado cuando se empezó a trabajar sobre el concepto de una idea rigurosa de información.

En 1928, Ralph Hartley probó que el contenido de información de un mensaje de longitud n elegido de un alfabeto de longitud L depende, por un lado, de la longitud del mensaje, n y por otro, de cierta función f que es dependiente de la longitud del alfabeto. No obstante, esta función debe satisfacer que si cierto mensaje tiene longitud n en el alfabeto de longitud L y en el alfabeto de longitud L' tiene longitud n', pero es el mismo mensaje, entonces se debería satisfacer que $I_H = nf(L) = n'f(L')$ [6]. Además, también probó que esta función que preserva el contenido de información entre alfabetos de diferentes longitudes es $f = \log_K$ donde K es una base arbitraria del logaritmo que resulta en escalar de manera diferente el contenido en información. Usamos base 2, que es la entropía de Shannon y a partir de ahora usamos la notación $\log$ para referirnos a $\log_2$. Así,

se define el contenido en información de cierto mensaje como [6]:

$$I_H(n, L) = n \cdot \log(L), \tag{2.1}$$

de donde se puede definir la entropía por símbolo como

$$I_H(L) = \log(L). \tag{2.2}$$

Sin embargo, hay un problema principal con este enfoque, que Shannon identificó y resolvió. El problema principal viene de que en el enfoque de Hartley se asume que todos los símbolos del alfabeto de donde se construyen los mensajes tienen las mismas probabilidades [6]. No obstante, en la mayoría de los casos prácticos este no es el caso.

Shannon tomó el siguiente enfoque. Primero consideró las probabilidades a priori de cada símbolo. Así, descubrió una forma más razonable de definir la ganancia de información de un símbolo que tiene una probabilidad a priori p [6]:

$$I(p) = \log \frac{1}{p}. \tag{2.3}$$

De esta definición, es fácil deducir que si la probabilidad a priori del símbolo es alta, por ejemplo, $p = 1$, entonces, la ganancia de información cuando se saca ese símbolo es muy pequeña, siendo el límite el de que para $p = 1$ se obtiene $I = 0$. Por otro lado, si la probabilidad a priori de un símbolo es pequeña, entonces $\frac{1}{p}$ es grande y por tanto, $\log \frac{1}{p}$ también es grande lo que implica que la ganancia de información cuando se observa el símbolo es elevada.

Observación 2.1.1. *Cuando se toma que todos los símbolos tengan la misma probabilidad se recupera la definición de Hartley a partir de la de Shannon como tiene que suceder, es decir, tomando $p = \frac{1}{L}$ obtenemos $I(p) = \log L = I_H(L)$.*

La descripción previa se ha hecho sobre símbolos en un alfabeto $\mathscr{X}$, pero se puede generalizar fácilmente a cualquier variable aleatoria discreta. Se hace mediante la siguiente definición:

Definición 2.1.1. [6] *Sea X una variable aleatoria discreta definida el espacio probabilístico $(\mathscr{X}, \sigma, \mathbb{P})$, se define la ganancia de información cuando se extrae un símbolo x del alfabeto $\mathscr{X}$ como:*

$$I_X(x) = \log\left(\frac{1}{\mathbb{P}_X(x)}\right). \tag{2.4}$$

Antes de nada es importante notar que este concepto de información parece diferir del concepto que tenemos en nuestra sociedad de información y es por ello que, a través del siguiente ejemplo, se ilustra lo que realmente significa el concepto de información introducido en la Definición 2.1.1.

Ejemplo 2.1.1. *Supongamos que tenemos dos cajas diferentes. La primera caja tiene 3 bolas rojas, 3 azules y 3 negras, mientras que la segunda caja tiene 9 bolas rojas, 0 bolas azules y 0 bolas negras.*

Siguiendo el concepto de información presente en la sociedad uno pensaría que la segunda caja tiene mayor información que la primera, ya que si queremos sacar una bola de cada una de las dos cajas, mientras que en la primera caja no sabemos con seguridad de qué color la vamos a sacar, en la segunda caja sabemos con seguridad que vamos a sacar una bola roja y por ende, concluiríamos que tenemos mucho más conocimiento sobre la segunda caja que sobre la primera y así concluir que la segunda caja tiene mucha mayor información que la primera.

No obstante, vayamos a la Definición 2.1.1 y saquemos una bola de cada una de las dos cajas. Cuando saquemos una bola de la primera caja, obtendremos un valor de información de $I = \log(3)$ ya que los 3 colores de bolas tienen probabilidad $p = \frac{1}{3}$. Por otro lado, cuando saquemos una bola de

la segunda caja será una bola roja y como esta tiene probabilidad $p = 1$, entonces $I = \log(1) = 0$, de donde concluimos que de la primera caja extraemos más información que de la segunda. De hecho, de la segunda no extraemos ninguna información.

Por lo tanto, lo que se debe concluir de este ejemplo es que el concepto de información de un sistema de la teoría de información no es el conocimiento de un observador externo sobre ese sistema, sino que es la información que un observador externo puede extraer de un sistema al realizar una medición sobre él.

De hecho, ambos conceptos son opuestos, ya que a mayor conocimiento de un observador externo sobre el sistema, menos información puede extraer de él y a menor conocimiento del sistema, más información puede extraer de él. Así que hay que tener cuidado, ya que por cómo nos influye nuestra cultura podríamos incurrir en interpretar mal algunos resultados de este TFG.

Como ya hemos observado, el caso previo considera la situación en la que extraemos algo sobre el sistema o realizamos una medición del sistema y analizamos cómo el resultado de esta extracción o medición nos ha ayudado a conocer más el sistema.

No obstante, podríamos preguntarnos si podemos cuantificar la información que esperamos obtener de un sistema sin haber extraído todavía nada de él. Es decir, si previamente a realizar la medición podemos cuantificar cuánta información esperamos obtener del sistema. La respuesta es afirmativa y para hacer esto no hay más que aplicar la definición de la esperanza matemática a la información. Al concepto que nace de computar la esperanza a la información se le denomina entropía.

Definición 2.1.2. [6] *Sea X una variable aleatoria discreta $\mathscr{X}$ definida en el espacio probabilístico $(\mathscr{X}, \sigma, \mathbb{P})$, se define la entropía de Shannon o la ganancia de información esperada como*

$$H(X) = E[I(X)] = \sum_{x \in \mathscr{X}} \mathbb{P}_X(x) \log\left(\frac{1}{\mathbb{P}_X(x)}\right). \tag{2.5}$$

Observación 2.1.2. *La anterior es una definición más teórica, sin embargo, suponiendo que se está trabajando con un sistema o alfabeto con L elementos o símbolos con distribución de probabilidad $\{p_i\}_{i=1}^{p}$, se define la entropía de Shannon del sistema como*

$$S[P] = \langle \log \frac{1}{p_i} \rangle = \sum_{i=1}^{L} p_i \log \frac{1}{p_i}, \tag{2.6}$$

donde p_i con $i \in \{1, ..., L\}$ son las probabilidades de cada símbolo del sistema.

Lema 2.1.1. *La entropía de Shannon es no negativa, es decir, $S[P] \geq 0$.*

Demostración. Por la definición de probabilidad $p_i \in [0, 1] \Rightarrow \frac{1}{p_i} \in [1, \infty] \Rightarrow \log \frac{1}{p_i} \geq 0$. Y por tanto, $p_i \log \frac{1}{p_i} \geq 0$, por lo que $S[P] \geq 0$. □

Observación 2.1.3. *Debido a que se cumple la relación*

$$-\log a = \log \frac{1}{a} \tag{2.7}$$

entonces, las definiciones

$$\sum_{i=1}^{n} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log p_i \tag{2.8}$$

son equivalentes y las vamos a usar de manera indistinta según interese.

Lema 2.1.2. *La entropía de Shannon para un sistema de L elementos está acotada superiormente por* $\log(L)$.

Demostración. Estamos en el caso discreto donde consideramos L posibilidades, es decir, $i \in \{1, 2, ..., L\}$ y no tenemos otra información que la de que la suma de las probabilidades p_i debe ser 1. Así, nuestro problema de maximización es:

$$\begin{cases} H(p_i) = -\sum_{i=1}^{L} p_i \log p_i, \\ \text{sujeto a:} \\ \sum_{i=1}^{L} p_i = 1, \end{cases}$$

al cual le aplicamos el método de los multiplicadores de Lagrange [7].

El método de los multiplicadores de Lagrange consiste en usar la relación entre el gradiente de la función y los gradientes de las restricciones para llegar de manera natural a una reformulación más simple del problema original a través de una función conocida como la función Lagrangiana o Lagrangiano que nada tiene que ver con el concepto del mismo nombre usado en física. En su forma original el Lagrangiano tiene la forma

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x) \tag{2.9}$$

donde $f(x)$ es la función a maximizar, en nuestro caso, $f(x) = -\sum_{i=1}^{L} p_i \log p_i$ y $g(x)$ son las restricciones, en nuestro caso, $g(x) = \sum_{i=1}^{L} p_i - 1$ y λ son los multiplicadores de Lagrange.

Entonces, derivamos el Lagrangiano con respecto a p_i y lo igualamos a cero al ser una maximización [7]:

$$\frac{\partial \mathcal{L}(p_i, \lambda_0)}{\partial p_i} = \frac{\partial}{\partial p_i}\left(-\sum_{i=1}^{L} p_i \log p_i + \lambda_0\left(\sum_{i=1}^{L} p_i - 1\right)\right) = -\log p_i - 1 + \lambda_0 = 0 \tag{2.10}$$

Nótese que la segunda derivada es $\frac{\partial^2 \mathcal{L}(p_i, \lambda_0)}{\partial p_i^2} = -\frac{1}{p_i}$ con $p_i \geq 0$ para cualquier $i \in \{1, ..., L\}$, es decir, la segunda derivada es negativa, verificando así que estamos ante un máximo. Así, resolviendo la ecuación igualada a cero recordando que denotábamos como $\log$ el logaritmo con base 2 nos queda que:

$$p_i = 2^{(\lambda_0 - 1)}. \tag{2.11}$$

Ahora, sustituyendo esta expresión en la condición de normalización $\sum_{i=1}^{L} p_i = 1$ nos queda que

$$\sum_{i=1}^{L} p_i = \sum_{i=1}^{L} 2^{\lambda_0 - 1} = 1 \Rightarrow L \cdot 2^{\lambda_0 - 1} = 1 \Rightarrow 2^{\lambda_0 - 1} = \frac{1}{L} \Rightarrow p_i = \frac{1}{L}. \tag{2.12}$$

De aquí concluimos que la situación de entropía máxima para un sistema de L elementos con distribución de probabilidad $\{p_i\}_{i=1}^{L}$ es cuando $p_i = \frac{1}{L}$, por lo que la entropía que puede alcanzar un alfabeto de L elementos siempre está acotada superiormente por el valor

$$H_{max} = -\sum_{i=1}^{L} \frac{1}{L} \log \frac{1}{L} = \sum_{i=1}^{L} \frac{1}{L} \log L = L \cdot \frac{1}{L} \log(L) = \log(L). \tag{2.13}$$

□

Teorema 2.1.1. *La entropía de Shannon para un alfabeto de L símbolos está acotada,* $0 \leq S[P] \leq \log L$.

Demostración. Directo a partir del Lema 2.1.1 y el Lema 2.1.2. □

2.2. La Entropía como Herramienta para la Detección de Correlaciones: Entropía de Bloque

En la sección previa hemos introducido el concepto de información. No obstante, todavía no hemos explicado cómo se puede utilizar este concepto para determinar el grado de correlación o de memoria de un mensaje dado como una secuencia temporal de símbolos.

Para hacer esto primero se debe resaltar que la información codificada en un mensaje se puede descomponer en dos términos, un término de entropía que cuantifica el desorden del sistema y otro término que cuantifica el orden del sistema. Al término ordenado se le denomina redundancia y significa que por ejemplo tomando en cuenta cuáles han sido los caracteres que ya han salido, hay una mayor probabilidad de saber qué caracter va a salir a continuación. Este orden depende de las correlaciones del sistema. La información desordenada, no obstante, es la entropía de la secuencia de símbolos y cuantifica la incertidumbre restante que queda en promedio cuando se han considerado ya todas las correlaciones antes de observar el siguiente caracter. Así, se puede afirmar que esta es una medida de la incertidumbre promedio por símbolo en la secuencia [6].

Nos basamos en este hecho para desarrollar la teoría que nos permita a través de los conceptos de la teoría de la información, conocer el grado de memoria de un mensaje.

Para desarrollar esta teoría necesitamos una serie de suposiciones sobre nuestro mensaje a analizar. La primera de ellas es que vamos a suponer que el sistema o mensaje consiste en una cadena infinita de símbolos $x_1...x_n...$, con $x_i \in \mathscr{X}$ donde $\mathscr{X}$ es un alfabeto del espacio probabilístico $(\mathscr{X}, \sigma, \mathbb{P})$. También suponemos que los símbolos son invariantes ante traslaciones, es decir, la probabilidad de encontrar cierta secuencia finita de símbolos en cierta posición del sistema no dependen de la posición, sino solo de los símbolos que ya se han observado, lo que para el caso de una serie temporal se traduce en que la probabilidad sea estacionaria. Además, vamos a suponer que en casi todos los casos o lo que es lo mismo, en un conjunto de casos con probabilidad 1, se puede obtener la estadística correcta de una única secuencia de símbolos infinita, es decir, se puede conocer a la perfección el sistema a partir de esta secuencia [6].

Entonces resumiendo, tenemos una secuencia de símbolos infinita y nuestro conocimiento previo del sistema solo refleja que cierto símbolo pertenece a cierto alfabeto $\mathscr{X}$, que contiene $|\mathscr{X}| = L$ caracteres diferentes. Nuestro desconocimiento previo por símbolo es entonces $S = \log L$. Esto es debido a que al principio no conocemos nada sobre nuestro sistema, luego el desconocimiento es el máximo posible, el cual según el Lema 2.1.2 es $\log(L)$, ya que cabe recordar por lo dicho en el Ejemplo 2.1.1 que conocimiento del sistema e información eran inversamente proporcionales.

Añadiendo sucesivamente distribuciones de probabilidad para secuencias de cada vez mayor longitud $\mathbb{P}_n$ con $n = 1, 2, ...$ podemos tener en cuenta las correlaciones para reducir el desconocimiento del siguiente símbolo en la secuencias. La entropía que todavía queda cuando incluimos todas las longitudes ($n \to \infty$) es la entropía de Shannon de la secuencia de símbolos [6].

Un posible enfoque para cuantificar esta entropía de las secuencias de longitud n es la entropía de bloque.

Definición 2.2.1. [6] *Sea $(\mathscr{X}, \Omega, \mathbb{P})$ el espacio probabilístico del sistema $\mathscr{X}$ y sea $x_1...x_m$ una secuencia de m símbolos con $x_i \in \mathscr{X}$, entonces se define la entropía de bloque como*

$$S_m = S[P_m] = \sum_{x_1...x_m} p(x_1...x_m) \log \frac{1}{p(x_1...x_m)}. \tag{2.14}$$

A partir del cual se puede definir la entropía por símbolo s como

$$s = \lim_{t\to\infty} \frac{1}{m} S_m. \tag{2.15}$$

Otro posible enfoque para cuantificar la entropía de las secuencias de longitud n está basado en un procedimiento mediante el cual formamos probabilidades condicionadas del siguiente símbolo teniendo en cuenta los anteriores. Este enfoque se basa en el hecho de que cuando extendemos la secuencia precedente que se usa para determinar la probabilidad condicionada del siguiente símbolo en salir, mejoramos la descripción probabilística. La información ganada por esta mejora se cuantifica a través del concepto de información relativa [6].

No obstante, antes de introducir el concepto de información relativa debemos introducir otros conceptos. El primero de ellos es la probabilidad condicionada de un símbolo x_2 habiendo salido antes un símbolo x_1:

Definición 2.2.2. [6] *Sea $(\mathscr{X}, \Omega, \mathbb{P})$ el espacio probabilístico del alfabeto o sistema $\mathscr{X}$ y sean $x_1, x_2 \in \mathscr{X}$ dos símbolos, se define la probabilidad condicionada de un símbolo x_2 habiendo salido antes un símbolo x_1 como:*

$$p(x_2|x_1) = \frac{p(x_1 \text{ y } x_2)}{p(x_1)} = \frac{p(x_1 x_2)}{p(x_1)}, \tag{2.16}$$

donde $p(x_1 \text{ y } x_2)$ denota la probabilidad de que salgan ambos símbolos. Para el caso general

$$p(x_m|x_1 ... x_{m-1}) = \frac{p(x_1 ... x_m)}{p(x_1 ... x_{m-1})}. \tag{2.17}$$

Así, se puede cuantificar la incertidumbre promedio h_m para un símbolo posterior dado una secuencia precedente de $m-1$ símbolos:

Proposición 2.2.1. [6] *Sea $(\mathscr{X}, \Omega, \mathbb{P})$ el espacio probabilístico del sistema $\mathscr{X}$ y sean $x_i \in \mathscr{X}$ símbolos del sistema, entonces la incertidumbre promedio h_m para el siguiente símbolo dada una secuencia precedente de $m-1$ símbolos es*

$$h_m = \langle S[p(X_m|x_1 ... x_{m-1})] \rangle = \Delta S_m \tag{2.18}$$

donde ΔS_m es el gradiente en forma de diferencia finita de la entropía de bloque de m símbolos definida como $\Delta S_m = S_m - S_{m-1}$

Demostración. Véase [6]. □

Corolario 2.2.1. [6] *La entropía de bloque es una función creciente con la longitud de bloque m, es decir, $\Delta S_m \geq 0$*

Demostración. Véase [6] □

Una vez introducidos estos conceptos, ya podemos introducir el concepto de entropía relativa. El concepto de entropía relativa viene motivado porque queremos cuantificar la información en correlaciones de longitud m. Para ello suponemos que tenemos una distribución de probabilidad uniforme a la que denotamos por $\mathbb{P}^{(0)}(X_m|x_1 x_2 ... x_{m-1})$ para un símbolo x_m, dado que una secuencia precedente específica $(x_1 x_2 ... x_{m-1})$ es conocida.

Entonces, estamos interesados en la información que obtenemos cuando observamos un símbolo x_1 un instante después y lo usamos para cambiar nuestra descripción probabilística del símbolo x_m a $\mathbb{P}(X_m|x_1 x_2 ... x_{m-1})$.

Definición 2.2.3. [6] *Se define la información relativa entre las distribuciones de probabilidad* $\mathbb{P}^{(0)}$ *y* $\mathbb{P}$ *definidas anteriormente como la medida de correlación de la información de longitud* m *cuando una secuencia específica* $(x_1, x_2, ..., x_{m-1})$ *es observada:*

$$K[\mathbb{P}^{(0)}; \mathbb{P}] = \sum_{x_m} p(x_m|x_1x_2...x_{m-1}) \log \frac{p(x_m|x_1x_2...x_{m-1})}{p(x_m|x_2...x_{m-1})}. \tag{2.19}$$

Si tomamos el promedio entre todas las posibles secuencias precedentes $(x_1, x_2, ..., x_{m-1})$ obtenemos una expresión para el promedio del contenido de información en correlaciones de longitud m a lo cual denominamos como información de correlación k_m.

Definición 2.2.4. [6] *Sea* $p(x_1...x_{m-1})$ *la probabilidad de cierta secuencia de longitud* $m-1$*, entonces se define la información de correlación* k_m *como la información promedio en correlaciones de longitud* m

$$k_m = \sum_{x_1...x_{m-1}} p(x_1...x_{m-1}) K[\mathbb{P}^{(0)}; \mathbb{P}] = \tag{2.20}$$

$$\sum_{x_1...x_{m-1}} p(x_1...x_{m-1}) \sum_{x_m} p(x_m|x_1x_2...x_{m-1}) \log \frac{p(x_m|x_1x_2...x_{m-1})}{p(x_m|x_2...x_{m-1})} \tag{2.21}$$

Además, la siguiente proposición nos da una caracterización relativamente útil de la información de correlación:

Proposición 2.2.2. [6] *Sea* $p(x_1...x_{m-1})$ *la probabilidad de cierta secuencia de longitud* $m-1$*, entonces se cumple que*

$$k_m = -\Delta^2 S_m \geq 0 \tag{2.22}$$

donde k_m *es el promedio de las correlaciones a longitud m y* $\Delta^2 S_m = S_m - 2S_{m-1} + S_{m-2}$ *es la curvatura en forma de diferencia segunda de la entropía de bloque.*

Demostración. Véase [6] □

Con estas definiciones ya dadas podemos dar la definición que Grassberger introdujo de complejidad. Aunque esta definición no es única, es la que se emplea en este trabajo [6].

