TALLER: CIENCIA DE DATOS

XXVIII SEMANA NACIONAL DE INVESTIGACIÓN Y DOCENCIA EN MATEMÁTICAS,
UNIVERSIDAD DE SONORA
2018

El taller de Ciencia de los Datos es organizado por académicos del Departamento de Matemáticas de la Universidad de Sonora, en el marco de las actividades de la XXVIII Semana Nacional de Investigación y Docencia en Matemáticas. Este taller tiene como objetivo presentar y promover distintas herramientas para el análisis de datos (tales como el análisis topológico de datos, series de tiempo, regresión logística, curvas ROC, entre otros), a través de conferencias, cursos y sesiones de trabajo.

Entre los conferencistas invitados se encuentran el Dr. José Perea (Department of Computational Mathematics, Science and Engineering; Department of Mathematics of the Michigan State University) y el Dr. Pedro Miramontes (Facultad de Ciencias de la Universidad Nacional Autónoma de México).

Además de los cursos y ponencias por invitación se presentarán proyectos multidisciplinarios en análisis de datos, desarrollados entre el Departamento de Matemáticas de la Universidad de Sonora e instituciones públicas, en áreas de la salud, infraestructura urbana y aspectos climáticos.

El comité organizador local está integrado por los profesores:
Jesús F. Espinoza (jesus.espinoza@mat.uson.mx)
Gudelia Figueroa (gfiguero@mat.uson.mx)
Rosalía G. Hernández (rosalia.hdez@mat.uson.mx)
José A. Montoya (montoya@mat.uson.mx)

Horario de presentaciones

<table>
<tr><th></th><th>Lunes 5</th><th>Martes 6</th><th>Miércoles 7</th></tr>
<tr><td>09:00 – 10:00</td><td>TCD–01</td><td>TCD–03</td><td>TCD–04</td></tr>
<tr><td rowspan="2">10:00 – 11:00</td><td rowspan="4">TCD–02</td><td rowspan="4">TCD–02</td><td>TCD–05</td></tr>
<tr><td>TCD–06</td></tr>
<tr><td rowspan="2">11:00 – 12:00</td><td>TCD–07</td></tr>
<tr><td>TCD–08</td></tr>
</table>

Todas las sesiones del Taller de Ciencia de Datos se realizarán en la Sala Audiovisual del Departmamento de Matemáticas, Edificio 3K-3, primer piso.

Lunes 5 de marzo, 2018

TCD–01 La distribución Beta Discreta Generalizada en la Ciencia de datos

Dr. Pedro Miramontes
Universidad Nacional Autónoma de México.

Resumen: Aunque no se puede fijar con precisión la fecha lo que hora conocemos con el nombre de Ciencia de Datos, si podemos ubicar en esa categoría el trabajo de George Kingsley Zipf quien en las décadas de los años treinta y cuarentas del siglo pasado se dio a la tarea de analizar corpus lingüísticos con la finalidad de encontrar regularidades. Después de analizar textos en inglés, enunció la ley que ahora lleva su nombre y que se escribe como una ley de potencias:

$$f(r) = \frac{A}{r^{\alpha}}.$$

Las leyes de potencia tienen propiedades geométricas muy interesantes, entre las cuales destaca la invarianza ante cambios de escala. Siendo ésta una propiedad de los fractales autosemejantes, no es de extraar el impacto que la Ley de Zipf ha tenido.

Recientemente, al analizar enormes corpus literarios, se han encontrado desviaciones de la Ley de Zipf en las regiones de baja frecuencia en diagramas rango-orden. Nuestro grupo de trabajo ha propuesto una variante de la función Beta que corrige estas desviaciones [1,2,3].

En esta charla, se presenta dicha propuesta y se muestran sus aplicaciones a diversas situaciones; en particular, a la distribución de las unidades administrativas (municipios) de todo el mundo.

Referencias.

1. R. Mansilla, E. Köppen, G. Cocho, P. Miramontes. *On the behavior of journal impact factor rank-order distribution*. Journal of Informetrics. Vol 1, pp 155-160. 2007.
2. Gustavo Martínez-Mekler, Roberto Alvarez Martínez, Manuel Beltrán del Río, Ricardo Mansilla, Pedro Miramontes, Germinal Cocho. *Universality of Rank-Ordering Distributions in the Arts and Sciences*. Plos One. 2009. `https://doi.org/10.1371/journal.pone.0004791`.
3. Oscar Fontanelli, Pedro Miramontes, Germinal Cocho, Wentian Li. *Population patterns in Worlds administrative units*. Royal Society Open Science. 2017.
DOI: 10.1098/rsos.170281.

TCD–02 Análisis topológico de series de tiempo

Dr. José A. Perea
Department of Computational Mathematics, Science and Engineering
Michigan State University.

