HACIENDO CIENCIA DE DATOS

Posted on 16 marzo, 2017 by Celia del Carmen Escamilla Rivera y Verenice Escamilla Rivera

Para hacer ciencia de datos se requiere de al menos un modelo, de un proceso mental determinado y de la adopción de algunas hipótesis. Una vez hecho esto, se toman los datos y se aplican las herramientas estadísticas que abundan en la literatura.

Category: Ciencia

Tag: Ciencias Exactas

Para hacer ciencia de datos se requiere de al menos un modelo, de un proceso mental determinado y de la adopción de algunas hipótesis. Una vez hecho esto, se toman los datos y se aplican las herramientas estadísticas que abundan en la literatura.

Cuando empezamos a estudiar estadística una cosa es clara, el gran problema son los cálculos y las operaciones aritméticas. En el pasado se utilizaban papel pautado y formularios para realizar las operaciones que nos conducían hacia los resultados esperados. Los orígenes históricos de la estadística hay que buscarlos en los procesos de recogida de datos, censos y registros sistemáticos,

asumiendo un papel asimilable a una aritmética estatal para asistir al gobernante, que necesitaba conocer la riqueza y el número de sus súbditos con fines tributarios y políticos.

El gran desarrollo actual de la estadística y la probabilidad depende de la utilización de las computadoras, no necesariamente grandes computadoras, sino fundamentalmente de la computadoras personales, las cuales han permitido que una gran cantidad de profesionales puedan colaborar en múltiples plataformas de investigación, desde todos los rincones del mundo, y ayudar al desarrollo de las aplicaciones en muchas especialidades científicas.

Si actualmente las dificultades para la aplicación de la estadística no residen en los cálculos y además existen muchos programas informáticos para estas aplicaciones, entonces ¿en dónde residen los problemas? En general, la respuesta a esta pregunta la podemos expresar así: " hay que aplicar correctamente las pruebas respetando escrupulosamente las condiciones en las que estás deben realizarse". Es decir, debemos ajustar de forma correcta los modelos matemáticos que subyacen detrás de cada prueba.

La existencia de fenómenos o experimentos no determinísticos, donde el conocimiento de las condiciones en las que éstos se desarrollan no garantiza los resultados, hace imprescindible el uso de una función que asigne niveles de certidumbre a cada uno de los desenlaces del fenómeno, y ahí es donde aparece la probabilidad.

Una correcta proyección de estos conceptos es lo que va a permitir estudiar grandes colectivos a partir de pequeñas partes de ellos, llamadas muestras. Y como en la mayoría de los descubrimientos, la noción de todas estas ideas se ha ido desarrollando a lo largo del tiempo en función de la necesidad, de los recursos y de la aportación talentosa de los estudiosos del tema.

El poder de procesar y hacer estadistica de datos

¿En dónde podemos hacer uso de los conceptos anteriores? Existe una variedad de áreas en donde el empleo de la ciencia de datos es indispensable. Desde los negocios, pasando por el estudio de la ecología hasta a escalas del universo en que vivimos. Con el fin de ilustrar cómo funciona la idea de procesar datos presentamos tres ejemplos en diferentes áreas de investigación.

El primer ejemplo podemos enfocarlo al área de cosmología, la ciencia que se encarga del estudio del origen y evolución del universo. Sabemos, gracias a las misiones espaciales, que éste está en expansión, y ha estado comportándose así desde el momento en que nació. Este tipo de fenómeno podemos medirlo a través de una cantidad: la velocidad, la cual puede calcularse al mirar las numerosas explosiones de estrellas distantes. De esta manera podemos medir la rapidez con que se están alejando de nosotros, es decir, expandiéndose en el espacio. Estas observaciones generan una gran cantidad de datos que podemos analizar mediante la estadística. Casos interesantes surgen cuando al momento de proponer un modelo teórico (el cual se busca contenga todas las respuestas de la naturaleza del comportamiento del universo) pueda ajustarse a los datos que se

observan directamente, lo cual demostraría que ese "modelo" puede ser aplicable para situaciones futuras. Entonces, el aspecto gráfico-estadístico de este ejemplo podemos detallarlo como en la Figura 1, donde observamos una muestra de alrededor de 730 explosiones de estrellas conocidas como Supernovas y las comparamos con la estadística de un modelo teórico cosmológico. Vemos que ambos resultados logran ser equiparables.