Definición 2.2.5. [6] *Sea* k_m *la información de correlación promedio en correlaciones de longitud m, entonces se define la complejidad* η *como la suma ponderada de las contribuciones de la información de correlación de diferentes distancias*

$$\eta = \sum_{m=1}^{\infty} (m-1) k_m = -\sum_{m=1}^{\infty} (m-1) \cdot \Delta^2 S_m \tag{2.23}$$

donde la última igualdad se cumple por la Proposición 2.2.2 anterior.

El objetivo de este trabajo es cuantificar el grado de correlación de cierta señal temporal o el grado de memoria de dicha señal temporal, para así deducir si es una señal aleatoria o no correlacionada o en cambio es una señal correlacionada. La Definición 2.2.5 permite precisamente cuantificar lo anterior y comparar entre señales temporales para concluir cuál está más o menos correlacionada.

En la mayoría de los casos aplicar la definición directamente para calcular el valor de η no suele ser el método más usado. En la literatura se usa más la caracterización basada en el siguiente resultado.

Proposición 2.2.3. [6] *Sea* η *la complejidad definida en la definición anterior, entonces se cumple que* $\eta = \lim_{m\to\infty} (S_m - ms)$ *donde* S_m *es la entropía de bloque para longitud* m *y* s *es la entropía por símbolo teórica.*

La siguiente figura ilustra de manera esquemática lo que dice la proposición.

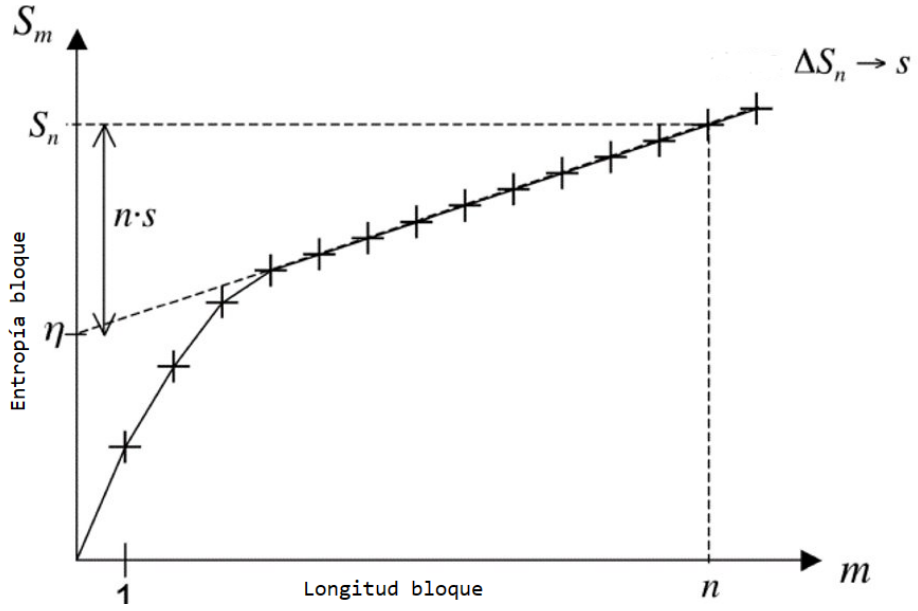


Figura 2.1: Representación de la entropía de bloque en función de la longitud del bloque para una secuencia de longitud infinita. Imagen tomada de [6].

Demostración. Véase [6]. □

Recordemos, que en la Definición 2.2.5 se definió la complejidad como una medida de la magnitud de la correlación que se encuentra en un mensaje pesada por la longitud a la que se encuentra.

Es por esto que la Proposición 2.2.3 permite distinguir entre mensajes que tienen correlaciones y mensajes completamente aleatorios. Por ejemplo, la Figura 2.1 nunca podría representar a una secuencia aleatoria, ya que una secuencia aleatoria debe tener complejidad 0 que se corresponde con la abscisa en el origen, es decir, no debe tener correlaciones y es evidente que la η del mensaje de la Figura no es 0.

Otra forma de ver este hecho desde un punto de vista de análisis desde la teoría de la información más básica es que en una secuencia aleatoria desde el primer símbolo al último la pendiente debe ser igual, ya que por definición en una secuencia aleatoria, el ver un símbolo no nos proporciona ninguna información sobre el sistema y por tanto, la pendiente del punto 1 al 2 debe ser la misma que la del 2 al 3 y así sucesivamente, es decir, la pendiente no puede decrecer como lo hacía en el caso correlacionado representado por la Figura 2.1, ya que seguiremos esperando la misma ganancia de información independientemente de cuántos símbolos hayamos sacado, ya que el saber qué símbolos han salido no nos ha proporcionado ningún tipo de información o conocimiento del sistema.

Resumiendo, la Proposición 2.2.3 nos permite calcular η de una forma práctica y sistemática a través del siguiente resultado, el cual es una consecuencia directa de ella.

Corolario 2.2.2. [6] *Sea m_0 la longitud de permutación para la cual se da la transición de región de curvatura no nula a región de curvatura nula de la Proposición 2.2.3, entonces se verifica que $S_m = \eta + s \cdot m$ para cualquier $m \geq m_0$ donde s representa la entropía por símbolo teórica.*

Demostración. Consecuencia directa de la Proposición 2.2.3. □

Es decir, podemos obtener la complejidad de cualquier secuencia temporal representando sus valores de S_m en función de la longitud del bloque m y haciendo el ajuste a la parte que se ajusta a una recta, la cual teóricamente para una señal temporal de longitud infinita, sería para todas las longitudes de bloque m de un m_0 en adelante.

La complejidad η es la ordenada en el origen de este ajuste y la entropía por símbolo teórica s, la pendiente de este ajuste.

2.3. Problema de Tamaño Finito

En las dos secciones previas nos hemos basado en [6] para desarrollar el marco teórico concerniente a la teoría de la información adaptado al objetivo de este trabajo que es la detección de correlaciones en señales temporales a través de la entropía de bloque.

No obstante, el desarrollo teórico de la sección anterior es válido para el caso de la entropía de bloque en el que estemos trabajando con una serie temporal de longitud infinita. Sin embargo, en aplicaciones prácticas no se dispone de señales temporales de longitud infinita. Y no se ha encontrado en la literatura ningún desarrollo teórico que sea válido ni para la entropía de permutación, ni para señales temporales que no sean de longitud infinita, por lo que vamos a tener que analizar cómo se modifica la discusión precedente para el caso de series temporales finitas, lo que denominaremos problema de tamaño finito.

Primero vamos a resolver el problema de que tenemos señales finitas y no infinitas, a lo cual se le denomina problema de tamaño finito. Para ello vamos a empezar suponiendo que seguimos trabajando con la entropía de bloque pero con señales de longitud finita. Este hecho va a afectar a nuestro estudio en el sentido de que la entropía de bloque no va a crecer hasta infinito como lo hacía anteriormente, sino que va a estar acotada por un valor S_{max} que va a depender únicamente de la longitud de nuestra señal temporal N.

Es decir, la recta de la Figura 2.1 no se va a prolongar indefinidamente sino que va a saturar a un valor $S_{max}(N)$ como se ilustra a través de la siguiente figura:

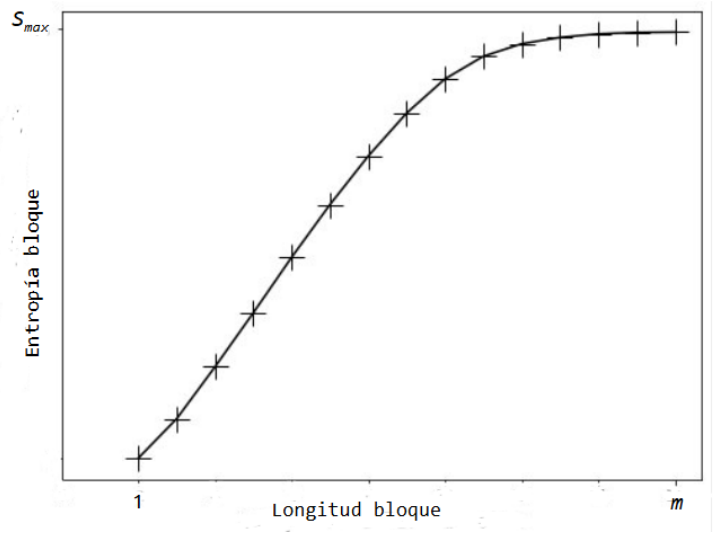


Figura 2.2: Representación de la entropía de bloque en función de la longitud del bloque para una secuencia de longitud finita. Se puede observar que se estabiliza en un valor S_{max} dado.

Todas estas afirmaciones se pueden demostrar y lo hacemos mediante el lema y el teorema siguientes.

Lema 2.3.1. *Sea $x_1 ... x_N$ una señal temporal generada por cualquier tipo de sistema físico. Sea S_m el valor de la entropía de bloque para la longitud de bloque m, con $k \in \{1, ...N\}$, entonces se verifica que $S_m \leq \log(N)$ para cualquier $k \in \{1, ...N\}$.*

Demostración. Vamos a hacer esta demostración en dos pasos. Lo primero de todo nos damos cuenta de que el Lema 2.1.2 nos dice que una entropía de Shannon con L símbolos está acotada superiormente por $\log L$.

Por otro lado, nos damos cuenta que nosotros estamos trabajando a la hora de calcular la entropía de bloque con una señal temporal $x_1...x_N$ de longitud N. Es inmediato deducir por tanto que si nos preguntáramos cuál es la cota superior de la entropía de Shannon de esta señal concluiríamos que $\log N$, ya que en este caso aplicaríamos el Lema 2.1.2 con $L = N$.

No obstante, lo que nosotros queremos saber es cuál es la cota superior de las entropías de bloque asociadas a la señal $x_1...x_N$, no de la entropía de Shannon. No obstante, si uno se da cuenta la entropía de bloque no es más que un conjunto de entropías de Shannon pero en vez de aplicadas a símbolos individuales, a bloques de cualquier longitud de símbolos adyacentes en la señal temporal. No obstante, a la hora de aplicar la definición de la entropía de Shannon (Definición 2.1.2) da igual lo que los elementos del alfabeto sean, bloques, símbolos o lo que sea.

Asimismo, nosotros al final el cálculo lo hacemos sobre nuestra señal $x_1...x_N$ que es la que tiene la información. Luego, como mucho el mayor número de bloques a los que poder calcular la entropía de bloque será N, ya que es absurdo pensar que de una señal de N símbolos podamos sacar más de N bloques no repetidos considerando que los símbolos de los bloques deben ser adyacentes en la señal temporal. Por lo tanto, la máxima entropía de bloque será la máxima entropía de Shannon asociada a N bloques o elementos del alfabeto que por el Lema 2.1.2 es $\log N$. □

No obstante, el problema de tamaño finito no se traduce en que únicamente S_m no puede ser mayor que el valor dado por el teorema anterior, sino que además, S_m es una función creciente la cual a partir de cierto valor satura en el valor dado por el teorema anterior y este es realmente el comportamiento debido al problema de tamaño finito de la entropía de bloque. Enunciamos este hecho a partir del siguiente teorema.

Teorema 2.3.1. *Sea $x_1...x_N$ una señal temporal generada por cualquier tipo de sistema físico. Sea S_m el valor de la entropía de bloque para la longitud de bloque m, con $m \in \{1, ...N\}$, entonces S_m es una función creciente con el parámetro m que satura en el valor $S_k = \log N$.*

Demostración. Para demostrar este teorema nos basamos en el Corolario 2.2.1 donde se prueba que la entropía de bloque es una función creciente con la longitud de bloque m, es decir, $\Delta H_k \geq 0$ y en el Lema 2.3.1 anterior donde se demuestra que $S_k \leq \log N$.

Juntando estos dos hechos es inmediato concluir que S_m es una función creciente que satura en el valor $S_{max} = \log N$. □

A continuación se exponen algunos valores de cotas superiores para longitudes N comúnmente usadas en el texto.

N	10^3	10^4	10^5	$2 \cdot 10^5$	10^6
S_{max}	9.966	13.287	16.610	17.610	19.932

Tabla 2.1: Valores máximos de la entropía para longitudes N de la señal temporal usadas a lo largo del texto.

Observación 2.3.1. *Cabe resaltar que en ningún momento ni en el Lema 2.3.1 ni en el Teorema 2.3.1 se ha usado ninguna propiedad particular de la entropía de bloque, es decir, se podría hacer exactamente el mismo argumento con la entropía de permutación. Así, aunque todavía no se ha introducido el concepto de entropía de permutación lo cual se hará en la siguiente sección, adelantamos que este desarrollo*

teórico será igualmente válido para la entropía de permutación.

Por ello, ya hemos resuelto el problema de la finitud de la señal y solo nos queda trasladar el marco teórico de la entropía de bloque para señales infinitas a la entropía de permutación para señales infinitas.

2.4. Optimización de la Entropía de Bloque: Entropía de Permutación

En esta sección introducimos primero a través de la Subsección 2.4.1 el por qué de usar la entropía de permutación en lugar de la de bloque cuando para la de bloque hay todo un desarrollo teórico detrás y para la de permutación no. Luego, damos la definición propia de la entropía de permutación a través de la Sección 2.4.2. El análisis sobre cómo adaptar el desarrollo teórico de la Sección 2.2 válido para la entropía de bloque para que sea válido para la entropía de permutación vendrá dado en la siguiente Sección 2.5.

Notación 2.4.1. *Dependiendo del autor y el ámbito donde se use la entropía se le suele denotar por la letra H o por la letra S. En este texto se ha escogido denotar a la entropía de Shannon y de bloque por la letra S y a la entropía de permutación se la ha denotado H. Con esto se ha querido facilitar al lector la distinción entre la entropía de bloque y la entropía de permutación al denotarlas mediante letras distintas.*

2.4.1. Problemas de la entropía de bloque

El análisis de la entropía de bloque en series temporales reales conlleva la cuestión que se denomina problema de discretización. Este hecho es una consecuencia indirecta del problema de tamaño finito, es decir, es una consecuencia de que las señales temporales tengan longitudes finitas.

Para determinar la entropía de bloque de una señal temporal física $x(t)$ necesitamos hacer una partición del rango de la función, ya que una señal temporal puede tomar un número infinito de valores reales entre $\min x(t)$ y $\max x(t)$. Dicho de otro modo, uno tiene que transformar una serie de datos de una variable continua en una serie de símbolos de un conjunto discreto. Cuánto más fina es la resolución deseada, más grande es el alfabeto necesario. Esto supone una limitación muy importante en las aplicaciones prácticas de la teoría de la información en señales reales. Para más detalles de esta problemática ver [1].

2.4.2. Entropía de permutación

Consideremos una serie temporal $\{x_t\}_{t=1}^{N}$. Para poder calcular la entropía de permutación H_n de un bloque de longitud n, necesitamos primero calcular la frecuencia con la que obtenemos cada una de las $n!$ posibles permutaciones (ordenaciones de mayor a menor) de n elementos en nuestro mensaje. Para ello definimos la aplicación $\mathscr{O}$ para ordenar de mayor a menor los bloques de longitud k como

$$\begin{array}{rccc} \mathscr{O}: & \mathbb{R}^k & \longrightarrow & \mathbb{R}^k \\ & (x_1, x_2, ..., x_k) & \mapsto & (o_1, o_2, ..., o_k), \end{array} \tag{2.24}$$

donde o_i representa cuántos x_j son mayores que x_i si $i \neq j$. Por ejemplo, si tenemos el bloque de longitud 4, $x = (7.45, -0.3, 98.7, -2.517)$, entonces $o(x) = (1, 2, 0, 3)$. Se puede demostrar que para una longitud k dada hay $k!$ posibles ordenamientos de mayor a menor posibles, o dicho de otra manera hay $k!$ permutaciones posibles. A continuación damos un ejemplo del cálculo de la frecuencia

de todas las posibles permutaciones de longitud 2 de un mensaje. Supongamos que nuestro mensaje es $x(t) = (1.4, 3.11, 2.9, 4.56)$, entonces los posibles bloques de longitud 2 son $(1.4, 3.11)$, $(3.11, 2.9)$ y $(2.9, 4.56)$, lo que se corresponde a las permutaciones $(1, 0)$, $(0, 1)$ y $(1, 0)$ respectivamente. Como estamos en $n = 2$, las dos únicas posibles permutaciones son $(0, 1)$ y $(1, 0)$, por lo que la frecuencia de la permutación $(0, 1)$ es $\frac{1}{3}$ y la de la permutación $(1, 0)$ es $\frac{2}{3}$.

Así, en sentido general, para cada permutación π_i determinamos su frecuencia relativa como [8]

$$p(\pi_i) = \frac{\#\{t|t \leq T - n : (x_{t+1}, ..., x_{t+n}) \text{ es de tipo } \pi_i\}}{T - n + 1}, \tag{2.25}$$

que no es más que aplicar la regla de Laplace de casos favorables entre casos totales.
Cabe destacar que esto no es más que una estimación de la frecuencia de cada tipo de permutación y que para obtener la expresión correcta se debería tomar una serie temporal infinita $\{x_1, x_2, ..., x_n\}$ y tomar el límite para en la Ecuación (2.25) anterior.

Con todo esto ya podemos dar la definición de entropía de permutación.

Definición 2.4.1. *Sea Σ el grupo de permutaciones de n elementos y sean $\pi_i \in S_n$ las $n!$ permutaciones, se define la entropía de permutación de orden $n \geq 2$ como*

$$H(n) = -\sum_{\pi_i \in \Sigma} p(\pi_i) \log p(\pi_i). \tag{2.26}$$

Proposición 2.4.1. *La entropía de permutación para longitud de bloque n elementos está acotada, $0 \leq H(n) \leq \log n!$*

Demostración. Directo a partir del Teorema 2.1.1 y del hecho de que para una longitud de bloque n hay $n!$ posibles permutaciones, con lo que el alfabeto del Teorema 2.1.1 en este caso tiene $L = n!$ elementos. □

Cabe resaltar que la cota inferior se obtiene para una secuencia periódica de valores y la cota superior para una secuencia completamente aleatoria de valores.

Para poder comparar secuencias de distintas longitudes se define la siguiente cantidad:

Definición 2.4.2. *Se define la entropía por símbolo de orden $n \geq 2$ como*

$$h_n = \frac{H(n)}{n - 1}. \tag{2.27}$$

Esta entropía por símbolo no deja de ser una estimación. Para obtener la entropía por símbolo real o teórica, debemos tomar el límite de $n \to \infty$:

Definición 2.4.3. *Se define la entropía por símbolo real o teórica como*

$$h = \lim_{n \to \infty} \frac{1}{n - 1} H_n. \tag{2.28}$$

2.5. Adaptación del Desarrollo Teórico de la Sección 2.2 a la Entropía de Permutación

Vamos a dividir esta sección en dos partes. En la primera vemos si es válida la teoría de la Sección 2.2 desarrollada para la entropía de bloque también para la entropía de permutación, lo cual ya adelantamos que la respuesta será negativa y en la segunda subsección desarrollamos un marco teórico propio para la entropía de permutación.

2.5.1. ¿Es válida la teoría de la Sección 2.2 para la entropía de permutación?

Lo primero de todo vamos a centrarnos en lo que nos interesa, para lo cual necesitamos saber qué es lo que nos interesa. A nosotros nos interesan los resultados que nos sirvan para poder calcular la complejidad de una serie temporal de una forma práctica.

Si volvemos un momento a la Sección 2.2 nos damos cuenta de que hay dos formas para calcular la complejidad de señales temporales de manera sistemática. La primera es a través de la Proposición 2.2.3 ya que nos asegura el cálculo de la complejidad a través de ajustes lineales y la segunda es juntando la Definición 2.2.5 y la Proposición 2.2.2 para calcular la complejidad a través del cálculo de curvaturas.

Así, de lo que más nos interesa comprobar la validez es en primer lugar de la Proposición 2.2.3 o bien de la Proposición 2.2.2 .

Empecemos con la Proposición 2.2.3, que desafortunadamente no es muy difícil de comprobar que es falsa para el caso de la entropía de permutación.

