Conectando lo discreto con lo continuo; propuesta metodológica para el cálculo de ancho de bin gráfico en el análisis estadístico de datos.

Mayra Angélica Bárcenas Castro[1*], Lorenzo Borselli[2], Ramón Díaz de León-Zapata[1] , Efrén Flores García[1] , Ariel Benjamín de la Rosa Zapata[1]

[1]Instituto Tecnológico de San Luis Potosí, Av. Tecnológico s/n, Soledad de Graciano Sánchez, San Luis Potosí, C.P. 78376, México.

[2]Universidad Autónoma de San Luis Potosí, San Luis Potosí, México.

*e-mail: *mayrabarcenas0422@gmail.com*

Resumen

Se presenta una propuesta metodológica para el análisis de resultados de gráficos estadísticos, en donde el ancho de bin es el principal factor de cambio para su representación. Se implementa un pseudocódigo que permite visualizar los pasos a seguir para obtenerlo. Se realizó un ejercicio para mostrar la pérdida o ganancia de información al implementar la Densidad de Kernel y también la ausencia de esta. La propuesta se hizo directamente con la metodología de Bowman y Azzalini (1997) en Matlab y el generado en este trabajo. Se utilizó el test de Kolmogorov-Smirnov para validar las ventajas del método propuesto. Se encontró que el método propuesto permite una mejor tendencia de los datos ya sea positiva, negativa o simétrica y poca o escasa pérdida de información con respecto evaluación estadística aplicando densidad de kernel; de este modo es una opción para obtener un mejor análisis estadístico cuantitativa y cualitativamente y además puede ser reproducido en temas similares.

Palabras clave: Ancho de bin, Kolmogorov-Smirnov, Densidad de Kernel, Teoría de información.

Abstract

A methodological proposal is presented for the analysis of statistical chart results, where bin width is the main factor of change for representation. A pseudocode is implemented that allows you to visualize the steps to be followed to obtain it. An exercise was generated to show the loss or gain of information when implementing Kernel Density and also the absence of it. The proposal was made directly with the methodology of Bowman and Azzalini (1997) in Matlab and the one generated in this work. The Kolmogorov-Smirnov test was used to validate the advantages of the proposed method. It was found that the proposed method allows a better trend of data either positive, negative or symmetrical and little or little loss of information with respect to statistical evaluation applying kernel density; in this way it is an option to obtain a better statistical analysis quantitatively and qualitatively and can also be reproduced in similar topics.

Key Words: Bin Width, Kolmogorov-Smirnov, Kernel Density, Information Theory.

1. Introducción

Un tema que ha sido de interés entre la comunidad científica es la dicotomía: continúo frente a lo discreto. De los pilares defensores sobre la idea de la continuidad se tiene a Isaac Newton y Gottfriend Leibniz quienes establecieron dicho cálculo diferencial e integral debido a que las funciones conllevan continuidad en los fenómenos descritos. En este sentido una pregunta que hacía referencia Riemann fue si la continuidad siempre se aplica a cualquier fenómeno [1]. Sin embargo Benoit Mendelbrot en 1982 dijo: Las nubes no son esferas, las montañas no son cono, las líneas costeras no son círculos y la corteza no es lisa, tampoco lo

hace el rayo viajar en línea recta" [2]. Es decir los problemas complejos solo se pueden contestar con métodos discretos, lo que significa que la naturaleza es compleja.

En este hilo conductor las metodologías para analizar datos estadísticos resultan de interés tanto en áreas de ciencias exactas como sociales. Si bien es cierto que existen propuestas para realizar con mejor entendimiento la distribución de los datos para visualizar su comportamiento, es importante destacar que los mismos autores redactan que dependerá de los datos y el fenómeno a estudiar para obtener una interpretación fidedigna [3].

