ANÁLISIS DE PATRONES APLICADO A PROBLEMAS DE TEORÍA DE NÚMEROS: EL PROBLEMA DE COLLATZ.

E. Bautista-Thompson, A. Villanueva-Becerril & J. Figueroa-Nazuno
Centro de Investigación en Computación.
Instituto Politécnico Nacional.
Unidad Profesional "Adolfo López Mateos" Zacatenco
Col. Lindavista, C.P. 07738
México D. F.
ebautista@correo.cic.ipn.mx jfn@cic.ipn.mx

Resumen

El problema de teoría de números planteado por L. Collatz en 1937 genera series numéricas (Series de Hailstone) que presentan comportamiento complejo difícil de caracterizar en forma analítica. Se presentan los resultados de aplicar técnicas para el Análisis de Sistemas Dinámicos No Lineales, a series obtenidas a partir de números enteros aleatorios entre 1×10^{50} y 1×10^{150}, con el objetivo de identificar patrones característicos en dichas series.

Palabras clave: Series de Hailstone, Mapas de Recurrencia, Análisis de Gramáticas.

1. Introducción.

Tradicionalmente, El Reconocimiento y Clasificación de Patrones son Teorías y Técnicas que se aplican para estudiar si en algún tipo de "señales" o datos registrados de alguna forma, se pueden encontrar Patrones o clases de Patrones. Normalmente se presupone y se sabe de su existencia y los resultados de esos análisis nos dan información o nos hablan de los fenómenos que están atrás de esas señales. Existen cierto tipo de situaciones como las que se analizan en este trabajo, en donde posiblemente las técnicas de análisis de patrones nos ayudan a entender cuáles son los mecanismos o fenómenos que pueden existir o no, en estas series matemáticas. Nuestro interés es aplicar las técnicas modernas de reconocimiento de patrones, en especial las del estudio de los sistemas dinámicos no lineales, para conocer un poco más estas series numéricas, ya que las técnicas analíticas hasta el momento nos han dado muy poca información.

A pesar de que se han desarrollado una gran cantidad de trabajos sobre el Problema de Collatz, (también conocido como Series de Hailstone, Algoritmo de Hasse, Algoritmo de Siracusa entre otros nombres) no ha sido posible resolver dicho problema **[1, 2]**; la simplicidad de las reglas matemáticas del mismo, contrasta con la dificultad para estudiar en forma analítica el comportamiento de las series numéricas que genera. El problema de Collatz se puede plantear de la siguiente forma, sea el algoritmo definido por la función $h: N \to N$ sobre el conjunto de los enteros positivos:

$$h(n) = \begin{cases} \dfrac{n}{2} & si \quad n \quad es \quad par \\ 3n+1 & si \quad n \quad es \quad impar \end{cases}$$

donde n es inicialmente cualquier entero positivo. Para generar la secuencia sea $h^k(n)$ la *kth* iteración de *h(n)*. Una secuencia de números de Hailstone es generada mediante un ciclo de retroalimentación donde:

$$h^k(n) = h(h^{k-1}(n)) \text{ para } k \in N$$

y puede ser representada como

$$H(n) = \{h^k(n)\}\ k = 1, 2, 3, \ldots$$

Ejemplos de conjeturas sobre el comportamiento de estas series que no ha sido posible demostrar en forma analítica son: que para todo número entero se genere una secuencia que converge a 1, y que existan ciclos característicos de números en la Serie de Hailstone **[2]**. La dificultad de este problema llevo al matemático Paul Erdös a afirmar: *"Mathematics is not yet ready for such problems"* **[1, 2]**.

Se ha comprobado numéricamente que el algoritmo genera secuencias de números que convergen a 1 para todos los números iniciales (enteros positivos) $\leq 3 \times 2^{53} \approx 2.702 \times 10^{16}$ **[3]**. El que se pueda generar secuencias de números que presentan un comportamiento complejo a partir de un algoritmo simple, y que no haya sido posible caracterizar dichas secuencias de números en forma analítica, es lo que ha motivado el presente trabajo, en el cual se aplican técnicas utilizadas para el análisis de series de tiempo de sistemas dinámicos no lineales a las secuencias de números, obtenidas mediante el algoritmo ya descrito, que son tratadas como si fuesen series de tiempo generadas por un algoritmo determinista **[4, 5]**. El objetivo es estudiar un conjunto de parámetros e identificar patrones, que caractericen a las Series de Hailstone.

