Curso-Seminario
Ciencia de Datos Geoespaciales

Carga Horaria y modalidad del Curso:

Se prevé desarrollar un curso con un mínimo de tres encuentros presenciales, acceso a un sitio web sobre contenidos del seminario, desarrollo de contenidos teóricos y elaboración de ejercicios prácticos. El curso posibilitará a los participantes confeccionar mapas web con R Studio y el paquete leaflet, y elaborar reportes web con datos geográficos provenientes de datos abiertos, como asi tambien de geoservicios de nodos de Infraestructura de Datos Espaciales.

La propuesta tambien contempla la posibilidad de recibir con anticipación bases de datos específicas que sean útiles para el grupo con el fin de trabajar directamente sobre ellas. He recibido un conjunto de datos shp de censos que serán utilizados en el curso.

Fundamentación

La Ciencia de Datos es una disciplina emergente que permite analizar datos con el objetivo de extraer conocimiento de los mismos como así también producir un mejor entendimiento.

El objetivo de este seminario en conjunto con otros impartidos en la Carrera de Geografía es comprender los conceptos y herramientas más importantes en permitan abordar un proyecto típico en Ciencia de Datos.

En el presente siglo se han introducido diseños curriculares de diferentes carreras destinadas a formar competencias de un científico de datos, que permitan desarrollar nuevas capacidades para describir, explotar y visualizar información existente como así también producir información agregada, combinarla o incluso predecir nueva información. El auge de la ciencia de datos se ve potenciada con el crecimiento exponencial de la información, su tratamiento, y los paradigmas de datos abiertos e infraestructura de datos espaciales.

La visualización de datos es una habilidad fundamental para cualquier persona que utilice habitualmente datos cuantitativos y cualitativos en su trabajo, constituyendo una herramienta que casi todos los profesionales necesitan en la actualidad. Una de las herramientas críticas para la visualización y análisis de datos en la actualidad es el lenguaje de programación R. Este lenguaje permite instalar numeroso paquetes de software como ggplot, tidyverse, flexdashboard, shiny, leaflet etc., convirtiéndose en

una plataforma extremadamente poderosa y flexible para hacer figuras, tablas, mapas web, informes de ciencia de datos y tableros de control.

El presente seminario introduce a la problemática de análisis de datos permitiendo profundizar sus conocimientos en visualización de datos. Ilustra y ejemplifica herramientas para construir diferentes tipos de gráficos sobre estadística descriptiva. La visualización de datos geográficos a través de mapas web, es otro de los contenidos del seminario, a través del paquete leaflet de R. El uso de datos geográficos es imprescindible ya que se conoce que (aproximadamente) más del 80 por ciento de la información tiene una vinculación directa o indirecta con información geográfica.

Para el caso de informes, tanto el paquete knitr como R Markdown, proporcionan un marco flexible de creación de informes que permiten integrar conceptos relacionados de ciencia de datos, admitiendo docenas de formatos de salida estáticos y dinámicos, incrustando visualizaciones y tablas, y facilitando la integración de código Látex y de múltiples lenguajes de programación.

Objetivos generales del seminario:

El seminario tiene diversos objetivos para los alumnos:

—Analizar datos existentes (ejemplo datasets de datos abiertos u otros provenientes de la actividad de investigación) empleando técnicas y paquetes que favorezcan el tratamiento, combinación, visualización y estructuración de los mismos.

—Aprender a producir informes, mapas web, consumir datos de geoservicios y construir tableros de control a partir de los datos analizados.

—Conocer herramientas el proceso del desarrollo de un proyecto de ciencia de datos.

—Involucrar a alumnos en el uso de literatura actualizada y relevante a la temática.

—Conocer y explorar las principales herramientas del ecosistema de R y R Studio.

—Colaborar con los alumnos de la carrera en la producción de sus tesis/informes, como así también en su formato, sistematización, documentación y redacción.

Contenidos

MÓDULO 1: Introducción

Propedéuticos de Ciencia de Datos y Data Wrangling. Introducción y Objetos. Trabajando con Datos. Data Management con R. Entorno de Trabajo en R. Instalación de Paquetes, Funciones básicas de Cálculo. Scripts. Importar datos abiertos y geoservicios WFS de IDE. Importar csv, xls, shp, WFS de datos geográficos. Mostrarlos en un mapa con leaflet.

MÓDULO 2: Tratamiento de Datos Alfanuméricos y Gráficos

Operaciones sobre los datos. Calidad de Datos. Formatos. Exportarlos. Tratamiento del dato: Crear Subsets de Datos, Agregar campos calculados, combinar, mezclar y ordenar datos. Data wrangling. Paquetes. Vectores y Dataframe. Manipulando datos (filtrar, reordenar y agrupar), juntando datos (distintos joins), estructurando datos.

MÓDULO 3: Visualización con Gráficos

Visualización de Datos. Abordaje general sobre Estadística descriptiva. El paquete ggplot2.