Proposición 2.5.1. *Sean H_k la entropía de permutación para bloques de longitud k y h_∞ la entropía por símbolo teórica, entonces es falso que se cumpla que $\eta = \lim_{m\to\infty}(H_k - k\cdot h_\infty)$, es decir, no es verdad que para la entropía de permutación exista un $k_0 \in \mathbb{N}$ a partir del cual $H_k = \eta + h_\infty \cdot k$ para cualquier $k \in \mathbb{N}$, es decir, no es verdad que a partir de cierto $k_0 \in \mathbb{N}$ la entropía de permutación siempre se ajuste a una recta.*

Demostración. Para demostrar esto basta observar el siguiente contraejemplo.

Supongamos que tenemos una secuencia completamente aleatoria, es decir, que cualquier bloque de longitud $\{1, 2, ..., i, ...\}$ es igualmente probable que los demás $i-1$ bloques, es decir, para todas las longitudes de bloque la distribución es uniforme.

Por tanto, por la Proposición 2.4.1 para cada longitud de bloque n la entropía alcanza el valor máximo que es de $\log(n!)$.

Por tanto, tenemos una función con el dominio en $\mathbb{N}$ de la forma

$$f(n) = \log(n!) = \log[n\cdot(n-1)...\cdot 1] = \log(n) + \log(n-1) + ... + \log(1) = \sum_{k=1}^{n}\log k, \quad (2.29)$$

de donde se puede deducir que

$$\Delta f(n) = f(n) - f(n-1) = \log(n), \quad (2.30)$$

lo cual solo es constante si $n = 2^k$ con $k \in \mathbb{Z}$ (recordemos que estábamos trabajando en base 2), que ya sabemos que no es el caso, ya que $n \in \mathbb{N}$. Cabe resaltar que hemos definido Δf como la diferencia finita centrada a derecha porque así el resultado es más simple, pero si se tomara la definición centrada o a izquierda el resultado es análogo.

Como estamos en el rango de $n \to \infty$ otra forma de ver esto sería la de realizar la aproximación de Stirling la cual afirma que si $n \to \infty$ entonces $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ de donde se deduce que

$$\log(n!) \approx log(\sqrt{2\pi n}) + nlog(n) - nlog(e). \quad (2.31)$$

Por tanto como estamos en el rango $n \to \infty$ se cumple que $n \gg \log(n) \gg \log(\sqrt{n})$ y así nos queda que

$$log(n!) \approx n\cdot(\log(n) - \log(e)), \quad (2.32)$$

y así

$$f'(n) \approx log(n) - log(e) + 1, \tag{2.33}$$

que es un resultado análogo al obtenido anteriormente.

Así, probamos que no es cierto que la entropía de permutación siempre tienda a una recta para bloques suficientemente largos. □

Este resultado es muy importante y es que en la literatura como [6] se basan en este hecho de ajuste a la recta para poder calcular la complejidad de una señal temporal de manera sistemática. A partir de aquí hay dos posibles soluciones, o bien enunciar y demostrar un resultado que determine en qué situaciones sí se ajusta a una recta y solo estudiar esos casos o bien intentar ver si la Proposición 2.2.2 resulta válida también para la entropía de permutación.

No obstante, a través de la siguiente proposición tampoco es muy difícil ver que la Proposición 2.2.2 tampoco es válida cuando pasamos de la entropía de bloque a la entropía de permutación.

Proposición 2.5.2. *Sea $\Delta^2 H_m = \frac{H_{m+2}-2H_m+H_{m-2}}{4}$ la segunda diferencia finita de la entropía de permutación, entonces es falso que se cumpla $k_m = -\Delta^2 H_m$ siendo k_m el promedio de las correlaciones a longitud m.*

Demostración. Supongamos por reducción al absurdo que $k_m = -\Delta^2 H_m$. Tomemos una señal completamente aleatoria que por definición cumple que $k_m = 0\ \forall m \in \mathbb{N}$. Entonces, si se cumpliera $k_m = -\Delta^2 H_m$ obtendríamos $\eta = \sum_{n=1}^{\infty}(n-1)k_n = -\sum_{n=1}^{\infty}(n-1)\Delta^2 H_n = 0$.

Comprobemoslo. Por la Proposición 2.4.1 sabemos que para una señal aleatoria $H_k = \log(k!)$, ya que sabemos que para una señal aleatoria la entropía es máxima. Entonces, calculemos $\Delta^2 H_n$ para la señal aleatoria.

$$\Delta^2 H_n = \frac{H_{n+2} - 2H_n + H_{n-2}}{4} = \frac{\log[(n+2)!] - 2\log(n!) + \log[(n-2)!]}{4} =$$

$$= \frac{\log(n+2) + \log(n+1) + \log(n!) - 2\log(n!) - \log(n) - \log(n-1) + \log(n!)}{4} =$$

$$= \frac{\log(n+2) + \log(n+1) - \log(n) - \log(n-1)}{4} > 0\ \forall n \in \mathbb{N}\backslash\{0,1\}$$

donde la desigualdad proviene de que $\log(n+2) > \log(n)$ y $\log(n+1) > \log(n-1)\ \forall n \in \mathbb{N}\backslash\{0,1\}$. Así, concluimos que $\eta = -\sum_{n=1}^{\infty}(n-1)\Delta^2 H_n < 0$, que es contradictorio con que $\eta = 0$, por lo cual hemos llegado a un absurdo y así $k_m = -\Delta^2 H_m$ es falso. □

Observación 2.5.1. *Cabe destacar que en [6] y por ende en la Proposición 2.2.2 se calcula la diferencia segunda como $\Delta^2 S_m = S_m - 2S_{m-1} + S_{m-2}$, mientras que en esta proposición se calcula como $\Delta^2 H_k = \frac{H_{k+2}-2H_k+H_{k-2}}{4}$, ya que según [9] conlleva un menor error al aproximar la curvatura que lo propuesto en [6]. No obstante, es inmediato observar que el resultado de esta proposición es exactamente igual usando cualquiera de las dos definiciones.*

Probablemente el lector ya se haya dado cuenta de la magnitud del problema y es que como consecuencia de la Proposición 2.5.2 se deduce la no concavidad de la entropía de permutación.

Corolario 2.5.1. *Sea $\Delta^2 H_n$ la diferencia segunda de la entropía de permutación, entonces es falso que $\Delta^2 H_m \leq 0$, es decir, la entropía de permutación no es cóncava.*

Demostración. Inmediato de que en la demostración de la Proposición 2.5.2 hemos llegado a que para una señal aleatoria se verifica que $\Delta^2 H_m > 0$. □

No obstante, la concavidad de la entropía como concepto es una de sus propiedades más notables [6] y que para la entropía de permutación, que es un tipo particular de entropía esto no se verifique parece de lo más extraño.

Sin embargo, aunque sea extraño, considerando estos resultados debemos concluir que tenemos un problema, ya que lo que se concluye es que el desarrollo teórico de la Sección 2.2 no es válido en su forma original para el caso de la entropía de permutación, por lo que se debería construir un marco teórico o al menos adaptar el existente de la Sección 2.2 para la entropía de bloque para que resulte válido para la entropía de permutación.

2.5.2. Marco Teórico Propio para la Entropía de Permutación

Empecemos por algo sencillo que es darnos cuenta de que mientras que la entropía de bloque de longitud 1, S_1, tiene sentido, no tiene sentido hablar de la entropía de permutación de longitud 1, H_1, ya que solo hay una posible permutación en este caso. En otras palabras, el bloque mínimo que tiene sentido considerar para la entropía de permutación es $m = 2$. Esto produce un desplazamiento $m \to m + 1$ en algunas de las fórmulas y definiciones que presentamos en la Sección 2.2 para la entropía de bloque.

Definición 2.5.1. *Sea k_m la información de correlación promedio en correlaciones de longitud m, entonces se define la complejidad η como la suma ponderada de las contribuciones de la información de correlación de diferentes distancias:*

$$\eta = \sum_{m=2}^{\infty}(m-2)k_m, \tag{2.34}$$

donde k_m representa las correlaciones promedio en longitudes m. Comparar con la Ecuación (2.23).

Una vez hecho esto vamos a hacer unas comprobaciones. Sabemos por [6] que para que una definición de complejidad sea correcta es condición necesaria que se verifique $\eta = 0$ tanto para una señal aleatoria como para una señal periódica, ya que ninguna de ellas tiene correlaciones, $k_m = 0$. Así que veamos cómo se comporta la definición

$$\eta = \sum_{m=2}^{\infty}(m-2)k_m = -\sum_{m=2}^{\infty}(m-2)\Delta^2 H_m, \tag{2.35}$$

aunque sepamos por la Proposición 2.5.2 que la segunda igualdad no es correcta.

Empecemos por ver que para una señal periódica se obtiene $\eta = 0$.

Proposición 2.5.3. *Sea $\eta = -\sum_{n=1}^{\infty}(n-2)\Delta^2 H_k$ una medida de la complejidad y H_k la entropía de permutación, entonces $\eta = 0$ cuando la señal es completamente periódica.*

Demostración. Dependiendo del criterio se podría decir que una señal completamente periódica es la periódica de periodo 1, es decir, $x(t) = (a, a, a, a, a, a, ...)$ con $a \in \mathbb{R}$ o la periódica de periodo 2, es decir, $x(t) = (a, b, a, b, a, b, ...)$ con $a, b \in \mathbb{R}$. El criterio depende en si considerar la de periodo 1 como constante y periódica o solo constante.

En cualquier caso, por un lado, la de periodo 1 sería por ejemplo $x(t) = (a, a, a, a, a, a, a, ...)$, de lo cual deducimos que $H_k = 0\ \forall k \in \mathbb{N}$, ya que para cualquier k solo hay un tipo de permutación posible y da igual que sean siempre igualdades, solo hay una posible secuencia. De este hecho deducimos que $\Delta^2 H_k = 0$ y por tanto $\eta = 0$.

Por otro lado, la señal periódica $x(t) = (a, b, a, b, a, b, a, b...)$ donde suponemos sin pérdida de generalidad que $a > b$. En este caso, para longitudes de permutación de bloque 2, vamos a obtener que las posibles permutaciones van a ser la de que el primer índice sea mayor $(0, 1)$ con probabilidad $p = \frac{1}{2}$ y la de que el segundo índice sea mayor $(1, 0)$ con probabilidad $p = \frac{1}{2}$. Así, según la Definición 2.4.1 $H_2 = 1$. Para H_3 vamos a tener la secuencia $(1, 0, 1)$ con probabilidad $p = \frac{1}{2}$ y la secuencia $(0, 1, 0)$ con probabilidad $p = \frac{1}{2}$, por lo que de nuevo H_3=1. Para H_4 vamos a tener la secuencia $(1, 0, 1, 0)$ con $p = \frac{1}{2}$ y la secuencia $(0, 1, 0, 1)$ con $p = \frac{1}{2}$, por lo que de nuevo $H_4 = 1$. Así sucesivamente tendremos $H_k = 1\ \forall k \in \mathbb{N}$ y así $\Delta^2 H_k = 0$ para todo $k \geq 2$, por lo que $\eta = 0$.

Cabe destacar que en este caso las igualdades se han considerado como tal ya que es una idea teórica y el razonamiento es el mismo permitiendo las igualdades o no permitiéndolas.

Por lo tanto, concluimos que nuestra definición funciona correctamente para el límite de las señales periódicas. □

Ahora, veremos como para una señal aleatoria no se cumple $\eta = 0$

Proposición 2.5.4. *Sea $\eta = -\sum_{n=2}^{\infty}(n-2)\Delta^2 H_k$ una medida de la complejidad y H_k la entropía de permutación, entonces $\eta = -\infty$ cuando la señal temporal es de longitud infinita y completamente aleatoria.*

Demostración. Para probar este resultado basta ver la demostración de la Proposición 2.5.2. En ella hemos visto que para una señal aleatoria se cumple que

$$\Delta^2 H_n = \frac{\log(n+2) + \log(n+1) - \log(n) - \log(n-1)}{4}, \tag{2.36}$$

de donde se deduce que

$$\eta = -\sum_{n=2}^{\infty}(n-2)\Delta^2 H_n \tag{2.37}$$

$$= -\sum_{n=2}^{\infty}(n-2) \cdot \frac{\log(n+2) + \log(n+1) - \log(n) - \log(n-1)}{4} \underset{n\to\infty}{\to} -\infty. \tag{2.38}$$

Como estamos trabajando en el marco teórico donde tenemos señales temporales infinitas, es decir, $n \to \infty$, entonces $\eta = -\infty$. □

Vamos a parar un momento y vamos a analizar todo lo visto en lo que va de sección. En general la mayoría se podrían calificar como resultados extraños, pero posibles. Sin embargo, el que desde un punto de vista teórico es muy difícil de admitir es el Corolario 2.5.1, ya que la concavidad de la entropía es un concepto teórico muy asentado [6].

Esto lo que nos sugiere es que la entropía de permutación debe de tener alguna propiedad especial que hace que tenga este comportamiento tan anómalo. Una pista de por dónde puede venir la solución a este comportamiento la encontramos en la Proposición 2.5.1, en el hecho de que la pendiente de la señal más aleatoria sea logarítmica.

Vamos a volver por un momento a la entropía de bloque, pero no a la entropía de bloque teórica sino al caso práctico. Para calcular de forma no teórica la entropía de bloque se debe dividir el eje Y en un número m de *bin*es, es decir, se debe hacer una partición de m elementos del eje Y, que es justamente el origen del problema de discretización comentado en la Subsección 2.4.1. De aquí el caso teórico se obtiene haciendo $m \to \infty$. Pero en cualquier caso, para la entropía de bloque de un elemento, S_1, hay m posibilidades, para la entropía de bloque, S_2, hay m^2 posibilidades y para

la entropía de bloque k, S_k, hay m^k posibilidades, de donde se deduce que la pendiente para la señal aleatoria (pendiente máxima) en este caso de la entropía de bloque es:

$$\Delta S_k = S_{k+1} - S_k = \log m^{k+1} - \log m^k = (k+1)\log m - k\log m = \log m \tag{2.39}$$

que es una constante ya que el número de elementos de la partición m es una constante. En el caso teórico era cuando $m \to \infty$, pero en este caso también es una constante y por tanto, es inmediato deducir cómo ya no tenemos ninguno de los problemas que hemos obtenido para la entropía de permutación a lo largo de esta sección, basta ver las demostraciones de todos los resultados dados en esta sección.

Resumiendo, el origen del problema con la entropía de permutación viene de que la pendiente máxima no sea constante (véanse las demostraciones de los resultados de esta sección), es decir, de que

$$\Delta H_k = H_k - H_{k-1} = \log k! - \log (k-1)! = \log k, \tag{2.40}$$

y esto a su vez si nos fijamos bien proviene de que el número de elementos de la partición no crezca en forma de potencia, sino que crezca de forma factorial, es decir, mientras que la entropía de bloque S_k crece como m^k, en el caso de la entropía de permutación H_k crece como $k!$ y la teoría que se desarrolló en la Sección 2.2 supone implícitamente que crece en forma de potencia.

Por lo tanto, se concluye que el problema que estamos teniendo no es un problema intrínseco de la entropía de permutación sino que aparece debido a que la entropía de permutación tiene una forma de crecimiento del *bin*eado de la forma $k!$ en vez de tenerlo en forma de potencia.

Es por ello que este problema lo va a tener cualquier tipo de entropía cuyo número de elementos de la partición crezca con la longitud del bloque k diferente a m^k donde m es el número de elementos de S_1 del tipo particular de entropía. Es decir, podría haber incluso entropías de bloque que tuvieran este problema. Bastaría definir el crecimiento de su *bin*eado de forma diferente a m^k. Cabe resaltar, no obstante, que hacer esto sería algo artificial para la entropía de bloque, pero no por ello imposible.

Por lo tanto, vamos a parar un momento y vamos a formalizar todo lo que acabamos de deducir en estas líneas previas.

Definición 2.5.2. *Sea H^* una entropía que verifica la Definición 2.1.2, es decir, cumple que $H^* = -\sum_{i=1}^{n} p_i \log(p_i)$ y sea $k \in \mathbb{N}$ un entero. Entonces, se define la sucesión de entropías H_k^* como $H_k^* = -\sum_{i=1}^{n_k} p_i \log(p_i)$, donde $n_k = n(k)$ es una función que depende de k que puede ser creciente, decreciente o constante a medida que aumenta k.*

Proposición 2.5.5. *Sea H_k^* la sucesión de entropías de la definición previa. Entonces,*

1.-La entropía de bloque corresponde a una sucesión de entropías con $n_k = n(k) = m^k$, siendo m el número de elementos de la partición del eje Y.

2.-La entropía de permutación corresponde a una sucesión de entropías con $n_k = n(k) = k!$.

Es decir, tanto la entropía de bloque como la entropía de permutación son una sucesión de entropías de Shannon con diferente número de elementos.

Demostración. 1.-Vamos a empezar pensando en el caso de una partición del eje Y finita (recordemos lo hablado en la Subsección 2.4.1). Supongamos que tenemos una partición de tamaño m del eje Y (nuestro alfabeto). En la entropía de bloque, k denota la longitud de los bloques seleccionados. Para $k = 1$ hay m posibilidades para el valor del eje Y del símbolo obtenido. Para $k = 2$

hay m posibilidades para el primer símbolo y m también para el segundo símbolo, por lo que hay m^2 posibilidades en general. Así, sucesivamente. Cuando hacemos $m \to \infty$ el argumento es el mismo.

2.-Para la entropía de permutación k denota la longitud del bloque seleccionado de nuevo. En este caso el número de elementos de cada partición coincide con el número de permutaciones posibles para esa longitud. Así, es inmediato ver que el número de elementos de la partición crece como $k!$, ya que el número de permutaciones posibles crece como $k!$. □

Ahora, enunciamos las propiedades especiales según cada tipo de crecimiento $n(k)$.

Proposición 2.5.6. *Sea H_k^* la sucesión de entropías de la definición previas. Entonces,*

1.-Si $n_k = n(k) = m^k$, entonces se cumple $\Delta H^*_{k\to\infty} = c \in \mathbb{R}^+$, $\Delta^2 H_k \leq 0$, es decir, tiene pendiente constante positiva de un índice k_0 en adelante y curvatura negativa.

2.-Si $n_k = n(k) = k!$, entonces $\Delta H^*_{k\to\infty}$ no es constante y $\Delta^2 H_k$ puede ser positiva.

Demostración. 1.-Proposición 2.2.3, Proposición 2.2.2.

2.-Proposición 2.5.1, Corolario 2.5.1. □

Y con esto ya podemos formular el resultado que buscábamos desde el principio de la sección.

Teorema 2.5.1. *La entropía de permutación detecta las correlaciones de forma equivalente a cómo las detecta la entropía de bloque pero mediante una escala diferente.*

Demostración. Este resultado se basa en que mediante la Proposición 2.5.5 se ha probado que la entropía de bloque y la de permutación miden lo mismo pero con diferentes escalas. Es absurdo pensar que una señal temporal esté más correlacionada que otra dependa de cómo se haga el *bin*eado.

Por poner un ejemplo, esta última afirmación sería lo análogo a afirmar que si se mide la temperatura en grados Celsius en el desierto del Sahara hace más calor que en Groenlandia, pero que si se hace en Kelvin entonces en Groenlandia hace más calor que en el desierto del Sahara. Esto es absurdo, ya que la temperatura no depende de la forma en que se calcule siempre que sea correcta. Lo mismo sucede con la entropía de bloque y la entropía de permutación.