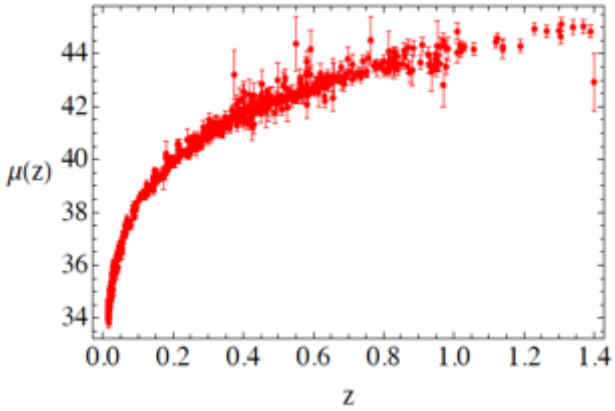

Figura 1a. Datos procesados usando el software Mathematica Wolfram.

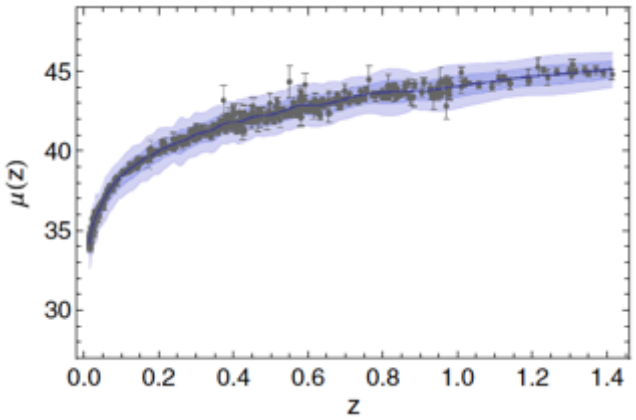

Figura 1b. Datos reales de supernovas que detallan la expansión acelerada del universo. Modelo teórico simulado con estadísticas bayesianas. (Copyright 2014. Phys.Rev.D89, no.4, 043007).

Un segundo ejemplo es en la ecohidrología, la ciencia que relaciona el comportamiento y propiedades del agua con las relaciones de los seres vivos entre sí y explica los complejos procesos del ciclo del agua. Los modelos matemáticos que usan datos numéricos de agua y vegetación han ayudado a conocer cuánta agua producen los bosques en diferentes contextos del país y a su vez han provocado tomas de decisiones políticas y económicas como son los programas de

conservación forestal y el desarrollo ganadero y agrícola. La Figura 2 es un claro caso de la disminución en producción de agua y tormentas tropicales en la Sierra Madre de Chiapas. Los ríos del Estado de Chiapas son las principales fuentes que aportan agua a México pero sus bosques han pasado de la deforestación, en el año 2005, a una conservación y reforestación, en el 2015. El Gobierno Federal ha implementado desde el 2007 programas forestales en donde relacionan que a mayor cantidad de bosques mayor producción de agua. Pero, realmente ¿estos programas han funcionado en el paso del tiempo?