MÓDULO 4: Reportes

Producción de informes con formatos de salida estáticos. Aplicación con R Markdown. Fragmentos de código. El paquete knitr. Tablas. Tipos de salida. Incluir tablas de contenido (toc). Formatos de Rmarkdown para generar reportes.

MÓDULO 5: Mapas Web y Tableros en la web

Información Geográfica y Mapas Datos Georreferenciados. Transformaciones de CRS (Sistemas de Refeencias de Coordendas). Tipos de geometría. Exploración de Datos Geográficos Visualización de Datos Geográficos. Combinación de fuentes de datos. Mapas con la librería leaflet. Tableros de control o comandos (dashboard) con flexdashboard.

Destinatarios

Estudiantes y graduados/as de la carrera de Geografía. Docentes e Investigados de la Universidad Nacional San Juan Bosco. Agentes invitados del sector público.

Propuesta Didáctica

Las actividades teóricas serán principalmente expositivas y prácticas. Todo concepto será ejemplificado con prácticas y ejercicios. En ellas también participarán los alumnos del seminario presentando aplicaciones o reportes sobre distintas temáticas (en los posible de aplicación en su tema de tesis/trabajo de intereés. Se fomentará el aprendizaje colaborativo y el trabajo en grupo.

Este curso seminario, se ha dictado con aterioridad en la Universidad del Centro de la Provincia de Buenos Aires UNCPBA (UNICEN), en la Universidad Nacional de San Juan UNSJ, en un taller de las XVI Jornadas IDERA en Córdoba, en taller de las XVII Jornadas IDERA de Santa Rosa, y en tres seminarios de maestría en la UNComahue, obteniendo muy buenos resultados en los alumnos en el desarrollo de capacidades para importar datos abiertos como de geoservicios, integrarlos y generar mapas en la web. Es una importante posibilidad práctica de aprender haciendo e incentivando a distintas disciplinas en la generación de mapas web. Al curso pueden participar: alumnos, docentes e investigadores de disciplinas relacionadas con la Información Geoespacial

(geografía, agrimensura, geologos, arquitectos, ingenieros forestales, etc) interesados en el tratamiento de la información geoespacial.

ACTIVIDADES PRÀCTICAS:
El seminario contiene mucha actividad práctica en lenguaje R que el alumno deberá realizar durante las clases y en forma autónoma fuera del horario de clase. Las actividades requerirán generar ploteos y mapas web, reportes y tableros utilizando el lenguaje R.

MATERIALES:

- El seminario dispondrá de una página web con apuntes sobre las unidades principales, actividades y enlaces a recursos en la web. Los alumnos deberán contar con notebook o en su defecto tener disponibilidad de un equipo PC con **un mínimo de 8 GB de RAM y espacio de disco suficiente para archivos de texto y archivos de imagen.** Privilegios de administrador para instalar y ejecutar utilidades de R-Studio.
- Deberán instalar en sus computadoras o notebooks:
    - 1. Instalar lenguaje R desde https://posit.co/download/rstudio-desktop/
    - 2. Instalar R Studio desde: https://posit.co/download/rstudio-desktop/

Condiciones de Acreditación

Para la acreditación de la materia se deberá asistir al 75% de las clases. Se tomará asistencia a las mismas. Realizar un trabajo integrador (simple) de un reporte o tablero en el cual se trabaje con un conjunto de datos geográficos locales, donde los participantes apliquen los conceptos desarrollados en el curso, pudiendo este trabajo ser entregado en grupos de dos o tres alumnos.

Bibliografía

Chambers, J. M. (2020). "S, R, and Data Science". The R Journal. 12 (1): 462–476. ISSN 2073-4859.

Chang, W. (2021) R Graphics Cookbook, 2nd edition, Disponible en: https://r-graphics.org/

Fox, J. ,Andersen, R. (2005). "Using the R Statistical Computing Environment to Teach Social Statistics Courses" (PDF). Department of Sociology, McMaster University. Retrieved 6 August 2018.

Li, J. (2021) The Exploration of the Approach to Data Preparation for Chinese Text Analysis Based on R Language. Open Access Library Journal Vol.8 No.9, September 3, 2021

Long, J. D., Teetor, P. (2021) R Cookbook, 2nd Edition, disponible en: https://rc2e.com/

Paradis, E..(2002) R para Principiantes. Institut des Science de l'Evolution. Universit Montpellier II. France. Traducción Jorge A. Ahumada. RCUCH University of Hawaii USGS/National Wildlife Health Center.

R Core Team (2017) R: A Language and Environment for Statistical Computing. https://www.R-project.org/

R Development Core Team (2021).disponible en https://cran.r-project.org/manuals.html

R Development Core Team (2021). R documentation, disponible en: https://www.r-project.org/other-docs.html
Vance, A. (2009). "Data Analysts Captivated by R's Power". New York Times.
Thieme, Nick (2018). "R generation". Significance. 15 (4): 14–19.
Tippmann, S. (2014). "Programming tools: Adventures with R". Nature News. 517 (7532): 109–110.