De hecho, el ejemplo de la temperatura es muy relevante, ya que al igual que la temperatura inferior en Kelvin tiene una cota inferior en 0 y en grados Celsius en -273,15, la entropía de bloque tiene cota inferior 0, pero la entropía de permutación tiene cota inferior $-\infty$ en el caso de una señal aleatoria de longitud infinita. No obstante, como nunca se alcanza el caso de una señal de longitud infinita, la entropía de permutación tiene una cota inferior dada por cierto número negativo. □

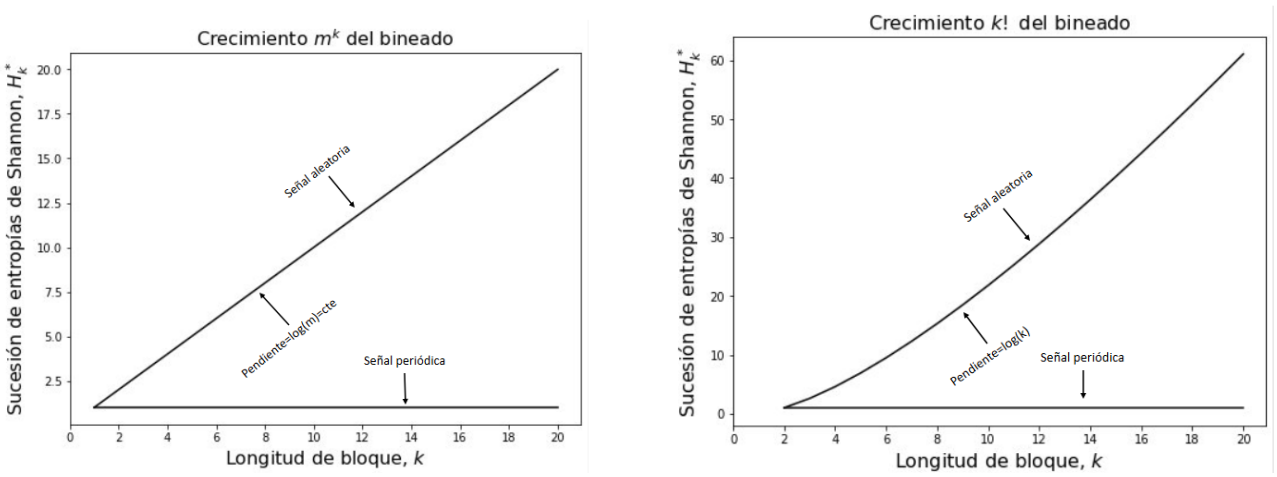
Corolario 2.5.2. *La definición*

$$\eta = \sum_{m=2}^{\infty}(m-2)k_m = -\sum_{m=2}^{\infty}(m-2)\Delta^2 H_m, \tag{2.41}$$

es válida en el caso de la entropía de permutación si el objetivo es ordenar las series temporales según su complejidad, la única diferencia es que $\eta \in [-c, \infty)$ con $c \in \mathbb{R}^+$, en vez de $\eta \in [0, \infty)$, donde el $-c$ proviene de la Proposición 2.5.4 y del hecho de que la longitud de las señales reales es finita. Es decir, el $-c$ depende de la longitud de la señal que se esté analizando. Por ejemplo, para $N = 2 \cdot 10^5$ se tiene que $-c \approx -1.84$.

Demostración. Inmediato del Teorema 2.5.1. □

A continuación vemos gráficamente la diferencia entre las dos situaciones.



(a) (b)

Figura 2.3: Representación de las propiedades de las curvas correspondientes a las señales completamente periódicas y aleatorias para cualquier sucesión de entropías de Shannon con un crecimiento del *bin*eado de la forma m^k en la subfigura (a) y de la forma $k!$ en la subfigura (b).

Se han proporcionado estas figuras con el objetivo de resumir gráficamente lo argumentado desde un punto de vista teórico en las líneas anteriores.

Probablemente las secciones previas le hayan resultado demasiado densas al lector, no obstante, era necesario hacerlas debido a que constituyen la base de nuestro método para detectar correlaciones. Sin embargo, es importante resaltar que de todas estas secciones lo importante para recordar es que la complejidad alcanza valores mínimos tanto para señales periódicas como para señales aleatorias y alcanza valores máximos a mitad de camino entre señales periódicas y aleatorias. A continuación, se proporciona un esquema gráfico sobre como varía la complejidad en función de la uniformidad de la señal:

Figura 2.4: Representación esquemática cualitativa sobre cómo varía la complejidad correspondiente a la entropía de permutación en función de la uniformidad de la señal para señales de longitud $N = 2 \cdot 10^5$.

La uniformidad de una señal se expresa de manera precisa como el grado de uniformidad promedio de la distribución de probabilidad de las permutaciones de la señal. Esta uniformidad alcanza el

mínimo para una señal completamente periódica donde solo hay un tipo de permutación y así la distribución de probabilidad de las permutaciones es una delta de Dirac, mientras que alcanza el máximo para una señal completamente aleatoria donde todas las permutaciones están representadas por igual, por lo que su distribución de probabilidad es la distribución uniforme. Cabe resaltar que se debe considerar la uniformidad promedio de la distribución de permutaciones, ya que por ejemplo una señal puede tener una distribución más uniforme que otra para una longitud de permutación k, pero tener una distribución menos uniforme para una longitud $k+1$.

Cabe destacar que dependiendo de la longitud de la señal, la cota inferior de la señal completamente aleatoria varía en el rango $[-\infty, 0]$, con $\eta = -\infty$ alcanzándose para una señal de longitud $N = \infty$ y $\eta = 0$ para una señal de longitud $N = 0$. En el caso de $N = 2 \cdot 10^5$ la cota inferior es de $\eta \approx -1.84$. La cota superior siempre es positiva y depende de cada sistema. Esta figura es característica de la entropía de permutación, ya que para la entropía de bloque recuérdese que tanto para una señal periódica como para una señal aleatoria se cumple $\eta = 0$.

Hay dos regímenes diferenciados, el régimen periódico y el régimen caótico. La delimitación de cada uno de ellos se ha hecho de manera cualitativa, no obstante, siempre se cumple que el máximo de la complejidad se alcanza para una señal perteneciente al régimen caótico. Asimismo, el crecimiento no tiene por qué ser en forma de parábola.

2.6. Método

Por todo lo argumentado en la sección previa deducimos que la forma en que debemos calcular la complejidad es en la forma

$$\eta = \sum_{m=2}^{\infty}(m-2)k_m = -\sum_{m=2}^{\infty}(m-2)\Delta^2 H_m, \tag{2.42}$$

donde $\Delta^2 H_m$ es la diferencia de orden 2 de la entropía de permutación.

No obstante, al hacerlo nos damos cuenta que nos surge un problema y es que debido al Teorema 2.3.1 no podemos llegar hasta longitudes de permutación $m \to \infty$, sino que debemos alterar la Definición 2.5.1 en la forma

$$\eta = \sum_{m=2}^{m_{max}}(m-2)\cdot\Delta^2 H_m, \tag{2.43}$$

donde m_{max} es el último valor no afectado por el problema de tamaño finito.

Hay que resaltar que el m_{max} no es aquel donde primero se alcance el valor de saturación dado por el Teorema 2.3.1, sino que es aquel primer valor para el cual el cual el valor de la entropía de permutación se vea afectada aunque sea de forma mínima por el problema de tamaño finito o la falta de estadística.

Por tanto, como $\Delta^2 H_m = \frac{H_{m+2}-2H_m+H_{m-2}}{4}$ lo único que nos queda por determinar es cómo calcular m_{max}, es decir, el primer valor que ya se ve afectado por la falta de estadística asociada a la longitud finita de la serie temporal, lo cual hacemos en la siguiente sección.

2.6.1. Primer Índice Afectado por Problema de Tamaño Finito

Para resolver este problema vamos a tomar el siguiente enfoque. Vamos a considerar como en la Proposición 2.5.1 el caso en que estamos analizando una señal completamente aleatoria.

Hacemos esto ya que sabemos que el comportamiento teórico de una señal aleatoria es $f(n) = \log(n!)$ con $n \in \mathbb{N}$ como demostramos en la Proposición 2.5.1. Así, una posible forma de determinar cuál es el primer índice para el cual afecta es por un lado calcular esta $f(n) = \log(n!)$, por otro calcular computacionalmente los valores de la entropía de permutación para una señal aleatoria, establecer un criterio de separación como por ejemplo 0.01 y calcular para cada longitud de la señal N cual es el valor de H_k para el cual empieza la separación y luego generalizar este valor a cualquier tipo de señal temporal de longitud N basándonos en que el problema de tamaño finito solo depende de la longitud de la señal temporal N como se ha probado en el Lema 2.3.1 y en el Teorema 2.3.1.

Este enfoque es un enfoque válido, lo único que tiene el inconveniente por un lado de que para valores de N pequeños como se alcanza rápido el problema de tamaño finito, los puntos pueden distar $H_k = 1 - 2$ entre sí, lo que hace que este método sea impreciso, ya que estaríamos diciendo por ejemplo que el valor de separación es de $H_k = 12 \pm 2$ por ejemplo. El otro inconveniente es que habría que determinar esto para todo N con el que se quiera trabajar. A parte de que en general se podría calificar a este método como bastante bruto o rudimentario.

Sin embargo, hay una forma más precisa de hacer esto. Para ello primero identificamos el problema matemáticamente.

(P) Queremos determinar la distribución teórica que sigue $(p_1, ..., p_k)$ teniendo N muestras, es decir, tenemos las siguientes N muestras de vectores n-dimensionales:

$$N \begin{cases} (p_{1,1}, ..., p_{n,1}) \\ ... \\ (p_{1,i}, ..., p_{n,i}) \\ ... \\ (p_{1,N}, ..., p_{n,N}) \end{cases}$$

y queremos saber la distribución poblacional o teórica de nuestras entropías de longitud n a partir de las N muestras de las que disponemos, es decir, queremos determinar el N mínimo suficiente para cada longitud de permutación n, ya que si el N no es suficiente es que hay mala estadística y por tanto, aparece el problema de tamaño finito.

En otras palabras, se pretende inferir la distribución de probabilidad poblacional o teórica discreta $\mathbb{P}[X = x_i] = p_i$ con $i \in \{1, ..., n\}$ a partir de N muestras.

Este es un problema estudiado recurrentemente en la literatura, incluso en el ámbito de la entropía de permutación [10]. En [10] se da como condición que $5 \cdot n! \lesssim N$, con lo que resolvemos el problema y así tomamos el n_{max} mayor que verifique

$$5 \cdot n_{max} \leq N, \tag{2.44}$$

como el último hasta donde aplicar la definición

$$\eta = \sum_{n=2}^{n_{max}} (n-2) \cdot \Delta^2 S_n. \tag{2.45}$$

No obstante, por lo demostrado en el Teorema 2.3.1 el problema de tamaño finito es un fenómeno que solo depende de la longitud de la señal N y que se refleja en que hay una cota superior en el valor de la entropía de permutación, H_k.

Es por esto que todavía nos falta una última cosa por hacer antes de haber definido completamente el método, ya que debemos relacionar la longitud de permutación n_{max} donde tenemos la cota que

muestra hasta dónde nuestros datos no están afectados, con el valor de la entropía de permutación, ya que para una señal aleatoria para el n_{max} podemos estar en altura $H_k = 15$, mientras que para una señal más periódica podemos estar en altura $H_k = 1$ para una misma longitud de la señal N y la forma que tiene de reflejarse el problema de tamaño finito es a través del valor de la entropía de permutación H_k^{max}, no del n_{max}.

Para solucionar esto nos situamos en el caso peor en términos de tamaño finito que será aquella situación donde se alcance el valor mayor de H_k para un n menor. Por la Proposición 2.4.1 es inmediato deducir que este caso peor es cuando tenemos una señal aleatoria. Por tanto, se puede deducir que el H_k^{max} donde aparece el problema de tamaño finito para cualquier señal de longitud N es el H_k que se alcanza cuando en la señal aleatoria se sustituye el n_{max} obtenido de la Ecuación (2.44).

Por lo tanto, el método propuesto se basa en calcular

$$\eta = \sum_{n=2}^{n_{max}} (n-2) \cdot \Delta^2 H_n, \tag{2.46}$$

donde el n_{max} es el índice correspondiente a la altura H_k^{max} definida anteriormente. Cabe resaltar que mientras que el H_k^{max} es invariante para todas las señales de la misma longitud, el n_{max} no lo es, por todo lo argumentado anteriormente.

A continuación siguiendo el método dado en las líneas previas damos valores de H_k^{max} para algunos valores de N comunes en este texto.

N	10^3	10^4	10^5	$2 \cdot 10^5$	10^6
n_{max}	5	6	7	7	8
H_k^{max}	6.90	9.48	12.29	12.29	15.29

Tabla 2.2: Valores de n_{max} y H_k^{max} para diferentes valores de N usados en el texto.

El lector podría pensar que esta Subsección es una sutileza teórica sin mayor importancia al llevarlo a la práctica. Podría pensar, que el tomar un índice más o un índice menos no debe tener importancia en la práctica. Si el lector piensa esto, se le recomienda encarecidamente que observe la Figura 5 del Apéndice C.2 y lea su breve discusión posterior.

Capítulo 3

Datos Simulados: La Aplicación Logística

En este capítulo se comprueba la validez del método introducido en la Sección 2.6 mediante datos simulados con la aplicación logística. Se divide el capítulo en 3 secciones. En la Sección 3.1 se introduce la aplicación logística, en la Sección 3.2 se exponen los resultados obtenidos con nuestro método y en la Sección 3.3 se cuantifica la reducción del problema de tamaño finito al pasar de la entropía de bloque a la entropía de permutación.

3.1. Introducción a la Aplicación Logística

La aplicación logística es uno de los ejemplos más simples de una aplicación no lineal y así, uno de los modelos más simples de dinámica caótica. Esta ecuación logística se puede expresar en forma de ecuación diferencial en la forma

$$\frac{dx}{dt} = bx(1-x), \tag{3.1}$$

donde el tiempo se considera como una variable continua [11]. No obstante, también se puede expresar como una ecuación cuadrática de recurrencia, así considerando el tiempo como discreto

$$x_{n+1} = bx_n(1-x_n), \tag{3.2}$$

donde b es una constante positiva a la cual se le suele conocer por el nombre de potencial biótico. Este nombre proviene de que la aplicación logística a parte de ser una ecuación muy útil para describir un sistema complejo mediante una relación simple, es una ecuación que describe el crecimiento de una población con el tiempo. Valores pequeños hacen que el crecimiento de una población sea pequeño, lo que conlleva a su extinción. Valores más altos en cambio llevan a que la población se asiente en un valor dado o a que fluctúe en torno a ciertos valores que representan épocas de población extraordinariamente alta o épocas de población extraordinariamente baja [11].

Ya en los años 1940 John Von Neumann sugirió usar esta aplicación logística con $b = 4$, es decir, $x_{n+1} = 4x_n(1-x_n)$ como generador de números aleatorios. No obstante, no fue hasta el trabajo de W.Ricker de 1954 y los estudios de la década de 1950 de Paul Stein y Stanislav Ulam para que las notables propiedades de la aplicación logística se conocieran [12].

Cabe destacar que en general, para un cierto valor arbitrario de b no se sabe calcular analíticamente las soluciones, sin embargo, en [12] se postula que cualquier solución exacta debe ser de la forma

$$x_n = \frac{1}{2}\{1 - f[r^n f^{-1}(1-2x_0)]\} \tag{3.3}$$

donde f es cierta función, f^{-1} su inversa y x_0 el valor inicial.

La forma de estudiar las aplicaciones logísticas es a partir de los llamados diagramas de bifurcación, que muestran los valores visitados o a los que hay un acercamiento asintótico (puntos fijos, órbitas periódicas o atractores caóticos) de un sistema en función del parámetro de bifurcación del sistema bajo estudio, que en el caso de la aplicación logística es el parámetro b [13].

Los puntos fijos, órbitas periódicas y atractores caóticos son algunos de los posibles comportamientos que tiene un sistema dinámico [6], los cuales se definen a continuación.

Definición 3.1.1. *Supongamos que tenemos un sistema discreto en tiempo, con estados $x(t) \in \mathbb{R}$, con una dinámica dada por la aplicación:*

$$x(t+1) = f(x(t)), \tag{3.4}$$

siendo f una función real diferenciable, $f : \mathbb{R} \to \mathbb{R}$. Entonces,
1) Se dice que $x \in \mathbb{R}$ es un punto fijo de f si y solo si $f(x) = x$.
2) Se dice que es una órbita periódica si $x(T) = f^T(x(0)) = x(0)$ y $x(k) \neq x(0)$ para todo $0 < k < T$.
3) Se dice que un atractor es aquel valor o conjunto de valores en donde el sistema se asienta después de cierto paso de tiempo. Por otro lado, se dice que el sistema tiende a un atractor caótico cuando el sistema tiene una oscilación que no cesa, a través de la cual nunca visita el mismo punto dos veces y tiene una estructura fractal, lo cual significa que existen los mismos patrones independientemente del zoom que se haga [6].

Cabe resaltar que se suele tomar un valor inicial de la población $x_0 \in (0, 1)$ y $b \in [0, 4]$, para que así $x_m \in [0, 1]$ para cualquier $m > 0$. Además, se puede comprobar que para valores de b menores que 1, el sistema siempre tiende a 0 (extinción) y para valores de b entre 1 y 3, el sistema siempre se asienta en un único valor del intervalo $[0, 1]$.

A partir de $b = 3$ es donde la tendencia empieza a cambiar ya que aparece la primera bifurcación. Hasta $b = 3.4$ los posibles valores de la población siguen siendo 2. Justo después de $b = 3.4$, el diagrama vuelve a bifurcarse, dando lugar a 4 posibles caminos que puede trazar la población. Luego, ligeramente después de 3.5 vuelve a bifurcase de nuevo, dando lugar a 8 posibles caminos. Aquí, la población oscila en torno a 8 posibles valores [11].

A partir de aquí, en $b \approx 3.6$, se transiciona al régimen caótico de la aplicación logística. Aquí las bifurcaciones crecen tanto, que el sistema pasa de 4, a 8, a 16, 32... posibles valores de la población. Se dice que en $b \approx 3.6$ hay una transición del régimen periódico al régimen caótico. Si una población para cierto valor de b tiene k posibles valores, se dice que tiene periodo k. En $b = 3.9$, la bifurcación ha sido tan grande que el sistema puede tomar casi cualquier valor en el intervalo $[0, 1]$. Esto es lo que se conoce como ruta del caos por duplicación del periodo en el contexto de los sistemas dinámicos.

Por todo esto se puede considerar que la zona de interés de la aplicación logística se sitúa entre los valores de b de entre 3.5 y 4. En consecuencia, se muestra el diagrama de bifurcación para $b \in [3.5, 4]$.

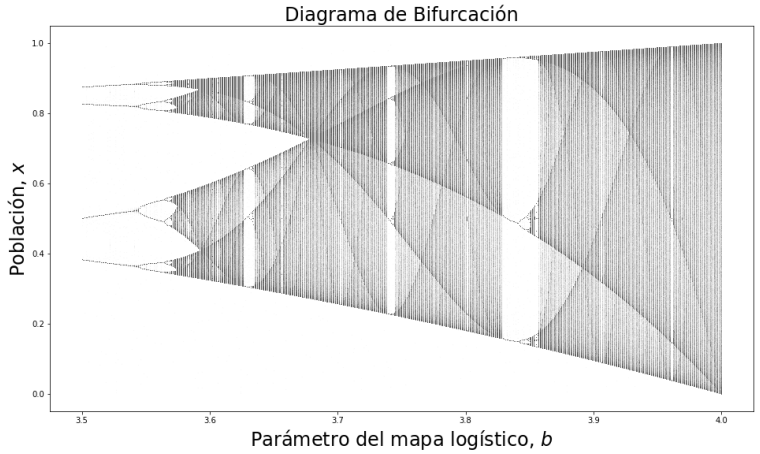


Figura 3.1: Diagrama de bifurcación para valores de b entre 3.5 y 4.

Como se puede observar en la figura a medida que aumenta el valor de b de 3.5 a 4 el número de posibles valores que visita el sistema dinámico va creciendo, hasta que en $b = 4$ los posibles valores de la población abarcan prácticamente todo el intervalo $[0, 1]$. No obstante, pese a esta tendencia hacia cada vez más caos, se puede observar como intermitentemente hay algunas ventanas de orden, en las cuales los posibles valores de la población se reducen a unos pocos. Por ejemplo, hay una pequeña ventana de orden sobre $b = 3.63$, otra sobre $b = 3.75$, no obstante, la ventana de orden más grande es la que se observa entre $b = 3.82$ y $b = 3.84$ donde se puede observar cómo los únicos posibles valores de la población al principio de esta ventana de orden son aproximadamente 0.15, 0.55 y 0.95. Si se genera una señal con la aplicación logística con un parámetro b de una ventana de orden, entonces esta señal será periódica o cuasiperiódica. Cabe resaltar que la figura muestra de manera muy evidente la transición del régimen periódico al régimen caótico de $b \approx 3.6$.

3.2. Uso de la Entropía de Permutación en la Aplicación Logística

En resumidas cuentas, el interés de la aplicación logística reside en que es una aplicación simple a través de la cual se puede estudiar el caos únicamente variando el parámetro b de su definición. A parte la aplicación logística ha sido muy estudiada por la literatura, por lo que se conoce perfectamente su comportamiento. Es por esto que se ha elegido esta aplicación para probar la validez del método propuesto en la Sección 2.6.