2. Técnicas de Análisis.

En esta sección se describen las técnicas que se utilizaron para el estudio de las Series de Hailstone. La primera técnica utilizada fue el Análisis de Mapas de Recurrencia, la cual genera un mapa (patrón) de la serie numérica. El mapa es una representación gráfica de la integral de correlación y hace posible visualizar la dependencia temporal y espacial de los datos utilizando una medida de distancia euclidiana entre los datos de la serie que es vista por medio de una escala de grises. El Mapa de Recurrencia permite detectar gráficamente (ver Figura 2 y 3) patrones escondidos y cambios estructurales en los datos, o identificar similitudes entre patrones de diferentes series, a su vez permite calcular los siguientes parámetros: Recurrencia, Determinismo y Entropía de Shannon **[5, 6, 7]**; los cuales son descritos a continuación:

Porcentaje de Recurrencia. Permite medir el grado de recurrencia (periodicidad y estructura) entre los datos de la serie, que es indicativo de patrones repetitivos en la misma **[5, 6]**.

Porcentaje de Determinismo. Permite medir el grado de determinismo en el sistema, por medio del análisis de mapas de recurrencia **[5, 6]**.

Entropía de Shannon. La entropía de Shannon, es una medida de la cantidad de información que se obtiene al tomar una medida para especificar el estado del sistema [7].

La otra técnica utilizada fue el Análisis de Gramáticas, la cual genera reglas de producción para la serie numérica. El Análisis de Gramáticas consiste en la generación de una gramática libre de contexto en la cual el conjunto de símbolos terminales (reglas de producción), representa la diferencia entre valores contiguos de la serie numérica. De esta forma, si una parte de la serie presenta el mismo comportamiento en repetidas ocasiones, éste quedará representado por una sola variable de la gramática, esta variable a su vez, puede formar parte de la definición de otra variable y así de manera recursiva. La generación de gramáticas a partir de una serie de tiempo permite dar una medida de complejidad (computacional) en la cual, a mayor número de reglas de producción necesarias para generar una gramática, mayor es la complejidad de la serie numérica **[8, 9]**.

3. Generación de las Series Numéricas.

Se construyeron series de números a partir de semillas generadas en forma aleatoria. Las series de tiempo utilizadas corresponden a semillas dentro del rango de 1×10^{50} a 1×10^{150}, en dicho rango el número de datos para cada serie varía entre 1000 a 5000 datos. Se seleccionaron al azar 10 series por cada incremento de 1000 de tal forma que se obtuvieron 5 grupos de 10 series cada uno. Los extremos del conjunto de 50, fueron series numéricas con 1067 y 5978 datos (ver Figura 1).

Para calcular el mapa de recurrencia y los parámetros derivados, se tuvieron que recortar las Series de Hailstone ya que las fluctuaciones de la secuencia de números son considerables (varios órdenes de magnitud entre valores contiguos) y el tamaño de los valores es sumamente grande (por ejemplo: números enteros del orden de 10^{19}), el recorte se hizo en función de evaluar la serie y observar a partir de qué valor era posible calcular numéricamente el mapa de recurrencia y sus parámetros asociados, lo mismo se hizo para el cálculo de las reglas de producción de gramáticas. Finalmente, el conjunto de 50 series tenía entre 331 y 755 datos por serie.

Una muestra de las series, formada por una de cada grupo, fue recortada nuevamente hasta quedar con los últimos 100 elementos, y se les aplicó el Análisis de Mapa de Recurrencia y de Gramáticas, para estudiar el comportamiento de las series a diferentes escalas. La Figura 1 muestra un ejemplo de una Serie de Hailstone.