El análisis de datos estadísticos ha tomado gran relevancia para la toma de decisiones; es fundamental que en las técnicas que se utilicen se logre obtener información que no refleje demasiado "ruido" como por ejemplo muchas modas. En este contexto desde tiempos de Silverman (1986) se han venido realizando esfuerzos por encontrar una longitud de ancho de bin optimizado, generando algoritmos de acuerdo a las necesidades que están presentes, por ejemplo programas como STATA, SPSS y Matlab por mencionar algunos. Este último Bowman y Azzalini (1997) también consideraban la necesidad de seguir fortaleciendo los algoritmos para una mejor versión de los datos. En este trabajo, se propone una metodología para el tratamiento de datos discretos, de tal forma que se pueda obtener mayor información tanto cualitativa como cuantitativa utilizando la estimación de Kernel.

Objetivos particulares del método

1.– Obtener un ancho de bin personalizado para cada muestra.

2.– Demostrar la pérdida o ganancia de información utilizando la Teoría de Información (Entropía de Shannon).

3.– Validar la información usando el Test de Kolmogórov-Smirnov.

Actualmente se ha venido observando fenómenos caóticos que con la implementación de métodos heurísticos comparado con métodos matemáticos, se dificulta predecir comportamientos para la toma de decisiones en diferentes áreas del conocimiento.

2. **Método y caso aplicativo**

Para fines de este trabajo se utiliza como herramienta de visualización el histograma. Los histogramas son funciones discontinuas. La herramienta para el análisis de los datos será la estimación de Kernel que fue introducido Rosenblatt (1956) y Parzen (1962) y han recibido una considerable atención a los estimadores no paramétricos de densidad de probabilidad para series de tiempo [4], en donde resulta ser una solución para la estimación continua de un histograma.

Cuando se trabaja con datos y se asume un determinado modelo estadístico, existen procedimientos para aceptar o rechazar el modelo, con cierto grado de incertidumbre; podemos recurrir a criterios lógicos o simplemente pragmáticos.

Para realizar este trabajo se utilizaron los resultados de las muestras obtenidas para analizar un depósito de sedimento en cuanto a su morfología y características físicas. Dentro de las muestras se cuantificó utilizando el método de intersección de Rosiwal el diámetro de las rocas. De este modo el geólogo que es el especialista en esta rama sugiere que con esas medidas se pueda observar la concentración o dispersión de dichas rocas y con su posición actual comparar en un futuro la morfología del depósito.

Siguiendo un criterio puramente lógico, por otro lado se define una distancia entre los parámetros de un modelo estadístico (en este caso densidades por ejemplo) cumpliendo ciertas condiciones razonables, de modo que aparece de forma natural la entropía de Shannon.

En este trabajo la entropía de Shannon también será un criterio para medir la eficiencia del modelo (Ver Tabla 1).

Muestra	**Diferencia máxima entre la entropía y el histograma sin Kernel de las muestras**
M0003	0.06720
M0004	0.04872
M0009	0.05347
M0011	0.05283
M0015	0.05872
M0018	0.06792
M0019	0.05964
M0024	0.06245
M0025	0.04642
M0032	0.04889
M0035	0.04480
M0037	0.08251
M0041	0.05185
M0044	0.06243
M0046	0.06685
M0047	0.05280
M0050	0.05146
M0052	0.04427

Tabla 1.- Muestra la diferencia máxima entre la entropía de Shannon y el histograma sin Kernel.

3. **Resultados**

Suponiendo que tiene que analizar un conjunto de datos de un depósito de flujo de escombros en un área delimitada en San Luis Potosí. Dichos datos son el diámetro de las

rocas. En primera instancia el geólogo pretende obtener la mayor información posible, sin embargo se visualiza que entre más pequeño es el ancho de bin, aparecen más modas, pero por otro lado se pierde información al utilizar ancho de bin de 0.5, 0.75 y 1 [6].