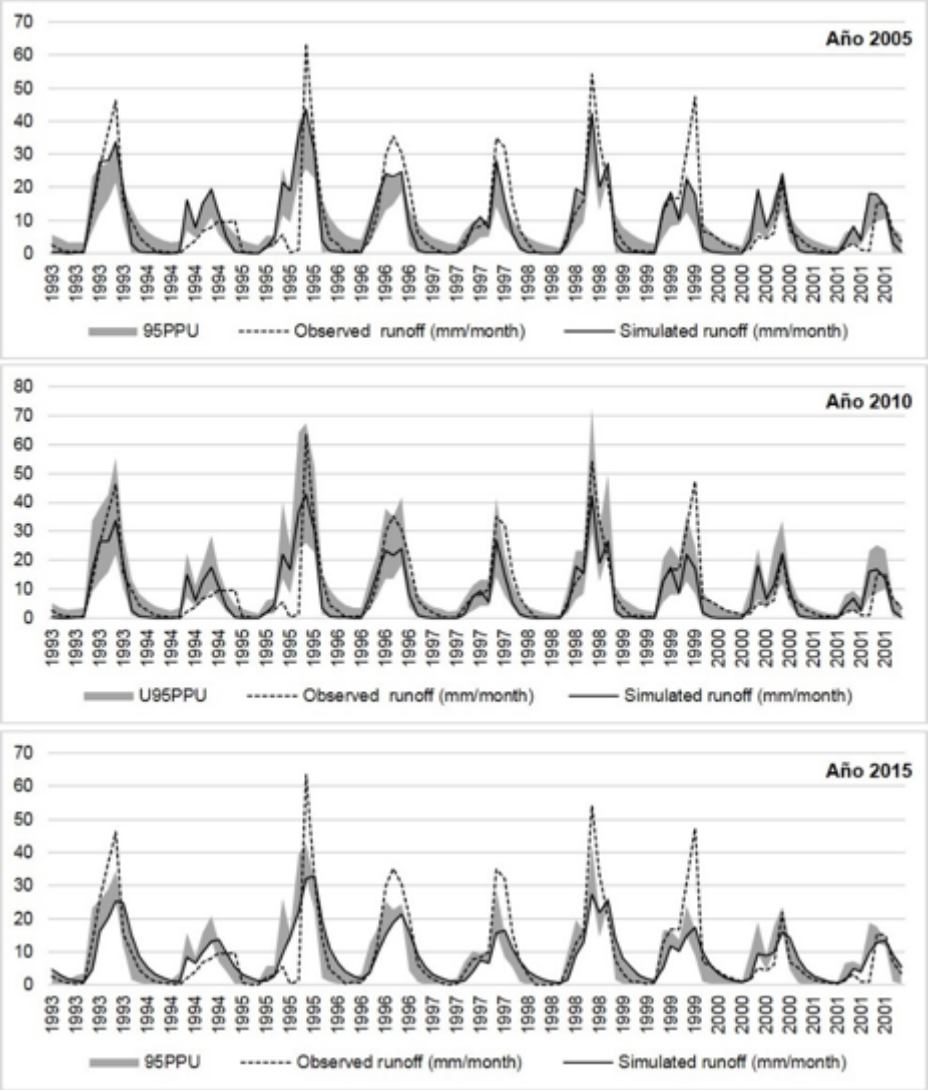

Figura 2. Comparación de datos de producción de agua simulados (simulated runoff) con datos de producción de agua observados en campo (observed runoff) medido en milímetros por mes, durante los años de cambios de vegetación (2005, 2010 y 2015), usando el software R y ArcSwat. (Copyright 2017. Doctoral Thesis Escamilla Rivera, V.)

Un tercer ejemplo nos lleva a la música. ¿Cómo colaboran los compositores de música clásica y se influyen mutuamente? La ciencia de datos ha sido capaz de mostrar cómo la cultura ha evolucionado e influido en el mercado de grabación. Un estudio publicado en la EPJ Data Science (2015) se basó en la mayor base de datos de grabaciones de música clásica hasta la fecha, utilizando el minorista en línea ArkivMusic y el sitio de referencia musical AllMusicGuide. Los autores se centraron en el análisis de redes de compositores contemporáneos en publicaciones de CD, utilizando modernas técnicas de análisis de datos. Primero determinaron cómo las propiedades fundamentales de la red de compositores occidentales de música clásica se correlacionan con los estilos artísticos y los períodos activos de su compositor. Encontraron que dicha red muestra la propiedad del pequeño mundo y una estructura modular. Luego observaron cómo una red de compositores clásicos se desarrolló con el tiempo. Específicamente, cómo los diferentes compositores son "escuchados juntos" por los consumidores de CD's de música clásica - un aspecto muy importante en los estudios culturales -. Luego demuestran cómo los consumidores se relacionan con diferentes compositores y estilos, lo que proporciona herramientas útiles para predecir el futuro paisaje del mercado de grabación clásico. Específicamente, encontraron que la red del compositor ha evolucionado concentrándose en compositores superiores mientras que su tamaño creció constantemente. En el futuro, es probable que el paisaje de la grabación musical se concentre en torno a unos pocos compositores con creciente prominencia y mayor diversidad gracias a un número creciente de compositores grabados.