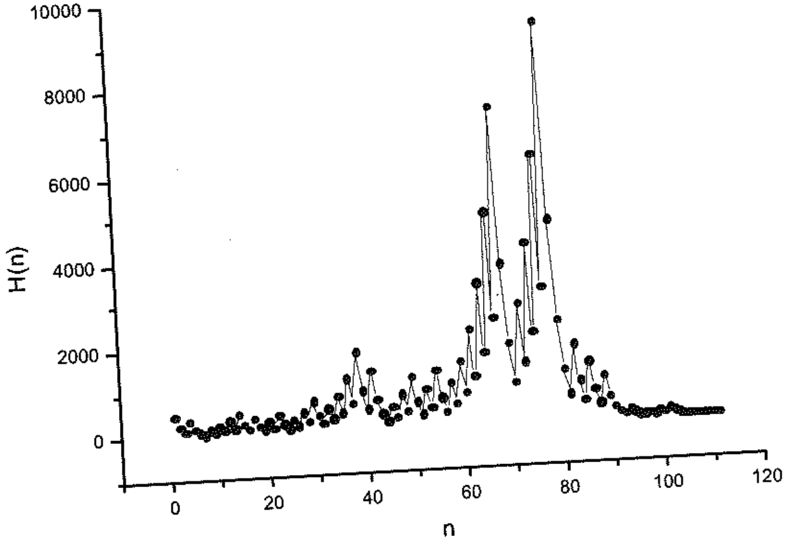

Figura 1. *Gráfica de Serie de Hailstone generada con valor de semilla=500.*

4. Análisis de los Resultados.

El Mapa de Recurrencia genera un patrón característico en el cual se observa que las fluctuaciones de varios órdenes de magnitud de los datos dominan. Lo interesante es que dicho patrón se repite para diferentes segmentos de la serie, lo que indica un comportamiento de autosimilitud en la serie, es decir, para diferentes escalas de las fluctuaciones de la serie el mismo patrón se repite a lo largo de la misma. Esto, además de proporcionar un nuevo patrón para caracterizar las Series de Hailstone nos indica que basta analizar un segmento de la serie para estudiar el comportamiento de su totalidad, las Figuras 2 y 3 muestran el patrón generado con el Mapa de Recurrencia y cómo es similar para diferentes series.

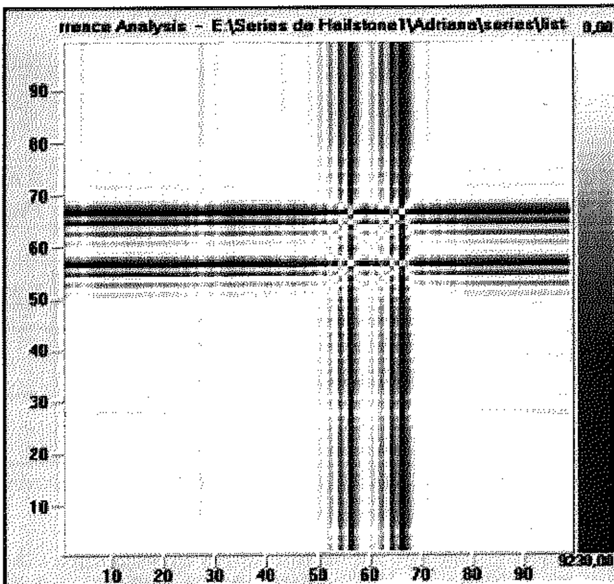

Figura 2. *Mapa de Recurrencia de Serie de Hailstone de 644 datos para los últimos 100 datos.*

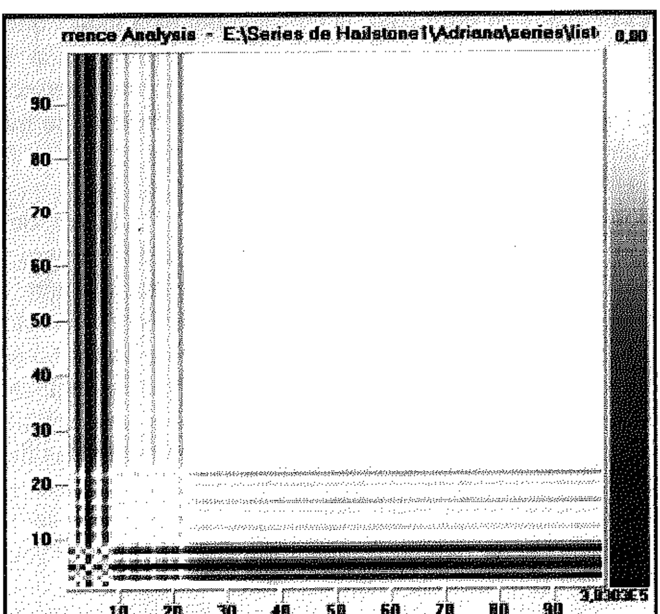

Figura 3. *Mapa de Recurrencia de Serie de Hailstone de 509 datos para los últimos 100 datos.*