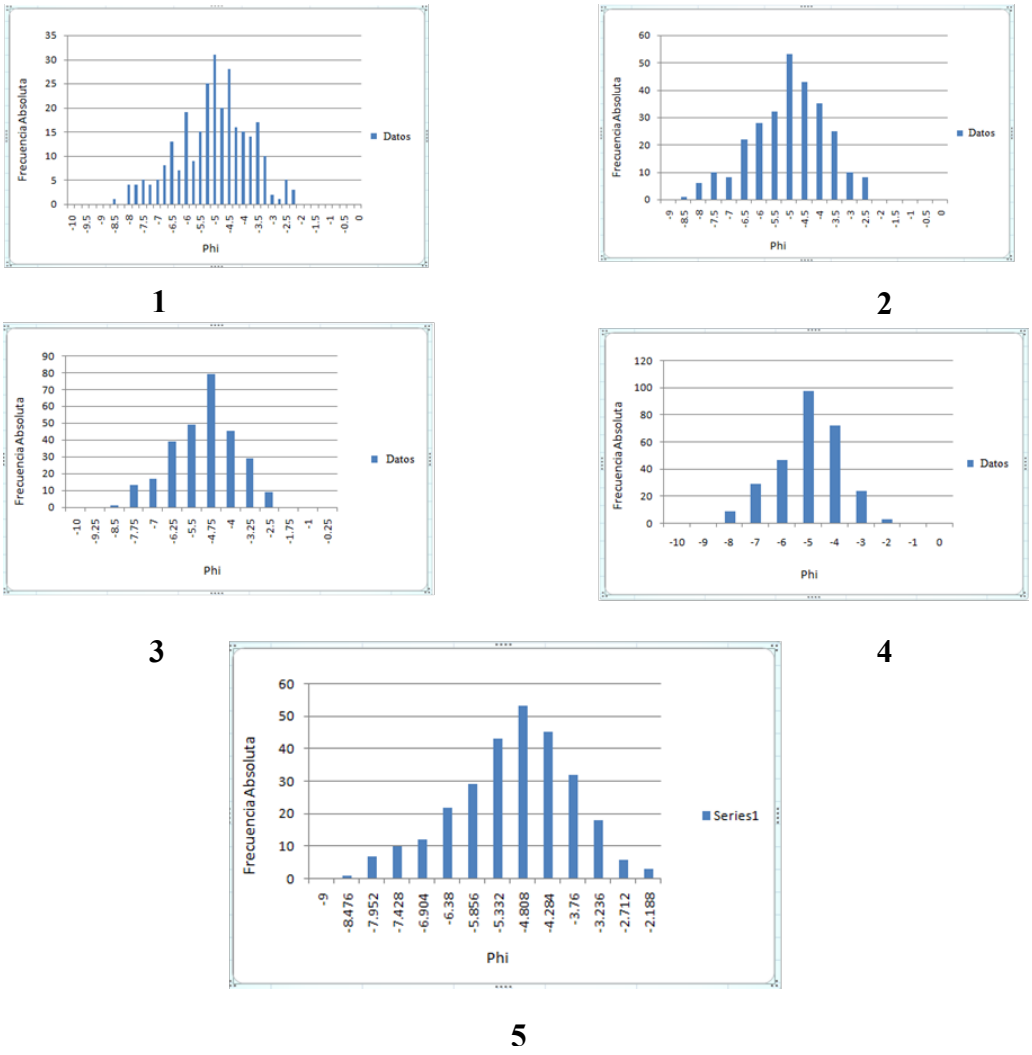


Figura 1.- La distribución 1 corresponde al ancho de bin de 0.25, la segunda a 0.50 la tercera a 0.75, la cuarta a 1 y la quinta se calculó con ancho de bin personalizado con un valor de 0.5240.

En la Figura 1, la distribución 5 se observa la diferencia entre los anchos de bin que usó el geólogo y el bin propuesto en este trabajo.

A continuación se muestran los anchos de bin obtenidos con los datos de las muestras del depósito de rocas.