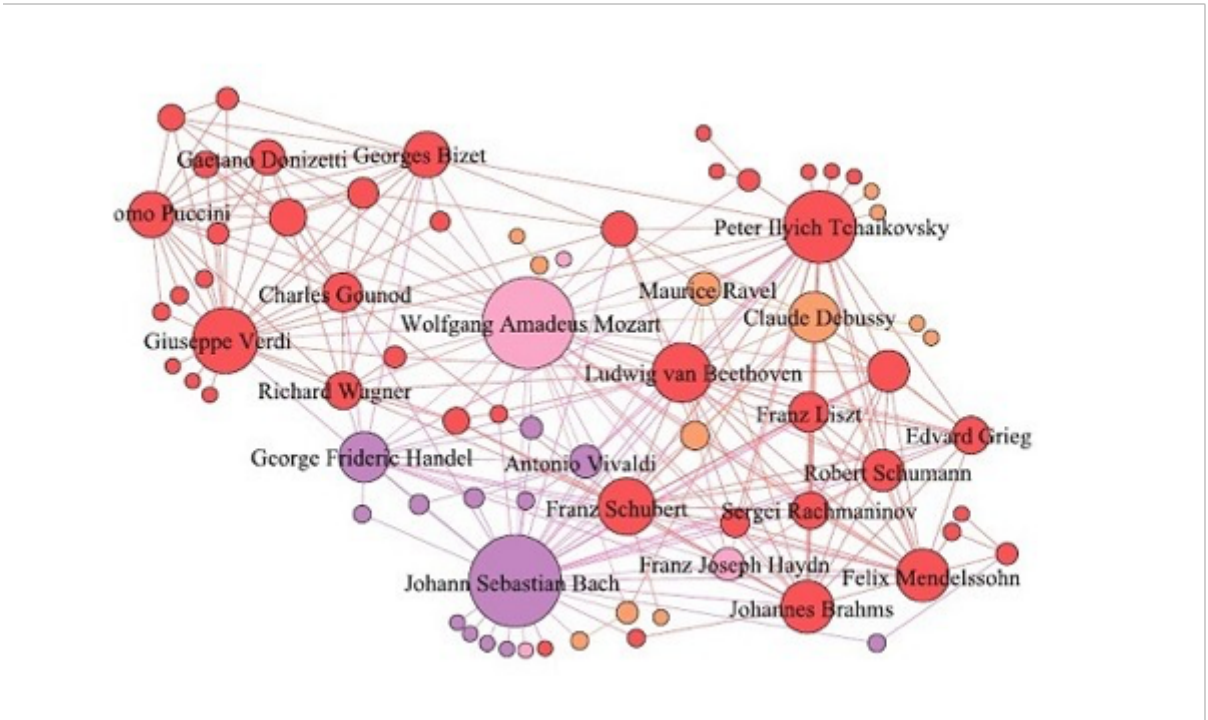

Figura 3. La columna vertebral de la red compositor-compositor, proyectado desde la red CD-composer, revela el principal componente de la red. Los tamaños de nodos representan los grados de los compositores, y los colores

representan sus períodos activos. (Copyright 2015. EPJ Data Science. Park, D. et al)

Ser un científico de datos

Según fuentes como El Financiero, esta profesión empezó a ser demandada desde el 2013 debido al gran avance de la tecnología. En esta profesión no se es un analista de datos convencional, que sólo suele mirar los datos obtenidos de una única fuente. Un científico de datos es parte analista, parte artista. Su trabajo consiste en obtener las respuestas para preguntas o problemas que se plantean en negocios, universidades, en áreas de investigación científica: física, matemáticas, energía, ecología, etc.

Entre las habilidades que requiere un científico de datos están:

- Saber extraer los datos independientemente de su fuente (como ya expusimos al principio, éstas pueden ser desde las proporcionadas por un experimento o inclusive de la observaciones de fenómenos en la naturaleza).
- Saber limpiar los datos para eliminar aquello que distorsiona los mismos.
- Dominar la técnicas de procesamiento de datos usando diferente métodos estadísticos (inferencias estadística, modelos de regresión, pruebas de hipótesis, etcétera).
- Diseñar nuevas pruebas o experimentos en caso de ser necesario.
- Visualizar y presentar gráficamente (cualitativa y cuantitativamente) los datos y resultados.

Figura 4. Las regiones "pictográficas" de las habilidades de un científico de datos.

Entonces, un científico de datos deberá ser capaz de tener sólidos conocimientos matemáticos, estadísticos e informáticos. Y lo mejor de todo, en tiempos actuales este científico es la persona más requerida en cualquier lugar del planeta.

¿Quieres conocer algunos códigos muy útiles?