La Tabla 1 muestra los resultados obtenidos (promedio, desviación estándar, valor máximo y mínimo) del cálculo de los parámetros de Recurrencia, Determinismo, Entropía de Shannon y el Número de Reglas de Producción para las 50 series numéricas seleccionadas para realizar este trabajo.

	Recurrencia (%)	*Determinismo (%)*	*Entropía (%)*	*# Reglas de Producción*
Promedio	97,363	96,296	6,306	15,86
Desviación Estándar	3,0531195	3,92280357	0,24648981	5,27995826
Valor Máximo	99,805	99,089	6,490	32
Valor Mínimo	86,476	85,629	5,594	8

Tabla 1. *Parámetros obtenidos del Análisis de Mapas de Recurrencia.*

La recurrencia promedio es cercana al 100%, esto indica que hay un alto grado de patrones repetitivos. El determinismo también es cercano al 100%, lo que se refleja en lo estructurado del mapa, esto indica que el generador de la señal es determinista. La entropía de Shannon es de 6.3% en promedio, la entropía tan baja indica que a pesar de poseer un comportamiento determinista, la ganancia en información que puede obtenerse de una medición en la serie de tiempo es pobre, consecuencia del comportamiento complejo de la serie (fluctuaciones de varios órdenes de magnitud y a diferentes escalas a lo largo de la serie).

Con respecto al análisis de gramáticas, se encontró que el número de reglas de producción no es alto, en promedio aproximadamente 16 reglas, (el número máximo de reglas de producción para las series fue de 32) en comparación con análisis de gramáticas realizados a diversos tipos de series de tiempo (por ejemplo, serie de electroencefalograma ECG: 82 reglas, serie económica de Down Jones: 65 reglas) **[10]**, lo cual indica que la Serie de Hailstone posee una complejidad computacional baja. Sin embargo se observa lo siguiente: la Figura 4 muestra que ocurre un crecimiento rápido de la cantidad de reglas de producción generadas, lo que corresponde a las fluctuaciones de la serie de Hailstone (ver la Figura 5), y posteriormente disminuye el ritmo de generación de reglas de producción. Este mismo comportamiento se repite para diferentes segmentos de la serie de tiempo (autosimilitud), tal como se muestra en la Figuras 6 y 7 para los últimos 100 datos de la misma serie.

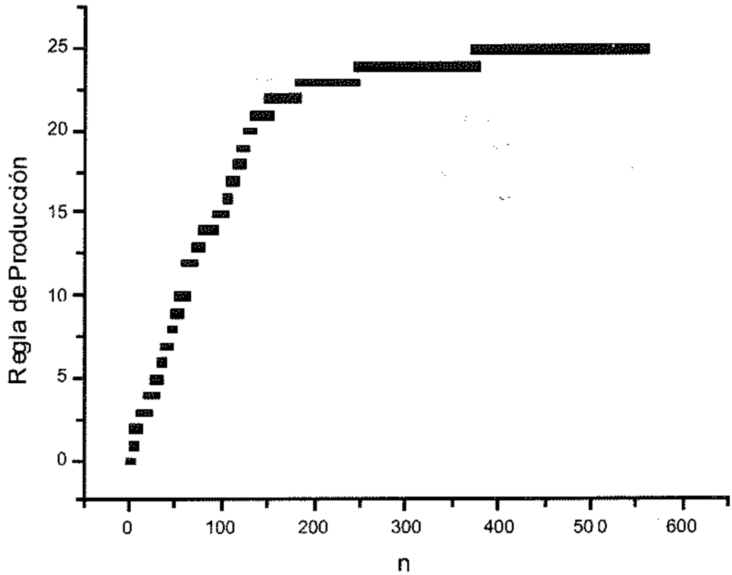

Figura 4. *Gráfica de la evolución de las Reglas de Producción para la Serie de Hailstone con n = 557 datos.*