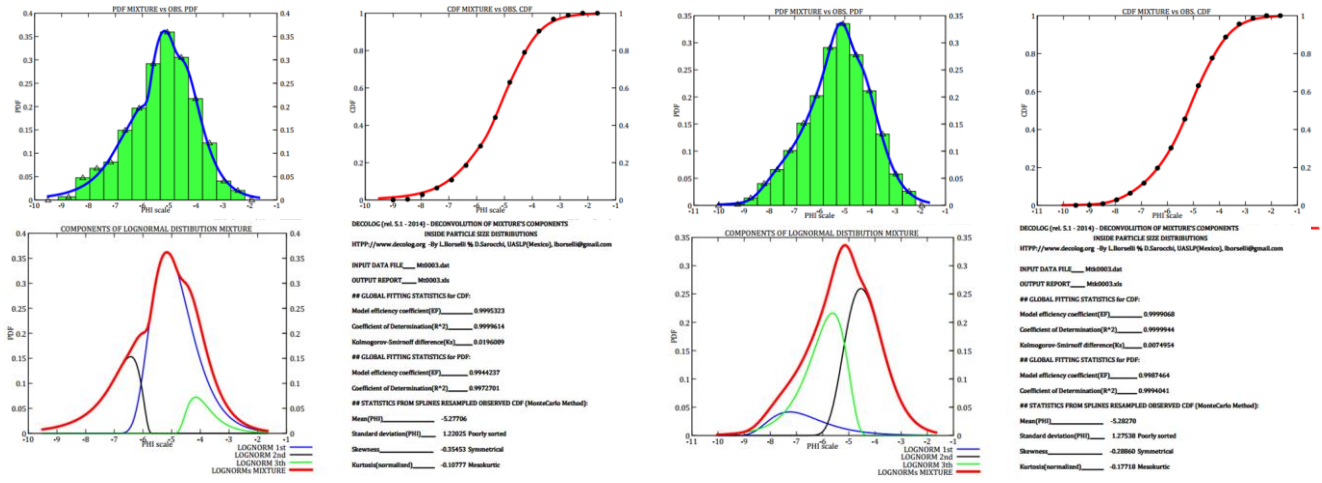
Numero de Muestra	[a,b,optphi,dife]=anchodebanda(x,phimin,phimax,steps)	Ancho de Banda óptimo	Resultado Test de Kolmogórov-Smirnov
M0003	[a,b,optphi,dife]=anchodebanda(x,-9,-2,[0.05:.001:1])	0.5240	0.0714
M0004	[a,b,optphi,dife]=anchodebanda(x,-10,-3,[0.05:.001:1])	0.3360	0.0952
M0009	[a,b,optphi,dife]=anchodebanda(x,-9,0,[0.05:.001:1])	0.6870	0.0769
M0011	[a,b,optphi,dife]=anchodebanda(x,-9.5,-2,[0.05:.001:1])	0.5160	0.0667
M0015	[a,b,optphi,dife]=anchodebanda(x,-9,-3,[0.05:.001:1])	0.4780	0.0667
M0018	[a,b,optphi,dife]=anchodebanda(x,-10,-2,[0.05:.001:1])	0.7210	0.1000
M0019	[a,b,optphi,dife]=anchodebanda(x,-9,-2,[0.05:.001:1])	0.5710	0.0769
M0024	[a,b,optphi,dife]=anchodebanda(x,-9,-2.5,[0.05:.001:1])	0.4630	0.0667
M0025	[a,b,optphi,dife]=anchodebanda(x,-9,-2.5,[0.05:.001:1])	0.4710	0.0714
M0032	[a,b,optphi,dife]=anchodebanda(x,-9.5,-1.5,[0.05:.001:1])	0.7720	0.0909
M0035	[a,b,optphi,dife]=anchodebanda(x,-8.5,-1,[0.05:.001:1])	0.6260	0.0833
M0037	[a,b,optphi,dife]=anchodebanda(x,-11,-2,[0.05:.001:1])	0.6250	0.0833
M0041	[a,b,optphi,dife]=anchodebanda(x,-9.5,-3,[0.05:.001:1])	0.6070	0.0909
M0044	[a,b,optphi,dife]=anchodebanda(x,-10,-0,[0.05:.001:1])	0.8250	0.0769
M0046	[a,b,optphi,dife]=anchodebanda(x,-10,-2,[0.05:.001:1])	0.8080	0.1000
M0047	[a,b,optphi,dife]=anchodebanda(x,-9.5,-3,[0.05:.001:1])	0.5500	0.0833
M0050	[a,b,optphi,dife]=anchodebanda(x,-9,-2,[0.05:.001:1])	0.6000	0.0833
M0052	[a,b,optphi,dife]=anchodebanda(x,-9,-1.5,[0.05:.001:1])	0.0909	0.0787