Resumen: Las observaciones que varían con el tiempo son omnipresentes en el mundo rico en datos en el vivimos. Ejemplos incluyen: series temporales de valores reales (como mediciones de sonido y temperatura), videos (considerados como sucesiones de imágenes) y redes dinámicas (nuevamente, sucesiones de grafos).

En los últimos años, las herramientas del análisis topológico de datos, los sistemas dinámicos y el análisis no lineal de series temporales han sido combinadas y adaptadas al análisis de datos de series temporales multimodales. En resumen, las series de tiempo pueden transformarse en nubes de puntos de alta dimensión (mediante encajes de dilación) y su forma puede ser cuantificada mediante análisis topológico de datos (por ejemplo, con homología persistente). Esto permite cuantificar características tales como periodicidad, cuasiperiodicidad, existencia de motivos, presencia de caos dinámico, etc. Este curso cubrirá

los principales aspectos teóricos detrás del análisis topológico de series de tiempo, los problemas computacionales asociados, y se explorarán aplicaciones que van desde la genética hasta las ciencias del habla.

Martes 6 de marzo, 2018

TCD–03 Análisis topológico de islas de calor urbanas

Jesús F. Espinoza[1], Gudelia Figueroa P.[1], Rosalía G. Hernández[1], José A. Montoya[1], Agustín Morúa[2], Javier Navarro E.[2], Hugo Valenzuela Ch.[1]
[1]Universidad de Sonora; [2]Instituto Tecnológico de Sonora.

Resumen: En esta ponencia se presenta un análisis del fenómeno de islas de calor urbana mediante una novedosa herramienta conocida como *análisis topológico de datos*. Presentaremos como caso de estudio el análisis realizado para la ciudad de Hermosillo, Sonora.

Una isla de calor urbana (UHI por sus siglas en ingés) es una región en la que la temperatura es más alta que la del área circunvecina. La existencia de estas regiones de mayor temperatura puede estar relacionada con distintos factores como la densidad habitacional, falta de áreas verdes, uso de suelo para infraestructura industrial, entre otros. Este fenómemo puede afectar la calidad de vida y la salud de la población en ciudades donde las temperaturas son significativamente altas. En consecuencia, es necesario un completo entendimiento de dicho fenómeno con el objetivo de promover políticas de salud e infraestructura urbana adecuadas, para subsanar sus efectos.

TCD–02 Curso (Parte II): Análisis topológico de series de tiempo, Dr. José A. Perea.

Miércoles 7 de marzo, 2018

TCD–04 Índice de riesgo de malignidad en tumoraciones anexiales

Carlos Dávila[1], Jesús F. Espinoza[2], Gudelia Figueroa P.[2], Rosalía G. Hernández[2], José A. Montoya[2]
[1]Hospital Integral de la Mujer del Estado de Sonora; [2]Universidad de Sonora.

Resumen: De acuerdo al Instituto Nacional de Cancerología, el cáncer de ovario es la primera causa de mortalidad entre los cánceres de origen ginecológico a nivel mundial, ocupando el tercer lugar en nuestro país; de aquí, la importancia de desarrollar un diagnóstico preciso para la evaluación de la malignidad de masas anexiales. Existen distintos modelos diagnósticos que permiten detectar la naturaleza maligna o benigna de tumores anexiales mediante una evaluación preoperatoria, esto es, sin la necesidad de contar con un reporte anatomopatológico postoperatorio, de tal forma que además de reducir costos, se evita deteriorar la salud fisiológica y psicológica de la paciente.

Tales evaluaciones preoperatorias son basadas en lo que se denomina ndice de Riesgo de Malignidad (IRM). Un índice de este tipo es por ejemplo, el índice de Jacobs, el cual se calcula como un producto poderado asociado al valor del puntaje ultrasonográfico, el estado menopáusico y la cantidad de CA-125.

En el presente trabajo, colaboración entre académicos de la Universidad de Sonora (UNISON) y el Hospital Integral de la Mujer del Estado de Sonora (HIMES), tiene como objetivo la evaluación de la utilidad del uso de un IRM para predecir malignidad en tumoraciones

anexiales en pacientes del HIMES, a través de un análisis estadístico que permita determinar un punto de corte óptimo en términos de especificidad/sensibilidad de distintos índices, y el correspondiente análisis comparativo.

TCD–05 Análisis no lineal de series de tiempo usando el Teorema de inmersión de Takens

Roxana Wendoline Ruíz
Universidad Nacional Autónoma de México.

Resumen: A lo largo del tiempo, muchos investigadores se han dado a la tarea de tratar de extraer información de las series de tiempo; en particular de series fisiológicas. Para ello se han utilizado muchas técnicas. En esta plática hablaremos del método usando el Teorema de inmersión de Takens, el cual se describe brevemente a continuacin.
Dada una serie de tiempo $\{x(t_i)\}_{i=1}^n \subseteq \mathbb{R}$ obtenemos una nube de N puntos $\{\vec{x}_i\}_{i=1}^N \subseteq \mathbb{R}^m$ de la siguiente manera:

$$\vec{x}_i = (x(t_i), x(t_i+\tau), x(t_i+2\tau), ..., x(t_i+(m-1)\tau)), \quad i \in \{1, ..., N\},$$

donde $\tau > 0$ es el *tiempo de retraso*, m es la *dimensión de inmersión* y $N = n-(m-1)\tau$. En esta plática discutiremos los detalles precisos para estimar los parámetros τ y m, e ilustraremos con series de tiempo provenientes de modelos.