Para ello, se ha calculado la entropía de permutación en función de la longitud de permutación para diferentes parámetros b de la aplicación logística para comprobar que el método funciona adecuadamente y que es capaz de discernir aquellas series temporales que están más correlacionadas de aquellas que lo están menos, usando las técnicas introducidas en el Capítulo 2 y luego comparando el resultado con lo dado en el diagrama de bifurcación de la Figura 3.1. Como la región de interés de la aplicación logística es el intervalo $[3.5, 4]$, se ha decidido estudiar los casos $b = 3.5, 3.55, ..., 3.95, 4$.

La aplicación logística muestra una mayor o menor correlación y por tanto, complejidad dependiendo de su parámetro b. Hasta aproximadamente $b = 3.58$ la complejidad es nula o aproximadamente nula, ya que las señales temporales generadas son periódicas o cuasiperiódicas. En, $b \approx 3.6$ al pasar del régimen periódico al régimen caótico es cuando alcanza la complejidad máxima y de $b = 3.6$ en adelante su complejidad va decreciendo de manera monótona, debido a que las series temporales generadas van tendiendo a ser cada vez más aleatorias (recuérdese la Figura 2.4 y que la complejidad es mínima para señales periódicas y aleatorias y máxima para señales a mitad de camino, es decir,

para señales localizadas al inicio del régimen caótico). Este decrecimiento monótono se rompe en algunos puntos. Estos puntos son el $b \approx 3.63$ donde la complejidad pasa a ser 0 repentinamente debido a la aparición de una ventana de orden que hace que las señales temporales generadas aquí sean periódicas. En el límite de esta ventana de orden en $b \approx 3.64$ se da lo que se conoce como el límite del caos o *edge of caos* que tiene asociada una correlación anormalmente alta y por tanto, una complejidad anormalmente alta. Finalmente, de $b \approx 3.65$ en adelante sigue el decrecimiento monótono.

Esta ventana de orden no es la única. En el diagrama de bifurcación de la Figura 3.1 se puede apreciar otra en $b \approx 3.74$ y otra más grande que las dos anteriores en $b \approx 3.82$. El comportamiento de la complejidad en estas dos es igual al de la ventana de orden de $b \approx 3.63$, es decir, una caída a aproximadamente 0 de la complejidad debido a la ventana de orden y luego una subida repentina debido a lo que se conoce como límite del caos.

Antes de empezar con el análisis en sí, vamos a comprobar la validez de nuestro método computacional descrito en el Apéndice C, rehaciendo las figuras de h_6 y h_{12} en 501 puntos equiespaciados en el intervalo $[3.5, 4]$ dadas en [8], donde h_6 representa la entropía por símbolo de longitud de permutación 6 de la Definición 2.4.2 y h_{12} la de longitud de permutación 12.

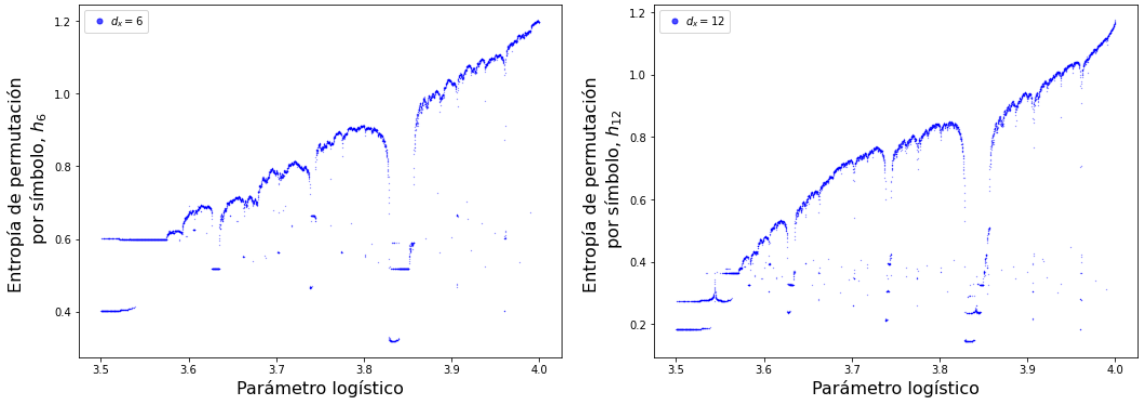


Figura 3.2: Representación de las entropías por símbolo para longitudes de permutación de 6 y 12 respectivamente. Estas figuras son exactamente iguales a las dadas en [8].

Una vez comprobada la validez de las técnicas computacionales descritas en el Apéndice C, comenzamos representando la entropía de permutación H_k en función de la longitud de permutación k para mostrar gráficamente algunas de las propiedades introducidas teóricamente en el Capítulo 2. Se van a representar 4 curvas correspondientes a aplicar la entropía de permutación a señales de longitud $N = 10^k$ con $k \in \{3, 4, 5, 6\}$. En estas representaciones según lo introducido en la Sección 2.3 a través del Teorema 2.3.1 se esperaría ver 4 curvas tomando valores similares, pero que se diferencian en que saturan en los valores de la Tabla 2.1 según su correspondiente longitud N. Comprobémoslo.

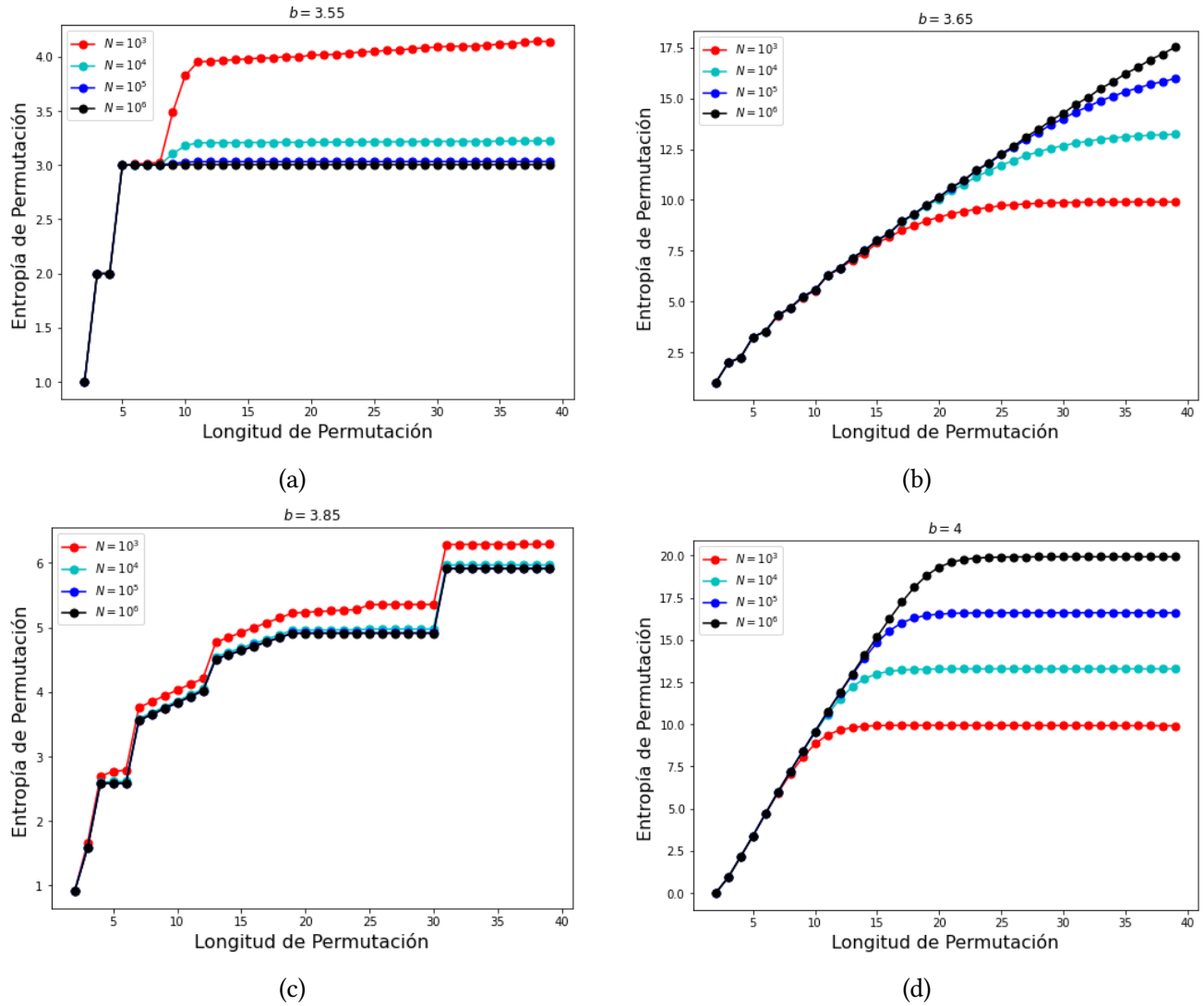


Figura 3.3: Representación de la entropía de la permutación H_k en función de la longitud de permutación k para $b = 3.55$ en la subfigura (a), $b = 3.65$ en la subfigura (b), $b = 3.85$ en la subfigura (c) y $b = 4$ en la subfigura (d), para señales temporales de $N = 10^k$ con $k \in \{3, 4, 5, 6\}$.

A primera vista estas 4 representaciones parecen muy distintas entre ellas y es por ello que nos sirve para ilustrar lo introducido teóricamente en el Capítulo 2.

Lo primero a destacar es que en los 4 casos estudiados las longitudes de permutación van de 2 a 40 y sin embargo, se observa que en el único donde se han estabilizado todos los valores de H_k según lo estipulado en el Teorema 2.3.1 y la Tabla 2.1 es el caso de $b = 4$.

Para el caso $b = 3.65$ se puede ver cómo solo han saturado las curvas de $N = 10^3$ y $N = 10^4$, mientras que para los casos de $b = 3.55$ y $b = 3.85$ se puede ver cómo todavía les queda un largo recorrido antes de saturar, es decir, el problema de tamaño finito aparece en un k notablemente mayor. Esto es debido a que las señales correspondientes a $b = 3.55$ y $b = 3.85$ son cuasiperiódicas, con complejidad aproximadamente nula.

Es importante subrayar este hecho y será relevante en lo sucesivo, ya que se verá cómo la longitud de permutación k donde se alcanza el problema de tamaño finito está íntimamente relacionado con la complejidad. Solo hace falta ir a la Definición 2.5.1 para deducirlo.

Es importante resaltar, a partir de la inspección de las Figuras 3.3b y 3.3d, que aunque parece existir una región clara donde se puede hacer un ajuste lineal, la Proposición 2.5.1 prueba que tal

asíntota lineal no existe para la entropía de permutación.

Observación 3.2.1. *Comparemos los valores de saturación de la Figura 3.3d con los predichos teóricamente en el Teorema 2.3.1.*

N	10^3	10^4	10^5	10^6
Teórico	9.9658	13.2877	16.6096	19.9316
Numérico	9.9248	13.2837	16.6091	19.9310

Tabla 3.1: Comparación entre los valores predichos por el Teorema 2.3.1 como $\log N$ y los valores determinados en simulación numérica calculados como el último valor de H_k de las curvas de $N = 10^3$, $N = 10^4$, $N = 10^5$ y $N = 10^6$ para el caso $b = 4$ de la Figura 3.3d.

Recordemos que lo que el Teorema 2.3.1 predice es una cota superior que H_k debe de alcanzar para longitudes de bloque k suficientemente grandes.

Una vez realizado este análisis de las Figuras 3.3a, 3.3b, 3.3c y 3.3d con el objetivo de transmitir visualmente algunos de los resultados más importantes de la teoría dada en el Capítulo 2, pasemos a lo que es realmente el objetivo de este capítulo que es comprobar que el método introducido en la Sección 2.6 da resultados de complejidad compatibles con el diagrama de bifurcación de la Figura 3.1, es decir, compatibles con los estudios previos de la literatura. Cabe destacar que solo se van a analizar señales correspondientes al régimen caótico.

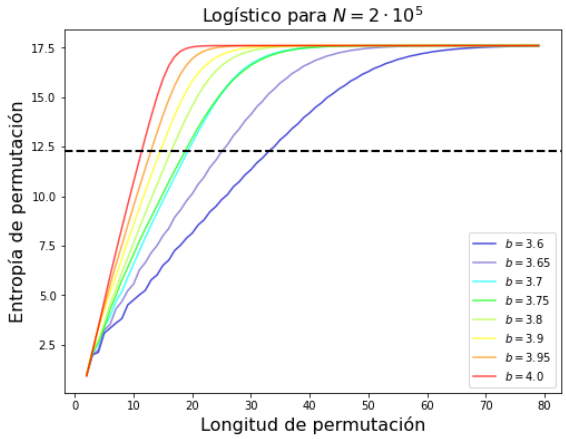


Figura 3.4: Representación de la entropía de permutación H_k en función de la longitud de permutación k para señales temporales de longitud $N = 2 \cdot 10^5$ generadas mediante la aplicación logística con $b \in \{3.6, 3.65, 3.7, 3.75, 3.8, 3.9, 4\}$. La línea discontinua negra representa el punto a partir cuál los valores están afectados por el problema de tamaño finito.

A continuación, calculamos la complejidad de la aplicación logística aplicando el método introducido en la Sección 2.6 a las curvas de la Figura 3.4 y comparamos con el diagrama de bifurcación de la Figura 3.1, para deducir la validez del método.

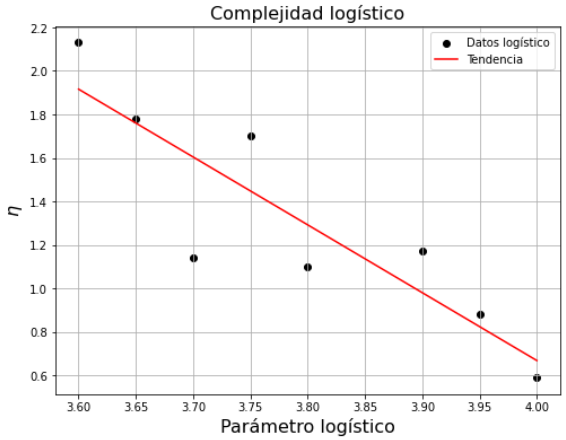


Figura 3.5: Representación de la complejidad de señales temporales generadas a través de la aplicación logística con los parámetros $b \in \{3.6, 3.65, 3.7, 3.75, 3.8, 3.9, 3.95, 4\}$.

Hay clara tendencia monótonamente decreciente a medida que aumenta el parámetro b de la aplicación logística, que concuerda con lo esperado. Se ha obtenido que para $b = 3.6$ y $b = 3.65$ la complejidad es notablemente mayor que el resto como se esperaba con la excepción del caso $b = 3.75$. No obstante, también se espera una complejidad alta para este caso, debido a que como se ha explicado antes en $b \approx 3.74$ hay una ventana de orden y en $b \approx 3.75$ es el límite del caos. En consecuencia, los resultados obtenidos son coherentes con los esperados.

Por último se representa el límite de la longitud de permutación antes de que aparezca el problema de tamaño finito, es decir, el k_{max} de cada b asociado a la altura H_k^{max} dado en la Tabla 2.1 que en el caso de $N = 2 \cdot 10^5$ es $H_k^{max} = 12.29$.

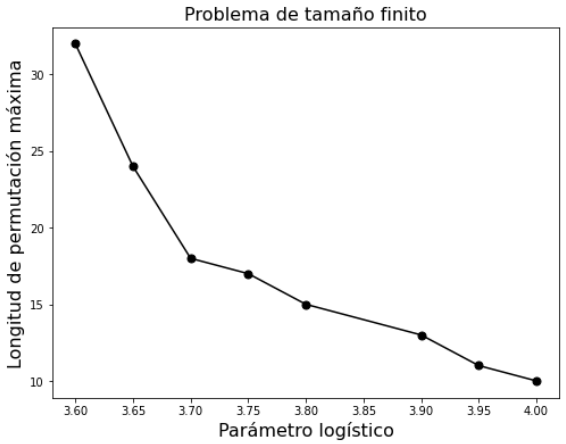


Figura 3.6: Representación de la longitud de permutación máxima k_{max} para series logísticas de longitud $N = 2 \cdot 10^5$, donde k_{max} representa el tamaño de bloque más largo para el que puede estimarse de manera fiable la entropía de permutación a partir de series de longitud finita N.

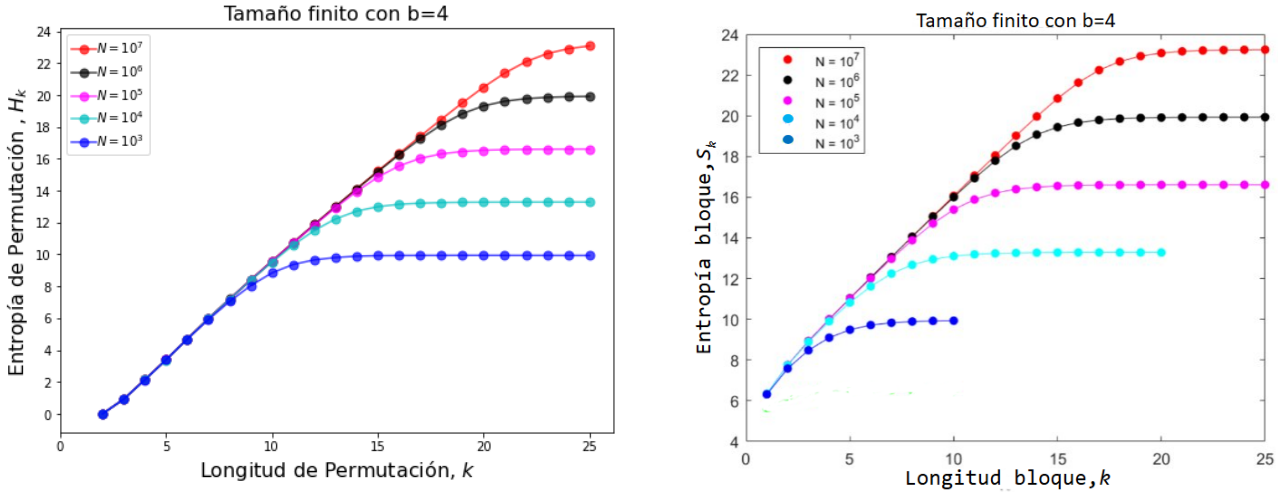
No es sorprendente la clara tendencia que se puede observar en esta figura, ya que no hay más que irse a la Definición 2.1.2 y a la Definición 2.4.1 de donde se puede deducir que para la señales más aleatorias en cada longitud de permutación la entropía de permutación es mayor y así las curvas de H_k en función de la longitud de permutación k son mayores, lo que hace que se alcance la altura límite dada por la Tabla 2.1 mucho antes para estas señales aleatorias que para las que están más correlacionadas.

3.3. Comparación de Problemas de Tamaños Finitos

Como ya se ha comentado en la Subsección 2.4.1 del Capítulo 2 las razones para escoger la entropía de permutación frente a la más clásica entropía de bloque, aunque para esta segunda se disponga de un marco teórico completo en la literatura mientras que para la primera no, es por un lado que se elimina el problema de discretización y por otro que se reduce el problema de tamaño finito. Veamos entonces cuánto se reduce el problema de tamaño finito.

Para ello nos basamos en que el Teorema 2.3.1 es igualmente válido tanto para la entropía de permutación como para la entropía de bloque, por lo que las alturas H_k^{max} son iguales en ambos casos y vienen dadas por la Tabla 2.1. Lo que cambia en ambos casos es el k_{max} asociado a estos H_k^{max}.

Así, escogemos un b particular como por ejemplo $b = 4$ y comparamos los k_{max} asociados a los H_k^{max} en el caso de la entropía de permutación y en el caso de la entropía de bloque.



(a) (b)

Figura 3.7: Representación del problema de tamaño finito al generar a través de la aplicación logística con $b = 4$ series temporales de longitudes $N = 10^k$ con $k \in \{3, 4, 5, 6\}$ para el caso de la entropía de permutación en la subfigura (a) y el caso de la entropía de bloque en la subfigura (b). La subfigura (b) se ha tomado de la referencia [1].

Observando la Figura 3.7 se deduce que las longitudes de permutación k donde se alcanza el problema de tamaño finito (saturación del valor de la entropía) en el caso de la entropía de permutación es aproximadamente un $50\,\%$ mayor que para el caso de la entropía de bloque. En consecuencia, el problema de tamaño finito se ha reducido en aproximadamente un $50\,\%$.

Capítulo 4

Datos Reales: Plasmas Generados en SLPM

Una vez comprobada la validez del método introducido en la Sección 2.6 mediante datos simulados, se procede a analizar señales temporales reales obtenidas del plasma del Santander Linear Plasma Machine (SLPM) para comprobar que nuestro método también da resultados coherentes para este caso. Se divide el capítulo en 2 secciones. En la primera Sección 4.1 se describe el tipo de señales que se van a analizar, y en la Sección 4.2 se exponen los resultados obtenidos con nuestro método.