Tabla 2 Se visualiza los diferentes anchos de bin para cada muestra con el método propuesto en este trabajo.

El test de Kolmogórov-Smirnov se utiliza como prueba no paramétrica que determina la bondad de ajuste entre la distribución con el ancho de bin propuesto en este trabajo y la distribución usando la entropía de Shannon (Ver Tabla 2).

Muestra	Número de datos	Test de Kolmogórov-Smirnov del histograma con ancho de bin personalizado	Test de Kolmogórov-Smirnov del histograma de densidad de Kernel con ancho de bin personalizado
M0003	281	0.0196089	0.0074954
M0004	180	0.0508785	0.0131471
M0009	303	0.0267919	0.0111684
M0011	614	0.0114390	0.0022671
M0015	309	0.0070604	0.0027111
M0018	261	0.0584591	0.0064221
M0019	440	0.0179237	0.0061904
M0024	326	0.0231141	0.0078664
M0025	259	0.0354217	0.0099706
M0032	258	0.0105814	0.0143452
M0035	267	0.0078935	0.0053063
M0037	205	0.0144139	0.0022295
M0041	277	0.0147551	0.0108268
M0044	449	0.0161349	0.0021035
M0046	281	0.0095966	0.0150991
M0047	303	0.0136268	0.0129752
M0050	354	0.0139128	0.0038347
M0052	441	0.0143547	0.0071734

Tabla 3 Se visualiza las diferencias utilizando el Test de Kolmogorov– Smirnov con y sin Kernel.



1 **2**

Gráfica 1.-Traslación de datos discretos a continuos usando el ancho de bin propuesto sin la estimación de Kernel, **Gráfico 2** .-Lo mismo que la primera pero con estimación Kernel, para su elaboración se utilizó Software Decolog 5.6.1 que se puede consultar de acuerdo a [7].

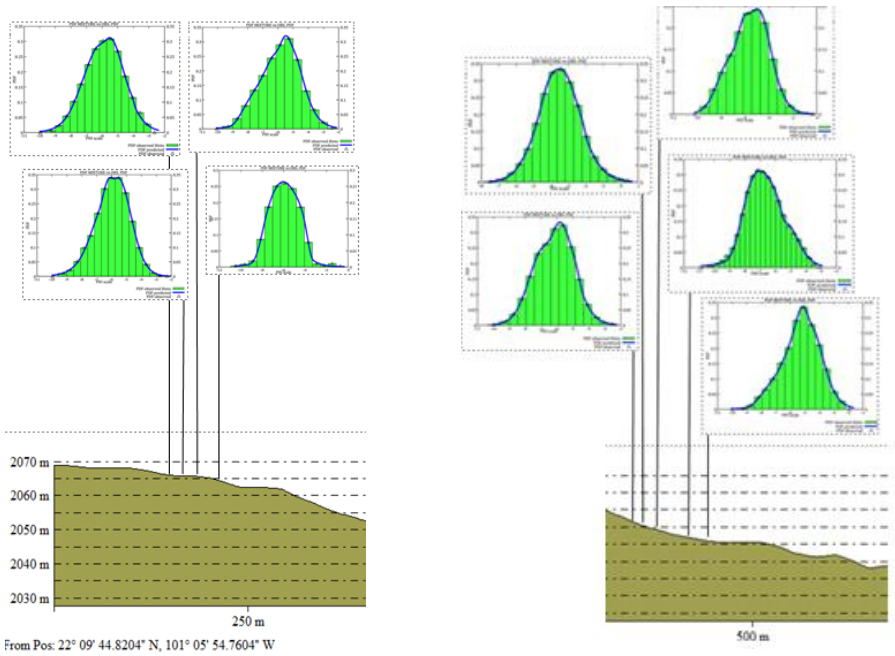


Figura 2. Resultado de la primera parte de las muestras usando el ancho de bin en este trabajo y su representación gráfica aplicado densidad de kernel.