TCD–06 Impacto de acciones realizadas frente a un brote de dengue: Dinámica de la enfermedad

Mayra Rosalía Tocto Erazo[1], José A. Montoya[1], Daniel Olmos Liceaga[1], Saúl Díaz Infante Velasco[1], Pablo Alejandro Reyes Castro[2] y Ana Lucia Castro Luque[2]
[1]Universidad de Sonora, [2]Colegio de Sonora .

Resumen: Ante situaciones de brotes de dengue, generalmente se realizan acciones tales como vigilancia epidemiológica, campañas de prevención, aplicación de insecticidas, entre otras, con el objetivo de reducir significativamente la poblacin susceptible a esta enfermedad.

En el 2010, ocurrió un brote de dengue durante 18 semanas en la ciudad Hermosillo, cuyo número de casos nuevos por semana presenta una compleja dinámica que puede haberse dado por diversos factores. En este trabajo se considera una área pequeña de la ciudad que abarca una de las colonias con mayor incidencia durante ese brote, y que tiene características socio-económicas y demográficas similares.

Se propone un modelo estadístico para describir matemáticamente el número de casos nuevos de dengue por semana, donde las observaciones vienen de una distribución de Poisson cuya media es la solución de un modelo tipo SIR (Susceptible - Infectado - Removido). Este modelo SIR modificado incluye un parámetro que considera las acciones realizadas ante un brote. Se empleó un enfoque de verosimilitud para estimar los parámetros del modelo. Los resultados muestran que el modelo estadístico planteado se ajusta a los datos. Además, se presenta una banda de posibles escenarios que hubieran ocurrido ante la ausencia de acciones, empleándose sólo los datos observados durante las primeras cinco semanas del brote.

TCD–07 Análisis estadístico de días de calor extremo en el Estado de Sonora

Jorge Alberto Espíndola Zepeda[1], José A. Montoya[1] y Javier Navarro Estupiñán[2]
[1]Universidad de Sonora; [2]Instituto Tecnológico de Sonora.

Resumen: Las variaciones de temperaturas extremas tienen una importancia muy particular ya que afectan a la salud pública, fuentes de agua, demanda de energía, biodiversidad, agricultura, ganadería, etc. En particular, Sonora es uno de los estados de México que presenta mayores condiciones secas y calientes en el país, llegando a registrar temperaturas máximas durante el verano por encima de los 48.5°C. Este periodo de calor excesivo en la

región, afecta la salud humana debido al estrés y condiciones adversas, provocando enfermedades cardiovasculares, cerebrovasculares y respiratorias. De hecho, los riesgos para la población se pueden incrementar aún más si se toman en cuenta factores como edad, nivel de ingreso, nivel de insolación social, si se trabaja con o sin aire acondicionado, etc.

En este trabajo se aborda el problema de analizar el fenómeno aleatorio llamado Días de Calor Extremo (DC) en el Estado de Sonora, donde los DC se definen como el número de días donde la temperatura máxima se encuentra por encima de un umbral establecido para el periodo de estudio (verano). Se propone considerar tanto la definición clásica de umbral como una novedosa definición desarrollada con base en la Teoría de Valores Extremos. En ambos casos, se propone un modelo lineal generalizado Poisson, donde la variable explicativa es la temperatura máxima promedio para el periodo de verano. Es importante enfatizar que este tipo de modelo propuesto es usado para analizar el comportamiento futuro (corto, mediano y largo plazo) de los DC en el Estado de Sonora. Para éste trabajo se utilizaron datos de estaciones meteorológicas de la Comisión Nacional del Agua (CONAGUA), ubicadas en el centro, norte, sur, este y oeste del Estado de Sonora.

TCD–08 Variabilidad oceanográfica de la zona costera del Estado de Sonora

José A. Montoya, Itchel Nathaly Osuna Llamas, Daniel Eduardo Fernández Villalobos y Carlos Manuel Robles Tamayo
Universidad de Sonora.

Resumen: Los patrones espaciales y temporales de variación de la temperatura superficial del mar (TSM) juegan un papel fundamental en la determinación de las condiciones para la supervivencia de los organismos que habitan en las aguas poco profundas. En esta ponencia se presentarán avances que determinan la variabilidad de la Temperatura Superficial del Mar (TSM), así como también la variabilidad de la biomasa Fito planctónica en base al contenido de Clorofila a (Chl a) a lo largo de la zona costera del Estado de Sonora.