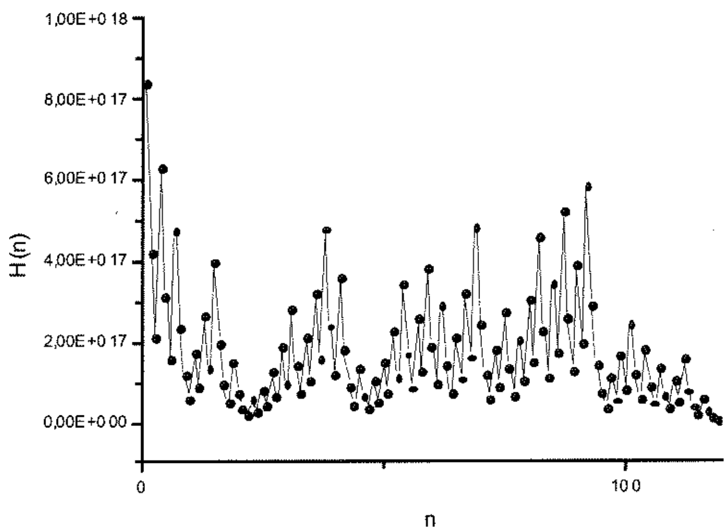

Figura 5. *Gráfica de la Serie de Hailstone con n = 557 datos, primeros 120 datos.*

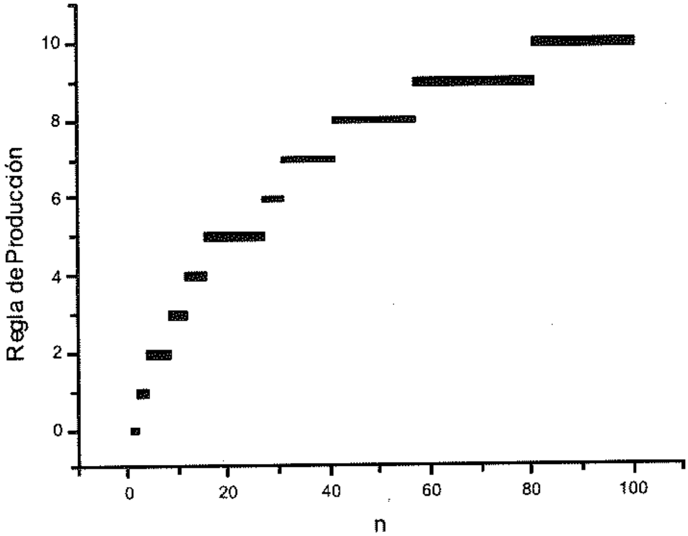

Figura 6. *Gráfica de la evolución de las Reglas de Producción para la misma Serie de Hailstone de 557 datos tomando los últimos n = 100 datos.*

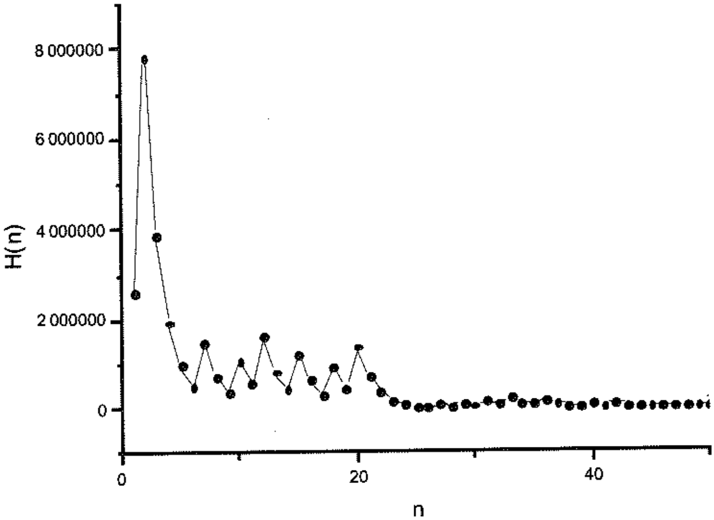

Figura 7. *Gráfica de la Serie de Hailstone de 557 datos tomando solo los últimos n = 100 datos, se muestran 50 datos.*

5. Conclusiones.

Se probó en forma experimental que series numéricas con comportamiento no suave (fluctuaciones de varios órdenes de magnitud, convergencia lenta) que son difíciles de estudiar en forma analítica, pueden ser caracterizadas con técnicas de análisis no lineal y es posible identificar patrones de comportamiento de dichas series.