4.1. Señales Temporales de SLPM

Las señales temporales disponibles de SLPM tienen longitud $N = 2 \cdot 10^5$ y han sido medidas en 10 radios diferentes de la columna de plasma mediante sondas de Langmuir, donde los 10 radios van desde el $r = 1.0$ cm hasta el $r = 2.8$ cm distribuidos uniformemente. En el Capítulo 1 se describió cómo el plasma se aloja dentro de una cámara de vidrio cilindríca de 1 m de longitud y 7 cm de diámetro. El rango de radios entre 1 cm y 2.8 cm abarca la mayor parte de los 3.5 cm de radio. El lector podría preguntarse por qué no se mide para radios inferiores a 1 cm o superiores a 2.8 cm o mejor dicho, aunque se midieran por qué no se consideran para el análisis en este trabajo.

Por un lado, para radios inferiores a 1.0 cm, la sonda Langmuir y su soporte perturban demasiado las mediciones, por lo que se descartan ya que no podemos asegurar su validez. Para los radios mayores a 2.8 cm se descartan simplemente porque se puede considerar que no hay plasma. Es decir, el plasma está en un recipiente cilíndrico de 3.5 cm de radio de manera que en el centro es donde se encuentra la densidad máxima de partículas y al alejarse del centro esta densidad va disminuyendo de forma que para radios mayores que $r = 2.8$ cm, el plasma está suficientemente enrarecido como para considerarlo como tal.

Por otro lado, cabe destacar que se trabaja con unidades arbitrarias del flujo a las que se denota como $u.a.$. La razón es que la entropía de permutación se basa en ordenar los bloques de la señal temporal de mayor a menor o de menor a mayor, por lo que las unidades no afectan. Matemáticamente hablando, la entropía de permutación es invariante ante cambios de escala de la señal temporal, lo que hace irrelevantes las unidades de las señal temporales.

Además, cabe resaltar que la señal de longitud $N = 2 \cdot 10^5$ ha sido medida en un intervalo temporal de $t = 0.2$ s con distribución uniforme. Así, se deduce que el paso de tiempo entre cada par de mediciones consecutivas ha sido de $t = 1$ μs, o equivalentemente la frecuencia de muestreo de 1 MHz.

A continuación, se representa la señal temporal para $r = 1.0$ cm, a partir de la cual se anali-

zarán sus propiedades generales.

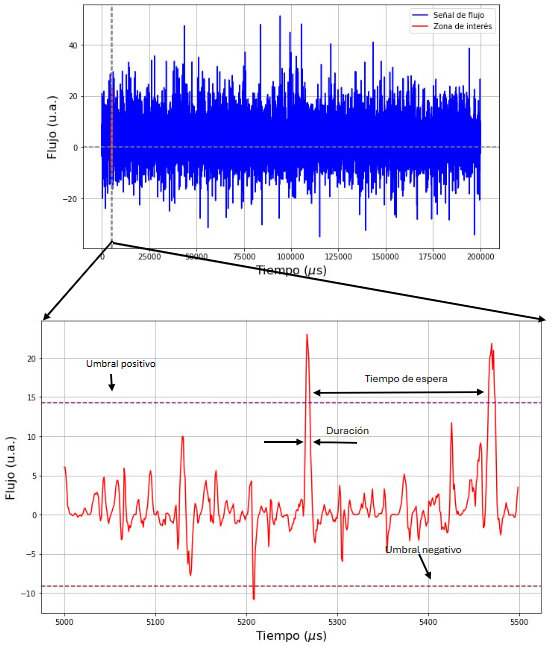


Figura 4.1: Representación de una señal temporal de longitud $N = 2 \cdot 10^5$ generada en SLPM para $r = 1.0$ cm. En la parte superior se representa la señal de flujo completa y en la parte inferior se representa únicamente una parte de longitud $N = 5 \cdot 10^3$. Las líneas horizontales discontinuas son umbrales necesarios para identificar avalanchas.

Los umbrales positivo y negativo establecen un criterio arbitrario pero con una razón sólida detrás dada en [4] para clasificar avalanchas. Todo lo que está por debajo del umbral negativo o encima del positivo se considera como avalancha. En este caso se observa que hay 3 avalanchas. La variable tiempo de espera cuantifica la distancia temporal entre cada par de estas avalanchas, la variable duración la duración de cada avalancha en el corte con el umbral positivo o negativo y además, se puede definir una tercera variable que cuantifica el área encerrada entre el umbral positivo o negativo y la curva para cada una de las avalanchas.

Cabe recordar que en el Capítulo 2 se hizo el desarrollo teórico para señales temporales tal que su probabilidad no cambia con el tiempo, es decir, señales estacionarias. Así, falta comprobar que las señales temporales del SLPM verifican esta propiedad de estacionariedad. Se define a continuación el concepto de estacionariedad.

Definición 4.1.1. [14] *Sea $X_{t_1}...X_{t_n}$ un proceso estocástico o señal temporal. Se dice que es un proceso estocástico estacionario si y solo si verifica que $X_{t_1}...X_{t_k}$ y $X_{t_1+\tau}...X_{t_k+\tau}$ tienen la misma función de distribución empírica (PDF) para cualquier conjunto de índices $\{1, ..., k\}$ y cualquier $\tau \leq n$.*

Siendo estrictamente rigurosos, para comprobar que un proceso estocástico o señal temporal es estacionaria, habría que comprobar que la función de densidad empírica no varía con el tiempo. No obstante, según [14] a efectos prácticos se suele comprobar la estacionariedad de los momentos de primer y segundo orden, lo que se traduce en comprobar que la media y la varianza no varían con el tiempo o equivalentemente, que la media y la desviación típica no varían con el tiempo.

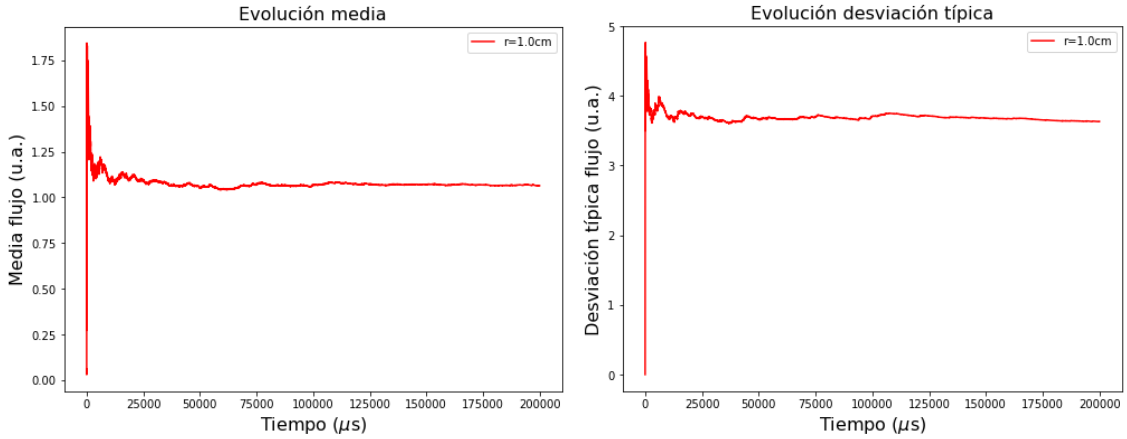


Figura 4.2: Evolución de la media y la desviación típica del flujo medido con la sonda colocada en $r = 1.0$ cm.

Como se desprende de la Figura 4.2, tanto la media como la desviación típica son constantes. Se ha procedido de forma análoga con los restantes 9 radios y se han obtenido resultados similares. Por lo tanto, se ha podido concluir que efectivamente se está trabajando con procesos estocásticos estacionarios como requiere la teoría del Capítulo 2.

4.2. Detección de Correlaciones en Señales Temporales del SLPM

En esta sección se trata de detectar el grado de correlación en secuencias temporales obtenidas en distintos radios de la columna de plasma generada en SLPM. Después de todo lo explicado en el Capítulo 2, es fácil deducir que hay un problema si se pretende detectar cuál de las 10 señales de flujos está más correlacionada ya que en el entorno central (aproximadamente entre valores de flujo de -10 u.a. y 10 u.a.) hay una gran concentración de fluctuaciones aleatorias, que se traduce en que las frecuencias de las diferentes permutaciones van a ser muy homogéneas, independientemente del radio, lo que supone un problema, ya que según la Definiciones 2.4.1 y 2.5.1 resultará en que todas las señales tendrán una complejidad muy baja. Es decir, las fluctuaciones del entorno central enmascaran la correlación asociada a las avalanchas. Esto es debido a que en porcentaje el transporte turbulento tiene muy poco peso frente al difusivo. Pero las correlaciones aparecen en el transporte turbulento y no en el difusivo.

Veamos que este argumento teórico en efecto se traslada a la representación de la entropía de permutación.

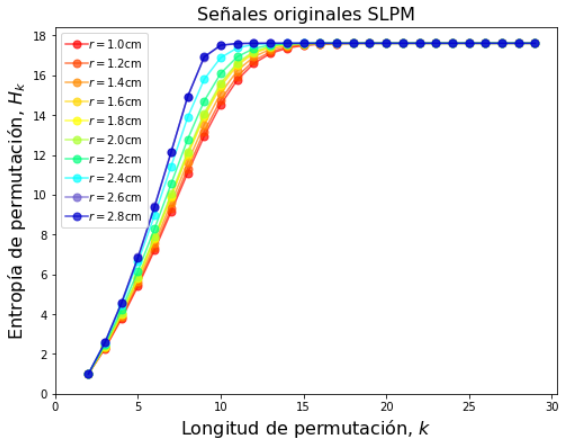


Figura 4.3: Representación de las entropías de permutación en función de la longitud de permutación calculadas a partir de las señales originales en crudo.

Como se puede observar en la Figura 4.3 los resultados no permiten distinguir las señales según su correlación. Recordemos que en el Capítulo 3 a través de la Figura 3.6 se muestra cómo a mayor complejidad mayor es el índice donde aparecen los primeros efectos del problema de tamaño finito (índice de corte con $H_k^{max} = 12.29$) y también que para el caso más aleatorio de la aplicación logística ($b = 4$) este índice es 10. Además, en resultados previos [1] se muestra que este caso es realmente aleatorio, ya que se representa su PDF y se observa como es casi uniforme. En esta Figura 4.3 estos índices están entre 5 y 10, y por lo tanto se deduce que las señales que se han analizado son casi completamente aleatorias, como se había previsto.

Para resolver este problema se debe pensar en qué se está analizando: señales temporales provenientes de un plasma cuya dinámica se puede describir mediante dos canales de transporte.

Por un lado está el transporte difusivo y por otro, el transporte no difusivo o mediante avalanchas. Mientras que el transporte difusivo reduce las diferencias y gradientes, eliminando las correlaciones entre distintos instantes temporales, el transporte por avalanchas en cambio genera correlaciones en la señal, ya que el transporte mediante avalanchas está generado por inestabilidades que vienen de instantes temporales precedentes. Inestabilidades que el transporte difusivo no ha sido capaz de eliminar.

Mientras que el transporte difusivo ocurre de forma continua en el tiempo, el transporte por avalanchas solo ocurre en determinados instantes en los que una fluctuación se hace lo bastante grande como para trasladarse en cascada por una buena parte del sistema, es decir, un gradiente lo suficientemente grande como para causar una avalancha.

Como se puede observar en la Figura 4.1 hay 3 variables que caracterizan este transporte por avalanchas [4], el tiempo de espera entre dos avalanchas sucesivas, la duración de la avalancha y el área de la avalancha.

Cuando se mide el flujo en un punto del espacio en diferentes instantes de tiempo, una avalancha se define como un crecimiento o decrecimiento anormalmente alto del flujo en ese punto [4]. Sin embargo, esta definición es arbitraria y no cuantitativa. Es por ello que se debe establecer un criterio para determinar qué es una avalancha y qué no lo es. En [4] se establece como criterio calificar como avalancha todo aquél proceso que supere el umbral positivo o el umbral negativo de la Figura 4.1 donde el umbral positivo se define como

$$\Gamma_t^+ = 0.3 \cdot \frac{\sum_{i=1}^{10} g(x_i)}{10} \tag{4.1}$$

donde $g(x_i)$ es la función de los valores del flujo ordenados de mayor a menor. Es decir, se toma el $30\,\%$ de la media de los 10 mayores valores del flujo para establecer el umbral positivo. Para el umbral negativo es análogo, pero se toma la media de los valores mínimos, es decir,

$$\Gamma_t^- = 0.3 \cdot \frac{\sum_{i=1}^{10} h(x_i)}{10} \tag{4.2}$$

donde $h(x_i)$ representa los valores de flujo ordenados de menor a mayor.

La elección del umbral del $30\,\%$ o $25\,\%$, etc, es arbitraria. Esencialmente debe elegirse un umbral suficientemente alto como para filtrar los eventos más grandes (avalanchas), pero no tan grande como para que se pierda estadística. El umbral óptimo es un compromiso entre eliminar el ruido (transporte difusivo) pero mantener una estadística suficiente del transporte cooperativo (avalanchas).

Se concluye por todo lo expuesto que el tratamiento del $30\,\%$ es de suma importancia a la hora de trabajar con señales temporales que tienen la forma de la Figura 4.1 y es que entre cada par de avalanchas como ya se ha explicado hay un tiempo característico denominado tiempo de espera en el que el único transporte es el difusivo que para un radio fijo se caracteriza por fluctuaciones aleatorias. Es por ello que se necesita un método que elimine este entorno central asociado a las fluctuaciones gaussianas para así poder analizar el transporte mediante avalanchas, el cual es el responsable de mostrar las correlaciones y así poder saber cuál de las señales de los 10 radios es más correlacionada y cuál menos.

Es importante resaltar que no se está afirmando que el tratamiento del $30\,\%$ sea la única manera de conseguir esta eliminación de las fluctuaciones gaussianas del entorno central, sin embargo, es una que se ha comprobado que funciona. En el Apéndice B se proporciona también un ejemplo de método de tratamiento de datos comúnmente usado en la literatura que no es efectivo para eliminar estas fluctuaciones aleatorias.

A continuación, se proporciona el pseudoalgoritmo para llevar a cabo el tratamiento del umbral del $30\,\%$ [4].

1.-Sea $F(t)$ el flujo de la señal temporal que se está analizando, entonces se calculan los umbrales positivo Γ^+ y Γ^- según la Ecuación 4.1 y la Ecuación 4.2 respectivamente.

2.-Se redefine la señal del flujo en la forma

$$\begin{cases} \text{si } \Gamma^- \leq F(t_0) \leq \Gamma^+, \text{ entonces } F^*(t_0) = 0 \\ \text{si } F(t_0) > \Gamma^+ \text{ entonces } F^*(t_0) = F(t_0) - \Gamma^+ \\ \text{si } F(t_0) < \Gamma^- \text{entonces } F^*(t_0) = F(t_0) - \Gamma^- \end{cases}$$

donde $F^*(t)$ representa la señal del flujo después del tratamiento del $30\,\%$

Resumiendo, se calculan los umbrales positivo y negativo con el $30\,\%$ de la media de los picos positivos y los 10 picos negativos y luego se resta a la señal temporal original positiva el flujo positivo y a la señal original negativa el flujo negativo, sin permitir un cambio de signo.

Cabe destacar que hay autores [4] que este segundo paso lo hacen de una manera ligeramente diferente.

2*.-Se redefine la señal del flujo en la forma

$$\begin{cases} \text{si } \Gamma^- \leq F(t_0) \leq \Gamma^+, \text{ entonces } F^*(t_0) = 0 \\ \text{si } F(t_0) > \Gamma^+ \text{ entonces } F^*(t_0) = F(t_0) \\ \text{si } F(t_0) < \Gamma^- \text{entonces } F^*(t_0) = F(t_0) \end{cases}$$

es decir, el flujo por encima del umbral no lo modifican. No obstante, ambas formas son equivalentes. A continuación, se muestra gráficamente la señal resultante después de aplicar el tratamiento del $30\,\%$ a la señal original.

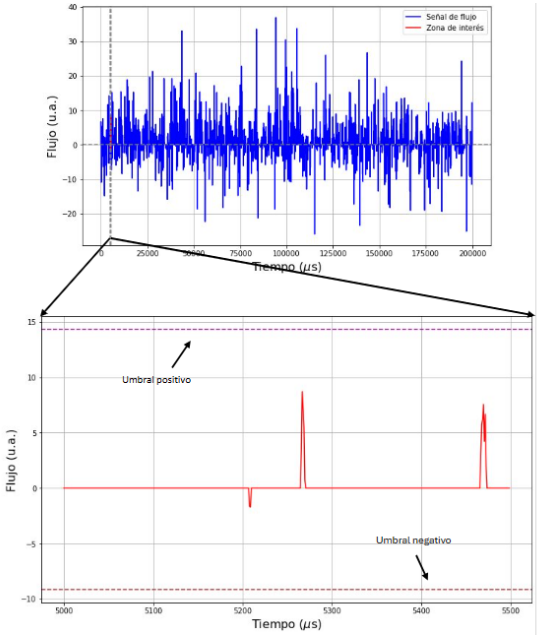


Figura 4.4: Representación de la señal temporal de flujo del SLPM correspondiente a $r = 1.0$ cm tratada a través del tratamiento del $30\,\%$.

La figura es exactamente igual a la Figura 4.1 pero con la única diferencia de que en esta se le ha aplicado el tratamiento del $30\,\%$. La diferencia entre ambas figuras es notable ya que mientras que en la Figura 4.1 se puede observar cómo el entorno central está cubierto de fluctuaciones, aquí se puede ver cómo la parte central está vacía, lo cual es todavía más apreciable al hacer zoom. De hecho, al hacer zoom se puede observar cómo solo hay 3 zonas de flujo no nulo, que son precisamente las 3 avalanchas mencionadas en la Figura 4.1. Nótese que las líneas punteadas representando el umbral positivo y el negativo son las mismas.

Observando esta figura y basados en la teoría del Capítulo 2 todo parece indicar que se ha resuelto el problema, gracias a que se ha eliminado el entorno central de las fluctuaciones aleatorias.

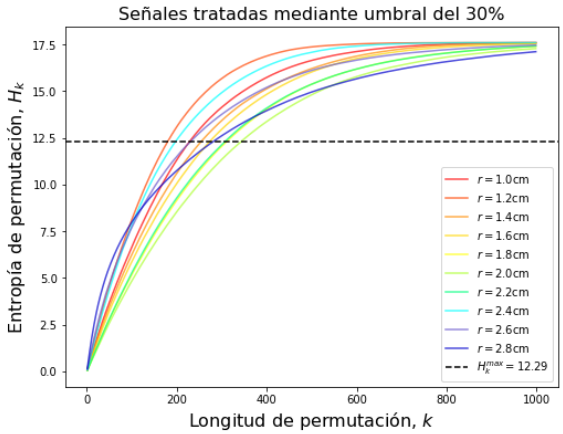


Figura 4.5: Representación de la entropía de permutación correspondiente a las señales tratadas mediante el tratamiento del $30\,\%$ para señales del SLPM en función de la longitud de permutación. La línea discontinua negra representa el punto a partir del cuál los valores están afectados por el problema de tamaño finito.

Lo primero que resalta es que se ha eliminado completamente el problema de las fluctuaciones aleatorias ya que como se puede observar el problema de tamaño finito empieza en índices entre 150 y 300 aproximadamente (donde corta con $H_k^{max} = 12.29$). Volviendo al Capítulo 3 recuérdese que para la aplicación logística estos valores estaban entre 10 y 32 y cuanto mayor era este índice, mayor tendía a ser la complejidad asociada. Por tanto, se espera que las complejidades asociadas a estas señales tratadas sean notablemente mayores que las del logístico. Antes de calcularlas, intentemos razonar con herramientas puramente teóricas del Capítulo 2 por qué sucede esto.

Para empezar se debe tener en cuenta cómo son las señales que se están analizando, para lo cual vamos a la Figura 4.4 y observamos que el mayor porcentaje de las señales son ceros, es decir, igualdades.