Consulta las páginas siguientes:

Programa R. (http://www.r-project.org) Es un software para el análisis estadístico de datos. Posee un lenguaje de comandos propio que lo dota de un potencial singular (aunque hay interfaces gráficas que permiten la realizar la mayor parte de las tareas a golpe de ratón), y es, sobre todo, un proyecto de colaboración a nivel mundial.

Al descargarlo e instalarlo se tiene acceso al programa, y a algunas aplicaciones sencillas de estadística y gráficos.

Programa Python. (https://www.python.org) Es un lenguaje de escritura independiente de plataforma y orientado a objetos, preparado para realizar cualquier tipo de programa, desde aplicaciones Windows a servidores de red o incluso, páginas web. Es un lenguaje interpretado, lo que significa que no se necesita compilar el código fuente para poder ejecutarlo, lo que ofrece ventajas como la rapidez de desarrollo e inconvenientes como una menor velocidad.

Posee una sintaxis muy visual, gracias a una notación con márgenes de obligado cumplimiento. En muchos lenguajes, para separar porciones de código, se utilizan elementos como las llaves o las palabras clave "begin" y "end". Para separar las porciones de código en Python se debe tabular hacia dentro, colocando un margen al código que iría dentro de una función o un bucle. Esto ayuda a que todos los programadores adopten unas mismas notaciones y que los programas de cualquier persona tengan un aspecto muy similar.

Mathematica Wolfram. (https://www.wolfram.com/mathematica/trial/) Es un programa que permite hacer cálculos matemáticos complicados con gran rapidez. Es como una calculadora gigante a la que no sólo podemos pedirle que haga cálculos numéricos, sino que también derivadas, cálculo de primitivas, representación gráfica de curvas y superficies, etcétera.

El lenguaje de Mathematica Wolfram integra muchos aspectos del análisis de datos estadísticos, desde obtener y explorar datos hasta construir modelos de alta calidad y deducir las consecuencias. Ofrece varias maneras de obtener datos, comenzando con fuentes de datos integradas, importando desde una variedad de formatos de archivo o conectándose a bases de datos. El procesamiento básico, incluyendo el cálculo de cantidades estadísticas, suavizado, prueba y visualización, proporciona un primer nivel de análisis.

Es importante que antes de realizar cualquier prueba, con un programa informático, nos aseguremos de conocer en profundidad cómo se maneja el programa; este primer paso ha de considerarse básico para aplicar correctamente las pruebas. Existen muchos programas informáticos en el mercado, comerciales y de distribución libre. Su grado de bondad no depende esencialmente de su precio, con toda seguridad el mejor de todos es de distribución libre, aunque exige invertir mayor tiempo en su adiestramiento.

¿Qué se pretende con la ciencia de datos?

Todo lo que nos rodea contiene datos de todas clases y sabores, con lo que su eminente análisis es crucial. Estamos en una nueva era, la de estudiar datos con una alta precisión y exactitud -dos conceptos muy diferentes entre sí. En esta fase, la nueva generación de programas computacionales jugará un papel crucial en la ya conocida era del "big data", donde la gestión y análisis de enormes volúmenes de datos que no pueden ser tratados de manera convencional superan la capacidad del software habitual para ser manejados y gestionados.C²

Celia Escamilla-Rivera es Doctora en Ciencias por la Universidad del País Vasco. Actualmente es Profesora Investigadora en el Mesoamerican Centre for Theoretical Physics. Su investigación se centra en la interacción entre la Cosmología Teórica y Observacional. Ha realizado algunos postdoctorados en la University of Nottingham (Reino Unido) y en el Observatorio de la Universidad Federal do Espirito Santo (Brasil). Es miembro del Instituto Avanzado de Cosmología en México.

cescamilla@mctp.mx

Verenice I. Escamilla Rivera es estudiante de doctorado en el Colegio de la Frontera Sur, Chiapas. Su investigación se centra en el manejo de cuencas costeras y forestales en diferentes escalas. Combina el análisis geográfico, usando herramientas de Sistema de Información Geográfica y percepción remota, con herramientas de las ciencias sociales y de la estadística. Ha laborado en diversos proyectos para la Secretaria de Desarrollo Urbano y Medio Ambiente del Gobierno del Estado de Yucatán, el CINVESTAV- Unidad Mérida, Yucatán, y el IMPLAN- Ciudad del Carmen,

Campeche.

viescamilla@ecosur.edu.mx