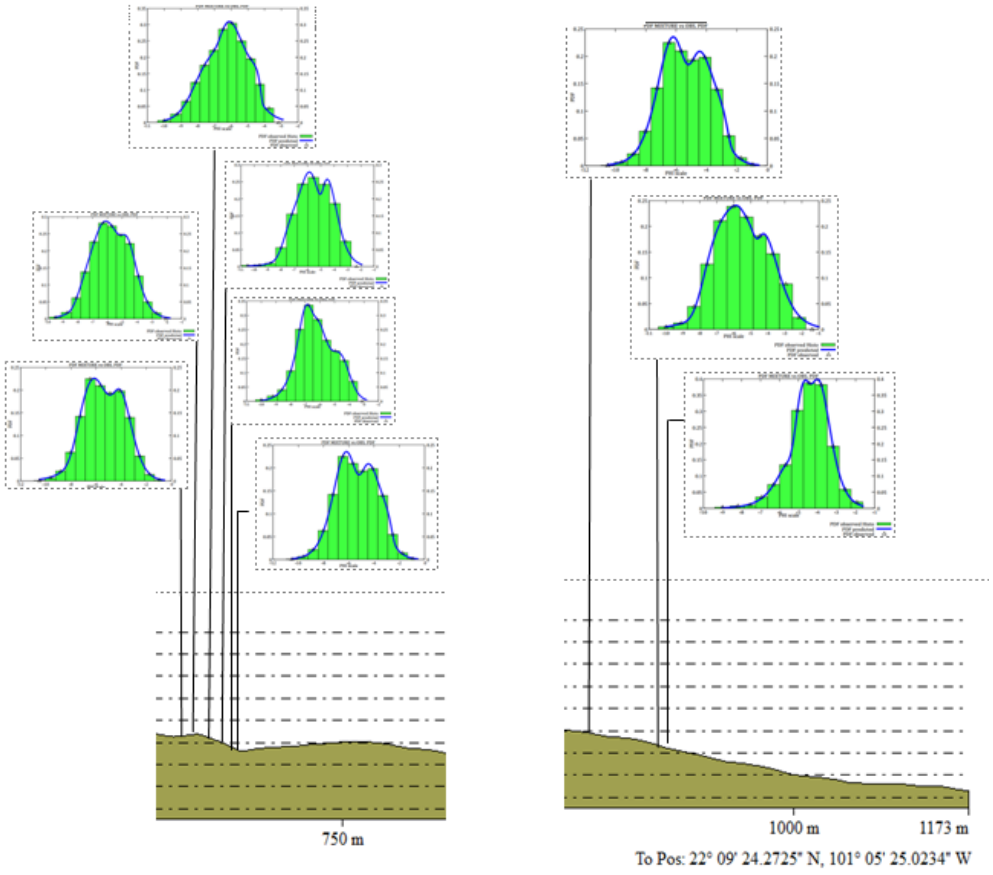


Figura 3. Resultado de la segunda parte de las muestras usando el ancho de bin en este trabajo y su representación gráfica aplicado densidad de kernel.

A continuación se muestra el Pseudocódigo para obtener el ancho de bin propuesto en este trabajo.