Ahora, debemos ir al Apéndice C y observar que la técnica computacional usada trata las igualdades como cadenas ascendentes. Esto significa que para calcular la entropía según la Definición 2.4.1 se obtiene un valor menor del que se debería haber obtenido, ya que el tener un gran porcentaje de un mismo tipo de permutación como son las cadenas ascendentes hace que según la Definición 2.4.1 se obtengan valores más pequeños de esta. Esto a su vez hace que el problema de tamaño finito de la Sección 2.3 aparezca para índices mayores ya que al ser para cada longitud de permutación k la entropía menor de la que debiera, entonces tarda más en alcanzar el valor dado H_k^{max} en la Tabla 2.2. Es por esto que se obtienen estos índices de 150 a 300, que comparando con el caso del logístico son tan elevados.

A su vez, este hecho es positivo ya que asegura que el crecimiento de la entropía es puramente debido al mayor o menor número de correlaciones en el transporte por avalanchas y no hay ninguna contribución de la aleatoriedad del transporte difusivo, por lo que es posible identificar qué señales están más correlacionadas o menos, ya que cómo se ha mencionado anteriormente las correlaciones de una señal de flujo de un plasma aparecen en el transporte por avalanchas y no en el transporte por difusión.

Volviendo a la expresión obtenida anteriormente para la complejidad:

$$\eta = -\sum_{i=2}^{n_{max}} (m-2)\Delta^2 H_m, \tag{4.3}$$

las curvas de mayor curvatura en el rango $H_k \in [0, 12.29]$ serán las de mayor complejidad. En la figura 4.5 se ve claramente cómo las dos curvas asociadas a los radios más externos $r = 2.6$ cm

y $r = 2.8$ cm son las que mayor curvatura tienen, ya que por ejemplo la de $r = 2.8$ cm empieza tomando valores H_k mayores que las restantes 9 curvas para un k fijo pequeño, pero para k grandes las restantes curvas toman valores mayores que ella. Sucede lo mismo pero en menor medida para la curva correspondiente a $r = 2.6$ cm. Luego, se puede intuir que cómo estas dos curvas tienen la mayor curvatura negativa, entonces tendrán la mayor complejidad.

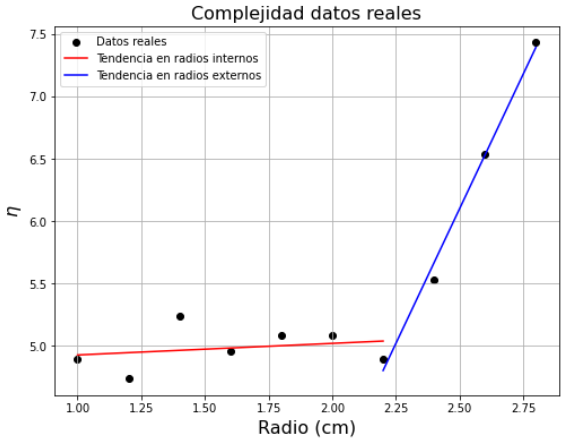


Figura 4.6: Representación de la complejidad para cada una de las 10 posiciones radiales del SLPM entre $r = 1.0$ cm y $r = 2.8$ cm, tratadas mediante el tratamiento del umbral del $30\,\%$.

Se concluye que en los radios internos, desde $r = 1.0$ cm hasta $r = 2.2$ cm, la tendencia es constante, mientras que en los radios externos la tendencia es ascendente.

En [4] se hace un estudio del mismo problema, pero mediante un análisis alternativo. Lo que concluyen es que en los 2 radios más externos, es decir, en $r = 2.6$ cm y $r = 2.8$ cm es donde hay una apreciable mayor complejidad, mientras que para los 7 radios más internos la complejidad es aproximadamente constante, que es la misma conclusión obtenida en la Figura 4.6. Por lo tanto, como los resultados son análogos a los obtenidos en [4] y además estos tiene un respaldo teórico sólido, entonces se concluye que el método funciona adecuadamente para estos datos reales también.

Cabe destacar que en el Apéndice A se han rehecho los resultados de [4] usando el análisis alternativo usado ahí.

Capítulo 5

Conclusiones

En este trabajo se ha desarrollado un método basado en la entropía de permutación de la teoría de la información que permite el cálculo del grado de correlación o memoria para distintas series temporales de una manera sistemática. Se ha deducido su validez teórica en el Capítulo 2 mediante una estructura matemático-deductiva, que es original de este TFG y se ha comprobado su eficiencia práctica con datos simulados por la aplicación logística y con datos reales provenientes de medidas de flujo radial turbulento en los plasmas generados en la SLPM.

El uso de sucesiones de la entropía de bloque como forma de detección de correlaciones es una técnica que ha sido estudiada desde el punto de vista teórico [6] y práctico [1] en la literatura. No obstante, debido al gran número de datos que necesita esta técnica para obtener una estadística fiable y debido a la finitud de las señales reales, la entropía de bloque no sirve para detectar correlaciones a medio y largo plazo como se puede ver en [1]. Sin embargo, la denominada entropía de permutación introducida por primera vez en [8] requiere un número considerablemente menor de datos para obtener una estadística fiable, lo que ahora permite detectar correlaciones en señales temporales a medio y largo plazo.

El principal problema de esta entropía de permutación es que a día de hoy no se dispone de una base teórica necesaria para respaldar los resultados numéricos relativos a las correlaciones y la complejidad de series temporales finitas. En este trabajo, en la Sección 2.5 se ha desarrollado este marco teórico, a partir del cual se describe un método que permite detectar todo tipo de correlaciones en señales temporales.

Se ha comprobado que el método desarrollado teóricamente detecta correlaciones, cumpliendo con el objetivo principal de este trabajo.

En primer lugar, se han usado datos simulados procedentes de la aplicación logística, muy usada en el ámbito de los sistemas dinámicos. Se han medido las correlaciones de señales temporales de las cuales previamente se conocía cuál era su relación de complejidades. Los resultados muestran cómo algunas señales temporales son claramente más correlacionadas que otras, con una tendencia descendente general respecto al parámetro b, como se esperaba obtener. Además, el método detecta las ventanas periódicas, en las que la complejidad disminuye y luego aumenta repentinamente en su límite.

Por otro lado, se ha empleado el método para detectar correlaciones en señales temporales asociadas a los flujos en distintas localizaciones de la SLPM, para la cual se conoce [4] que las señales asociadas a los radios más internos son menos correlacionadas que las de los radios externos, especialmente que los dos más externos. Justamente, al medir estas correlaciones con nuestro método se ha obtenido a través de la Figura 4.6 cómo la complejidad o grado de correlación aumenta

súbitamente a partir de $r = 2.2$ cm, tal como se esperaba.

Por tanto, se ha probado que el método desarrollado cuenta con un respaldo teórico sólido desarrollado en este TFG, y un respaldo numérico/experimental. Se ha comprobado que detecta todo tipo de correlaciones, de bajo, medio y largo plazo. Por lo tanto, es un método ideal si se desea caracterizar el grado de correlación general de una señal temporal, así como su complejidad.

Una posible línea de investigación a futuro sería medir señales de flujo en múltiples posiciones radiales por disparo, lo que permitiría medir propiedades como la velocidad de propagación de las avalanchas. El método desarrollado en este TFG puede ser de gran utilidad como técnica de análisis en estos experimentos.

Bibliografía

[1] L. Cavada de la Riva. Teoría de la información aplicada al análisis de la dinámica de plasmas de fusión. *TFG Univ.Cantabria*, (2023). URL https://hdl.handle.net/10902/30351.

[2] E. R. Sadik-Zada, A. Gatto, and Y. Weißnicht. Back to the future: Revisiting the perspectives on nuclear fusion and juxtaposition to existing energy sources. *Energy*, **290**, 129150, (2024). ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2023.129150. URL https://www.sciencedirect.com/science/article/pii/S0360544223025446.

[3] O. Castellanos. Estudio de la turbulencia en un plasma linealmente magnetizado. *PhD thesis, Univ.Cantabria*, (2007).

[4] J.A. Mier, J. Blanco, R. Sanchez, O. Castellanos, D.E. Newman, E Anabitarte, and J.M. López. Avalanche statistics of fuctuation-induced fluxes from the slpm and the w7-as stellarator. *Plasma Phys.Control.Fusion*, **66**, 065015, (2024). doi: 10.13140/RG.2.2.23859.20007.

[5] R. O. Dendy. Information Theory and Plasma Turbulence. *AIP Conference Proceedings*, **1188**, 247-257, (2009). ISSN 0094-243X. doi: 10.1063/1.3266803. URL https://doi.org/10.1063/1.3266803.

[6] K. Lindgren. *Information Theory for Complex Systems*. Springer Berlin, Heidelberg, Germany, (2024).

[7] J. Goings. Maximum entropy distributions, (2021). URL https://joshuagoings.com/assets/MaximumEntropyDistributions.pdf.

[8] C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Phys.Rev.Lett*, **88**, 174102, (2002). doi: 10.1103/PhysRevLett.88.174102.

[9] A. Quarteroni, R. Sacco, and F. Saleri. Numerical mathematics. **37**, (2007). doi: 10.1007/b98885.

[10] J. M. Amigó, S. Zambrano, and M. A. F. Sanjuán. Combinatorial detection of determinism in noisy time series. *Europhysics Letters*, **83**, 60005, (2008). doi: 10.1209/0295-5075/83/60005. URL https://dx.doi.org/10.1209/0295-5075/83/60005.

[11] E. W. Weisstein. Logistic map. *MathWorld–A Wolfram Web Resource*, (2024). URL https://mathworld.wolfram.com/LogisticMap.html.

[12] S. Wolfram. *A New Kind of Science*. Wolfram Media, Champaign, IL, (2002).

[13] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, **261**, 459-467, (1976). doi: 10.1038/261459a0. URL https://doi.org/10.1038/261459a0.

[14] A. Grami. Chapter 4 - probability, random variables, and random processes. In Ali Grami, editor, *Introduction to Digital Communications*, pages 151–216. Academic Press, Boston, (2016). ISBN 978-0-12-407682-2. doi: https://doi.org/10.1016/B978-0-12-407682-2.00004-1. URL https://www.sciencedirect.com/science/article/pii/B9780124076822000041.

[15] A. B. Pessa and H. V. Ribeiro. ordpy: A python package for data analysis with permutation entropy and ordinal network methods. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **31**, (2021). ISSN 1089-7682. doi: 10.1063/5.0049901. URL http://dx.doi.org/10.1063/5.0049901.

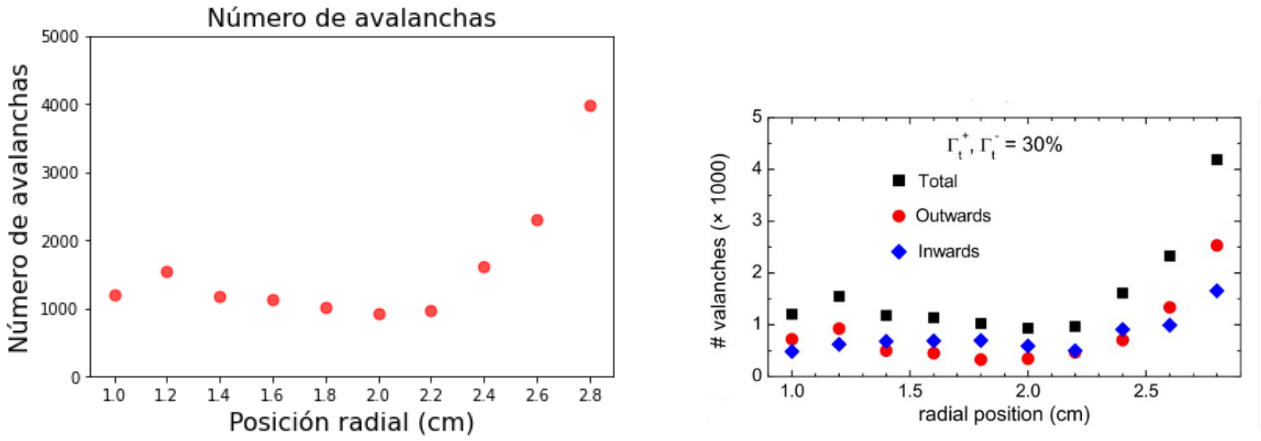
Apéndice

A. Método Alternativo de Detección de Correlaciones

Lo primero de todo recordamos que la función de este apéndice es únicamente la de eliminar toda posible duda sobre cualquier error en el procedimiento que llevan a los resultados dados en la Figura 4.6 al obtener los mismos resultados mediante otro método dado en la literatura [4].

Como ya se ha comentado anteriormente las correlaciones aparecen a través del transporte por avalanchas. En [4] se analizan estas correlaciones mediante el análisis de los parámetros característicos de este transporte por avalanchas, que como se ha comentado antes son los tiempos de espera entre cada par de avalanchas, la duración y área de cada una de las avalanchas y obviamente el número total de avalanchas que aparecen en la señal.

En primer lugar, se va a analizar que el número de avalanchas obtenido en cada una de las 10 señales es coherente con lo obtenido en [4].



(a) (b)

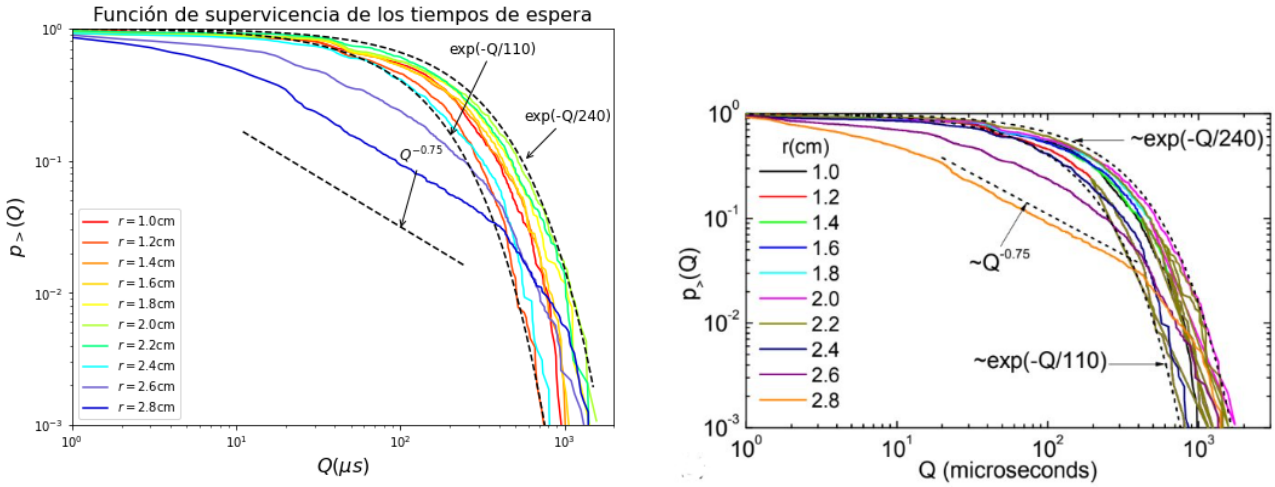
Figura 1: Representación del número de avalanchas en función del radio en la subfigura (a) y representación análoga pero tomada de [4] en la subfigura (b). Cabe destacar que en [4] se clasifican las avalanchas como salientes (*outwards*) o entrantes (*inwards*), por lo que para la comparación se debe mirar el número total dibujado en negro.

Dos cosas son destacables de esta figura. La primera y principal es que ambas figuras muestran el mismo resultado y por otro que, se ve cómo para radios más externos el transporte por avalanchas cobra mayor importancia, lo que en [4] se asocia con una mayor complejidad. Así, este resultado es coherente con lo obtenido en la Figura 4.6.

Una vez hecho esto nos quedaría comprobar la estadística asociada a los tiempos de espera,

duraciones y áreas de las avalanchas, lo cual en [4] se hace a través de las funciones de supervivencia empírica S(t), las cuales se definen como $S(t) = 1 - F(t)$ donde $F(t)$ es la función de distribución empírica.

Como nuestro objetivo es solo comprobar que el procedimiento es correcto, nos conformamos con rehacer una de las funciones de supervivencia, como por ejemplo la de los tiempos de espera.



(a) (b)

Figura 2: Representación de la función de supervivencia obtenida para los tiempos de espera en la subfigura (a) y función de supervivencia tomada en [4] en la subfigura (b).

Es inmediato comprobar que ambas figuras son iguales con la única diferencia de que la línea punteada señalada como $Q^{-0.75}$. No obstante, esta línea solo representa que la curva azul correspondiente a $r = 2.8$ cm tiene una región la cual en escala logarítmica se ajusta a una recta y la diferencia reside en que en [4] se traslada esta línea para ponerla sobre la curva correspondiente a $r = 2.8$ cm.

La interpretación física de estos resultados es que aquellas curvas a las que se les puede hacer un ajuste lineal en escala logarítmica a alguna de sus regiones son las mas correlacionadas [4]. Así, concluimos de nuevo que los 2 radios más externos están claramente más correlacionados, de nuevo, verificando nuestros resultados de la Figura 4.6.

B. Tratamiento de Suavización

Como se ha comentado en la Sección 4.2, el tratamiento del umbral del $30\,\%$ no es el único posible. Otro tratamiento de datos que se suele hacer comúnmente en la literatura es el tratamiento por suavización. No obstante, como se ha anticipado anteriormente en la Sección 4.2 este tratamiento no sirve para arreglar nuestro problema ilustrado en la Figura 4.4 de fluctuaciones aleatorias dominando el entorno central. Así, lo que queremos ilustrar es que nuestro problema no se puede arreglar con cualquier tratamiento común en la literatura, sino que se debe identificar la raíz del problema y arreglarlo con un tratamiento que se adecua a él como es el tratamiento del $30\,\%$.

Lo primero de todo explicamos en qué consiste el método de suavización. Sea $F(t_i)$ la señal de flujo original y $F^*(t_i)$ la correspondiente señal de flujo suavizada donde $i \in \{1, ...N\}$ siendo N la

longitud de la señal original. Entonces, se verifica que

$$F^*(t_i) = \frac{\sum_{j=i-m}^{j=i+m} F(t_j)}{2m+1} \tag{1}$$

donde m es el grado de suavización que se desea obtener. Lo usual es tomar $m \in \{2, 4, 8, 16\}$, ya que para m mayores la suavización es de tal magnitud que borra las propiedades de la señal original. De hecho, $m = 8$ y $m = 16$ ya se suelen considerar como suavizaciones demasiado elevadas. En este trabajo se toma $m = 4$.

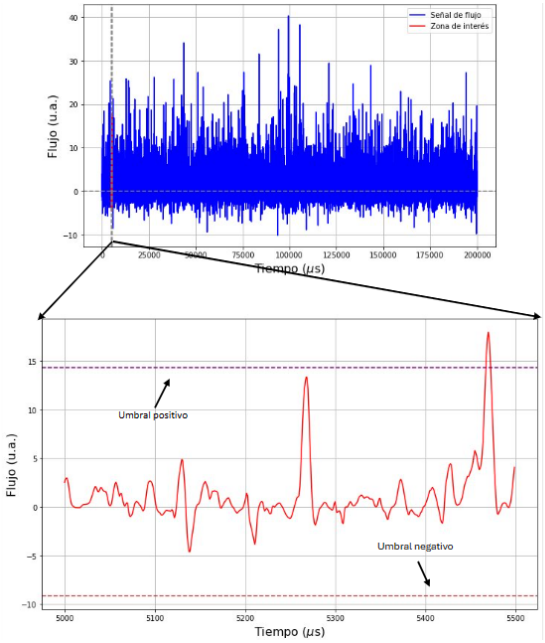


Figura 3: Representación de la señal temporal de flujo del SLPM correspondiente a $r = 1.0$ cm, después de haberle sido aplicado el tratamiento de la suavización con $m = 4$.

La figura es exactamente igual a la Figura 4.1 y a la Figura 4.4 pero con la única diferencia de que en esta se le ha aplicado el tratamiento de la suavización. Vamos a comparar esta figura con la Figura 4.1 de la señal original. Al hacerlo lo que vemos es que no se ha eliminado la parte central correspondientes a las fluctuaciones aleatorias y que además se ha reducido el tamaño de las avalanchas, es decir, se han suavizado. Es decir, no solo no hemos arreglado el problema de no eliminar las fluctuaciones aleatorias, sino que además hemos añadido otro problema que es que hemos reducido el tamaño y hemos alterado la forma de las avalanchas, así dificultandonos el trabajo de detección de las correlaciones, ya que como hemos dicho antes estas correlaciones aparecen en las avalanchas. Luego, ya podemos anticipar que los resultados cuantitativos que obtengamos mediante este tratamiento van a dar malos resultados.