Mediante los Mapas de Recurrencia se identificó un patrón característico que muestra que la Serie de Hailstone es generada en forma determinista y que existe una estructura que se repite a lo largo de la serie, dicha estructura consiste en las fluctuaciones de varios órdenes de magnitud las cuales dominan en la generación del patrón, dicho comportamiento se observa para diferentes escalas a lo largo de la serie; esto es un fenómeno de autosimilitud.

Analizando los resultados de las series no se observó relación alguna entre la semilla generadora de la serie y los patrones encontrados, es decir los Mapas de Recurrencia son similares para todas las series.

A pesar de que las series poseen un alto grado de recurrencia y determinismo y baja complejidad computacional (expresado en el bajo número de reglas de producción), las fluctuaciones extremas de las series ocasionan la baja Entropía de Shannon observada.

Los resultados en general nos indican que es posible utilizar técnicas de análisis no lineal, en el estudio de estas series numéricas y el hecho de que existan fenómenos de autosimilitud, alta recurrencia y determinismo y baja complejidad computacional son resultados importantes que no se tenían con el estudio exclusivamente analítico de El Problema de Collatz.

Referencias

[1] Pickover, C. A. (1990). *Computers, Pattern, Chaos and Beauty. Graphics from an Unseen World*, New York: ST. MARTIN'S PRESS.

[2] Lagarias, J. C. (1996). The 3x+1 Problem and its Generalizations, ATT Bell Laboratories Internal Report.

[3] Oliveira e Silva, T. (1999). Maximum Excursion and Stopping Time Record-Holders for the 3x + 1 Problem: Computational Results, Math. Comput. 68, 371-384.

[4] Espinosa-Contreras, A., Suárez-Núñez, R. & Figueroa-Nazuno, J. (1999). Análisis del Comportamiento de Diferentes Sismos en la Ciudad de México con Técnicas de la Dinámica No-Lineal. XLII Congreso Nacional de Física, Villahermosa, Tabasco.

[5] Espinosa-Contreras, A. & Figueroa-Nazuno, J. (2000). Análisis del Comportamiento de la Pérdida de Paquetes en la Red Internet con Técnicas de la Dinámica No-Lineal. Congreso Internacional de Computación CIC 2000, México D. F..

[6] Acevedo-Mosqueda, M. E., León-Vega, C. G. & Figueroa-Nazuno, J. (2001). Medición de Complejidad de Series de Tiempo, XLIV Congreso Nacional de Física, Morelia, Michoacán.

[7] Urbach, R. M. A. (2000). *Footprints of Chaos in The Markets*, Financial Times-Prentice Hall.

[8] Menchaca-Méndez R., Sánchez-Rodríguez C., Espinosa-Contreras, A. & Figueroa-Nazuno, J. (2000). Modelado y Caracterización de Series de Tiempo por medio de Análisis Gramatical, XLIII Congreso Nacional de Física, Puebla, Puebla.

[9] Menchaca-Méndez, R., Sánchez-Rodríguez, C. & Figueroa-Nazuno, J. (2001). Predicción de Series de Tiempo Mediante Análisis Gramatical, XLIV Congreso Nacional de Física, Morelia, Michoacán.

[10] Bautista-Thompson, E. & Figueroa-Nazuno, J. (2002). Análisis de Parámetros que Caracterizan la Dificultad de Predicción en Series de Tiempo, XLV Congreso Nacional de Física, León, Guanajuato.

Agradecimientos

Los autores desean expresar su agradecimiento al compañero Ricardo Menchaca por su apoyo con la Técnica de Análisis Gramatical.