PSEUDOCÓDIGO

```
Proceso Ancho_de_bin_personalizado
     mostrar "Inserte los datos de su muestra (mínimo
100)"
     leer x;
     escribir
"[histx,pt]=hist(x,phimínimo:[0.05:0.001:1]:phimáximo)";
     escribir
"[histk,r]=kdensity(x,phimínimo:[0.05:0.001:1]:phimáximo]"
;
     escribir "histx=histx/max(histx)";
     escribir "histk=histk/max(histk)";
     escribir "[h,p,ks2stat]=kstest2[histx,histk]";
     escribir "a(i)","Este es el valor de ancho de bin
personalizado";
     escribir "b(i)" ,"Este es el valor del test de
Kolmogorov-Smirnov";
     escribir "[dife,ll]=min(b)","posición del vector en
b";
     escribir "optphi=a(ll)","posición del vector a";
     escribir "repetir a partir del paso 4 hasta el 7";
          escribir "bar(pt,histx)";
          escribir "plot(pt,histk)"
FinProceso
```

4. **Conclusión**

Esta metodología se puede reproducir en otras áreas y es de interés comparar resultados con trabajos similares. Se debe tomar en cuenta que para utilizar esta propuesta se necesita mínimo cien datos por muestra (Ver PSEUDOCÓDIGO).

Se obtuvo de la traslación de usar el histograma sin densidad de kernel con respecto a introducirla una mejor visualización de los datos de manera cualitativa y además cuantitativamente usando el Test de Kolmogorov-Smirnov donde se pudo notar que fue muy poca la diferencia la distorsión de ajuste.

Por otro lado para cada muestra se obtuvo un ancho de bin diferente y por lo tanto la tendencia de los datos no fue simétrica sino más bien para cada muestra se genera una tendencia positiva o negativa; en el caso de la Gráfica 1 y 2 la tendencia de los datos es positiva.

En la Figura 2 y 3 se puede observar los resultados de los análisis de todas las muestras y su respectiva gráfica que son diferentes tanto la concentración como dispersión del tamaño de diámetro de las rocas. En este sentido las rocas más grandes se encontraron en la primera parte y cerca del centro, mientras que las rocas más pequeñas como por ejemplo sedimento de arena la concentración se encontró en la segunda parte casi al final del depósito estudiado.

Esta metodología ha sido aceptada para continuar reforzándola en la línea de investigación "Connecting the Discrete and the Continuous: Model Generation, from Rule Models to Equational Models" con el Dr. Hector Zenil Co-líder del Laboratorio de Dinámica Algorítmica en el Instituto Karolinska en Estocolmo, Suecia [5]. La idea es que se realicen pruebas usando la Dinámica de Información Algorítmica como una aproximidad computacional causal. En este contexto se pretende encontrar la probabilidad algorítimica de Solomonoff-Levin conectado con la complejidad algorítmica de Kolmogórov-Chaitin. Entre mayor sea la probabilidad algorítmica menor será la complejidad algorítmica, de este modo esta teoría permite formular nuevas predicciones a cualquier sistema, lo que implica mayor información de los datos analizados.

Referencias

[1] Ozelim, L. C. D. S. M., Cavalcante, A. L. B., & Borges, L. P. D. F. (2012). Continuum versus discrete: a physically inter pretable general rule for cellular automata by means of modu lar arithmetic. arXiv preprint arXiv:1206.2556.

[2] Mandelbrot, B. B. (1982). The fractal geometry of nature (Vol. 1). New York: WH freeman.

[3] Galindo Huerta, A. (2018). Algoritmos de clasificación para datasets desequilibrados: análisis y comparativa.
[4] Lake, D. (2009). Nonparametric entropy estimationn using kernel densities. En M. L. Johnson, & L. Brand (Edits.), Methods in Enzymology, Computer Methods, Parte B (Vol. 467, pág. 531). United States of America: Elsevier
[5] Zenil, H., Kiani, N. A., Marabita, F., Deng, Y., Elias, S., Sch midt, A., ... & Tegner, J. (2018). An Algorithmic Information Calculus for Causal Discovery and Reprogramming Syst ems. Available at SSRN 3193409.
[6] Barcenas C.M. (2015). Desarrollo de una metodología estadística para el análisis sedimentológico de depósitos de flujos de escombro: aplicaciones en el área de San Luis Potosí.
[7] Página Web– Decolog 5.6.1 https://www.lorenzo-borselli.eu/decolog/