Aunque ya se haya anticipado que va a dar malos resultados, veamos qué es lo que obtenemos al representar la entropía de permutación en función de la longitud de permutación.

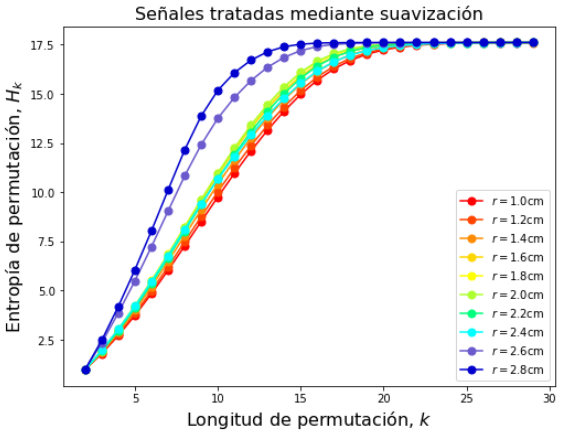


Figura 4: Representación de la entropía de permutación correspondiente a las señales tratadas mediante el tratamiento de suavización con $m = 4$ para los 10 radios del SLPM mencionados anteriormente en función de la longitud de permutación.

La figura obtenida aquí es muy similar a la Figura 4.3, por lo que por argumentos análogos a los dados para esa figura podemos concluir que el análisis realizado no tiene ningún sentido teórico ni práctico, por lo que se concluye que para analizar las señales de este capítulo no sirve el método introducido en la Sección 2.6 previo tratamiento de suavización.

C. Técnicas Computacionales

Dividimos este apéndice en dos partes. En la primera, Apéndice C.1, describimos la librería ordpy de Python que hemos usado y cómo la hemos usado, es decir, qué partes de ella hemos usado. En la segunda, especificamos exactamente cómo hemos calculado la complejidad en nuestro programa de Python, es decir, cómo se han obtenido las Figuras 3.5 y 4.6, ya que los resultados del texto se sintetizan en estas dos figuras, por si el lector deseara comprobar que los resultados efectivamente son compatibles con los esperados.

C.1. Consideraciones generales: Librería Ordpy

Debido a que el concepto de la entropía de permutación no es un concepto nuevo y ha tenido una repercusión notable, se han creado librerías que la aplican en lenguajes de programación como Python [15].

Como se expone en [15] esta librería tiene muchas funciones, no obstante, para este trabajo nos restringimos a trabajar únicamente con la función *permutation_entropy*(*data*, d_x, $d_y = 1$, $\tau_x = 1$, $\tau_y = 1$, *base*, *normalized*, *probs*, *tie_precision*). A continuación detallamos a qué rangos de las variables de entrada nos restringimos.

En primer lugar, nos restringimos al caso de series temporales unidimensionales, es decir, variamos d_x, la longitud de los bloques a permutar en horizontal, mientras que fijamos $d_y = 1$, la longitud de los bloques a permutar en vertical. La situación $d_y \neq 1$ es para el caso en que se quiera analizar por ejemplo imágenes.

Por otro lado, fijamos también $\tau_x = 1$ que es la distancia entre las diferentes posiciones de los símbolos a permutar en horizontal. Por ejemplo, si $\tau_x = 1$ y tenemos cierta señal temporal $\{x_i\}_{i=1}^N$, entonces los bloques de permutación de longitud n serán de la forma $x_i x_{i+1} ... x_{i+n-1} x_{i+n}$, pero en cambio para

$\tau_x = 2$ los bloques de permutación de longitud n serán de la forma $x_i x_{i+2} ... x_{i+n-2} x_{i+n}$ y para el caso general de $\tau_x = k \leq n$ tendríamos bloques de permutación de longitud $x_i x_{i+k} ... x_{i+n-k} x_{i+n}$.

Obviamente, también escogemos $\tau_y = 1$, ya que es la única posible si $d_y = 1$.

Por último, la entrada *base* que denota la base del logaritmo con la que trabajamos es 2 como ya se ha comentado en capítulos previos. Además, elegimos no tener resultados normalizados donde normalizar significa que los valores de la entropía de permutación de longitud k se dividan por el valor máximo de la entropía de permutación entre todas las posibles longitudes k.

Probs y *tie_precision* también los configuramos en *false*, ya que *probs* en false significa que metemos como entrada en *data* las señales temporales tal cual, ya que *probs=true* es para aquellos casos donde en *data* se introducen ya las frecuencias de cada posible permutación de una cierta longitud de permutación k y *tie_precision* solo es una herramienta para redondear los resultados obtenidos.

Otra cosa relevante para el trabajo que debe de ser comentada es cómo trabaja esta librería con las igualdades, ya que como se explicará más adelante en el texto en algunos análisis las igualdades son un factor importante. Como se puede leer en [15], en la librería *ordpy* las igualdades se consideran como una cadena ascendente de símbolos, es decir, si $x_{t_i} = x_{t_j}$ pero $t_i \neq t_j$, entonces por convención $x_{t_i} < x_{t_j}$ si $t_i < t_j$ y $x_{t_i} > x_{t_j}$ si $t_i > t_j$. Originalmente en el artículo original donde se introduce el concepto de entropía de la permutación [8], se propone resolver las igualdades introduciendo ligeras perturbaciones aleatorias, no obstante, lo más común en la literatura es considerar las igualdades como cadenas ascendentes [15].

Obviamente el considerarlas como cadenas ascendentes se traduce en una reducción de la entropía de permutación para todas las longitudes de permutación k. Esto es debido a que si cada vez que hay una cadena de igualdades lo trasladamos a que haya una cadena ascendente, entonces estamos haciendo que la frecuencia del tipo de permutación ascendente, es decir, la (12) para $k = 2$, (123) para $k = 3$, (1234) para $k = 4$ y así sucesivamente y así si la mayoría son igualdades lo traduciremos a que la mayoría son cadenas ascendentes y así tendremos un tipo de permutación muy dominante para cada longitud de permutación k y así de las Definiciones 2.1.1, 2.1.2 y 2.4.1 es inmediato deducir que obtendríamos un valor muy pequeño de la entropía de permutación. En cambio si introdujéramos perturbaciones aleatorias pequeñas, su efecto se traduciría en aumentar la entropía de permutación debido a que tendería a homogeneizar las frecuencias de las permutaciones que es justo el efecto contrario que teníamos antes.

Desde un punto de vista puramente teórico es claramente más adecuado considerar una señal constante como periódica que como aleatoria y por otro lado, debido al problema de tamaño finito de la sección 2.3 es mucho más beneficioso reducir la entropía que aumentarla, ya recordemos que estamos trabajando con señales temporales de longitud finita.

Por último, volvamos por un momento a la Observación 2.5.1 que señala cómo calcular las diferencias de orden 2, $\Delta^2 H_k$ con el menor error posible. Hay que destacar que computacionalmente no siempre se pueden calcular en la forma

$$\Delta^2 H_k = \frac{H_{k+2} - 2H_k + H_{k-2}}{4} \tag{2}$$

ya que por ejemplo para $k = 2$, H_0 no está definida. Para arreglar esto se ha tomado el criterio de calcular la diferencia de orden 1 en la forma

$$\Delta H_k = H_{k+1} - H_k \tag{3}$$

en vez de

$$\Delta H_k = H_{k+1} - H_{k-1} \tag{4}$$

para el caso en el que H_{k-1} no exista. De aquí se deduce cómo debe ser $\Delta^2 H_k$.

C.2. Cálculo de la Complejidad

En este apéndice detallamos cómo se han obtenido las Figuras 3.5 y 4.6 computacionalmente paso por paso, ya que son la síntesis de los resultados del texto, por si el lector deseara comprobar que los resultados efectivamente son los que son.

Vamos a hacerlo únicamente para el caso de los datos reales, ya que en este caso a parte de la complejidad se debe hacer un tratamiento del $30\,\%$ previo, pero el cálculo de la complejidad en sí es análogo en el caso del logístico.

Primero empezamos con el código que permite realizar el tratamiento del $30\,\%$ a nuestras señales temporales. Antes de nada cabe destacar que vamos a llevarlo a cabo solo para la señal temporal proveniente del radio $r = 1$ cm, al cual se le ha llamado como *data_plasma_1058*. Para el resto de radios el código es análogo.

```
#1.- Importamos los datos provenientes de la SLPM.
data_plasma_1058 = np.loadtxt('C:/Users/Iker León/OneDrive - UNICAN - Estudiantes
    /Escritorio/TFGfísica/data/DatosPlasmaOriginal/flujo1058.txt')
#2.-Hacemos un cambio de unidades para trabajar más fácil, ya que las unidades no
     nos afectan
data_plasma_1058 = [x / 1e19 for x in data_plasma_1058]
#3.- Obtenemos los 10 valores más grandes.
diez_valores_mayores = sorted(data_plasma_1058, reverse=True)[0:10]
#4.- Obtenemos los 10 valores más pequenos.
diez_valores_menores = sorted(data_plasma_1058, reverse=False)[0:10]
#5.-Ahora, calculamos el umbral positivo y el negativo.
upper_threshold=(sum(diez_valores_mayores)/10)*0.3
lower_threshold=(sum(diez_valores_menores)/10)*0.3
#6.-Calculamos la longitud para meterlo al for.
longitud=len(data_plasma_1058)
#7.-Hacemos el for para hacer el tratamiento del 30% a cada índice.
data_plasma_1058_threshold=np.zeros(longitud)
for i in range(1, longitud):
    if data_plasma_1058[i]<upper_threshold and data_plasma_1058[i]>lower_
       threshold:
        data_plasma_1058_threshold[i]=0
    elif data_plasma_1058[i]>0 and data_plasma_1058[i]>upper_threshold:
        data_plasma_1058_threshold[i]=data_plasma_1058[i]-upper_threshold
    elif data_plasma_1058[i]<0 and data_plasma_1058[i]-lower_threshold:
        data_plasma_1058_threshold[i]=data_plasma_1058[i]-lower_threshold
#8.-Guardamos el vector nuevo que cuenta ya con el tratamiento del 30%.
np.save("C:/Users/Iker León/OneDrive - UNICAN - Estudiantes/Escritorio/TFGfísica/
    data/DatosPlasmaAnálisis/Metodo30porciento/data_plasma_1058_threshold.npy",
    data_plasma_1058_threshold)
```

Lo único que hemos hecho en las líneas previas es seguir el pseudoalgoritmo dado en la Sección 4.2.

A continuación, proporcionamos el código requerido para calcular la entropía de estas señales temporales tratadas con el tratamiento del $30\,\%$. A partir de aquí el procedimiento para el caso de las señales temporales provenientes del logístico y de los datos del SLPM son análogas.

```
#1.-Importamos la senal del método del 30%.
data_plasma_1058_metodo30 = np.load("C:/Users/Iker León/OneDrive - UNICAN -
    Estudiantes/Escritorio/TFGfísica/data/DatosPlasmaAnálisis/Metodo30porciento/
    data_plasma_1058_threshold.npy")
```

```
#2.-Calculamos la entropía. Para el caso de las senales del SLPM llegamos hasta longitudes de 1000.
entropy_data_plasma_1058_metodo30 = []
for dx_ in range(2, 1000):
    entropy_data_plasma_1058_metodo30 += [ordpy.permutation_entropy(data_plasma_1058_metodo30, dx_, base=2, normalized=False)]
    np.save("C:/Users/Iker León/OneDrive - UNICAN - Estudiantes/Escritorio/TFGfísica/data/DatosPlasmaAnálisis/Metodo30porciento/Entropías1000/entropy_data_plasma_1058_metodo30.npy", entropy_data_plasma_1058_metodo30)
```

En las líneas previas hemos usado la función *permutation_entropy* de la librería ordpy según hemos explicado en el Apéndice C.1 anterior. Cabe destacar que el *for* hecho en el rango de 2 a 1000, en realidad solo considera las longitudes de 2 a 999 por cómo está configurado Python

Por último, calculamos la complejidad a partir de este vector de la entropía de permutación concerniente a la señal del radio $r = 1$ cm a la que se ha tratado con el método del $30\,\%$.

```
#1.-Importamos el vector de entropías.
entropy_data_plasma_1058_metodo30_N200000 = np.load("C:/Users/Iker León/OneDrive - UNICAN - Estudiantes/Escritorio/TFGfísica/data/DatosPlasmaAnálisis/Metodo30porciento/Entropías1000/entropy_data_plasma_1058_metodo30.npy")
#2.-Calculamos el primer índice que ya está afectado por el problema de tamano finito.
indice_1058=np.argmax(entropy_data_plasma_1058_metodo30_N200000>= 12.29)
#3.-Calculamos la longitud total del vector de entropías.
longitud_1058=len(entropy_data_plasma_1058_metodo30_N200000)
#4.-Definimos cuál es nuestro vector de entropías.
entropy=entropy_data_plasma_1058_metodo30_N200000[0:longitud_1058+0]
#5.-Calculamos las diferencias finitas primeras y segundas.
dy_dx = np.gradient(entropy, np.linspace(2, longitud_1058+1, num=longitud_1058+0))
dy2_dx2 = np.gradient(dy_dx[0:longitud_1058], np.linspace(2, longitud_1058+1, num=longitud_1058+0))
#6.-Insertamos dos ceros al principio para hacerlo más intuitivo, ya que la entropía
#de permutación empieza en 2 y no en 0.
dy2_dx2 = np.insert(dy2_dx2, 0, [0, 0])
#A partir de ahora habrá que tener cuidado ya que hemos hecho una translación de dos índices.
#7.-Calculamos la complejidad, variando del índice óptimo 5 a la izquierda y 5 a la derecha.
complejidad_1058=np.zeros(11)
for j in range(0,11):
    #El índice óptimo es indice_1058, no índice_1058+2 que podríamos pensar ya que
    #hemos hecho la translación de 2. No hay que hacer +2, ya que recordemos
    #que la diferencia finita va 2 índices hacia arriba y 2 hacia abajo. Luego,
    #el correspondiente a indice_1058 es el primero que se ve ligeramente afectado.
    for i in range(2,indice_1058-5+j):
        complejidad_1058[j] +=(i-2)*(-dy2_dx2[i])
```

Vamos a analizar estas últimas líneas con atención, ya que nuestros resultados se sintetizan en las Figuras 3.5 y 4.6 y estas líneas son las que han creado estas figuras.

Lo primero de todo nos fijamos en que *índice_1058* representa el primer índice en el que ya tenemos problema de tamaño finito. No obstante, también nos debemos fijar en que al calcular las diferencias finitas segundas hemos puesto dos ceros al principio, para hacerlo más intuitivo, ya que la entropía de permutación empieza en 2 y no en 0. Uno podría preguntarse por qué no hemos metido los dos ceros al principio de todo, es decir, antes de calcular las diferencias finitas. La respuesta a ello

es que si lo hubiéramos hecho, estos ceros hubieran afectado al cálculo de las diferencias finitas.

Entonces, al calcular la complejidad debemos tener en cuenta esta translación y por ello la complejidad la deberíamos calcular como

```
for i in range(2,indice_1058+2):
        complejidad_1058 +=(i-2)*(-dy2_dx2[i])
```

Y si lo hiciéramos así nos estaríamos confundiendo, ya que para empezar, cuando hacemos *for i in range(a,b)* el *b* no lo considera, es decir, trabaja hasta el *b-1*, por lo que en principio debería ser

```
for i in range(2,indice_1058+3):
        complejidad_1058 +=(i-2)*(-dy2_dx2[i])
```

Pero esto sería otro error, ya que recordemos cómo definíamos la diferencia finita segunda. Nosotros definíamos la diferencia finita segunda como

$$\Delta^2 H_k = \frac{H_{k+2} - 2H_k + H_{k-2}}{4} \tag{5}$$

es decir, en nuestra curvatura $\Delta^2 H_k$ intervienen tanto H_{k+2} como H_{k-2}, es decir, dos índices a la derecha también nos van a influir. Luego, uno pensaría en hacer

```
for i in range(2,indice_1058+1):
        complejidad_1058 +=(i-2)*(-dy2_dx2[i])
```

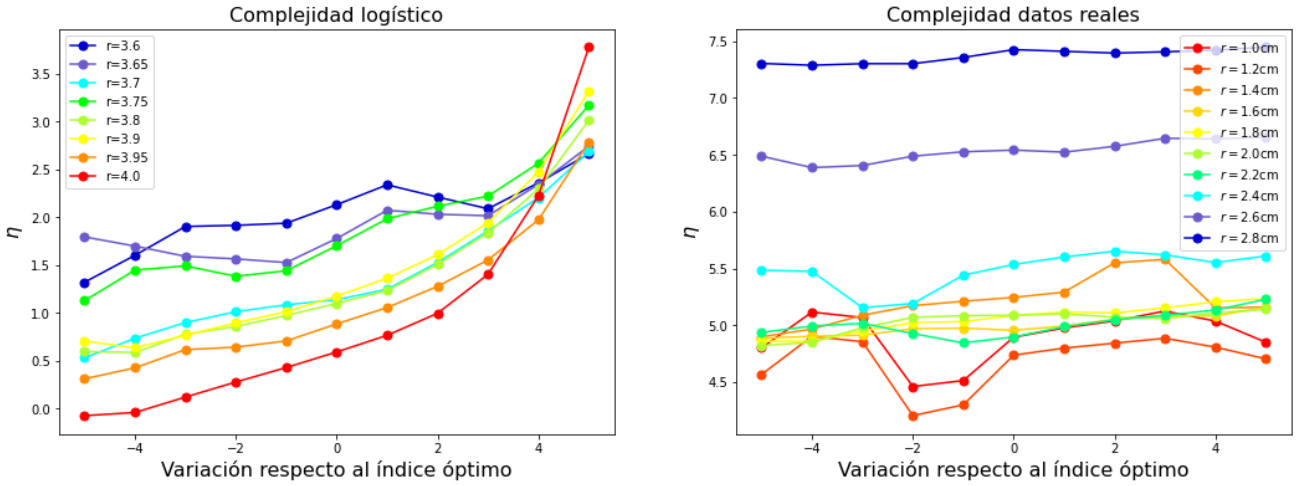
Y nos volveríamos a equivocar. El por qué de nuestra nueva equivocación reside en cómo definíamos el índice_1058:

```
indice_1058=np.argmax(entropy_data_plasma_1058_metodo30_N200000>= 12.29)
```

es decir, es el primer índice ya afectado por el problema de tamaño finito. Pero nosotros no queremos que haya ningún índice afectado por el problema de tamaño finito, luego lo que debemos hacer en realidad es

```
for i in range(2,indice_1058):
        complejidad_1058 +=(i-2)*(-dy2_dx2[i])
```

Uno podría pensar que estas sutilezas teóricas de tomar un índice arriba o abajo son cosas sin importancia, ya que al final esto no se va a ver reflejado en los resultados prácticos. Pero se equivocaría. Veamoslo a través de las siguientes dos figuras.



(a) (b)

Figura 5: Representación de cómo varía la complejidad si calculásemos la complejidad con menos o más índices de lo óptimo deducido en la Sección 2.3, para el logístico, subfigura (a), y para el plasma, subfigura (b). El 0 del eje X representa el índice óptimo, es decir, los valores de las complejidades con los que se han hecho las Figuras 3.5 y 4.6.

Como se puede observar en el caso de los datos reales, variar ligeramente hasta dónde se calcula la complejidad no afecta demasiado, no obstante, al hacerlo en el logístico sí afecta. En este caso, considerar menos de los debidos no hay problemas, pero en cambio, si consideramos más de los debidos vemos que los resultados nos cambian completamente tanto cuantitativa como cualitativamente. Vamos a analizar por qué sucede esto.

Cuando consideramos más índices de los debidos estamos incurriendo en que estamos metiendo índices ya afectados por el problema de tamaño finito. Y mientras que esto no tiene demasiada importancia en las curvas de las entropías de los datos reales ya que son curvas que tienen una extensión de en torno a 500, luego la suma de un termino más a los 500 anteriores no tiene tanto peso. En cambio, en el logístico que son curvas de extensión en torno a 20-25, entonces, la inclusión de un término más o menos sí afecta, y al meter términos afectados por el problema de tamaño finito, estamos ocultando los resultados de verdad, como se puede ver en la gráfica de la izquierda de la Figura 5.

Con esto lo que se quiere dejar claro es que hay que ser cuidadoso con cómo se calcula la complejidad si se está trabajando con curvas de extensión pequeña.

Por último, decir que el resto del código de Python no se da en el texto ya que o bien es repetido a lo dado ya, o bien es solo el código para hacer las representaciones o bien porque lo obtenido mediante el código no es de mayor relevancia para el texto.