1. ***Mapas Coropléticos***

## *Introducción:*

El objetivo del presente documento es analizar una serie de metodologías médiate la utilización de Python y la biblioteca Pandas con su extensión GeoPandas para el análisis de datos espaciales, herramientas comúnmente utilizadas en “Ciencia de Datos”, para la confección de mapas coropléticos a partir de la utilización de datos de casos de covid19 en los partidos de la cuenca del río Luján.

En el análisis de la distribución espacial de los fenómenos sobre la superficie terrestre los mapas coropléticos desempeñan un papel destacado en la ciencia de datos espaciales\*. En la actualidad con la pandemia del COVID19 estos han adquirido un gran protagonismo en los medios de comunicación, tal y como se vio con la publicación del mapa de la distribución espacial de casos de COVID-19 a escala mundial realizada por la universidad Johns Hopkins (https://coronavirus.jhu.edu/map.html).

La palabra coropleta proviene de la raíz “coro” que significa área o región. Un mapa coropético presenta datos en cada región que se encuentra asociada a la intensidad en una rampa de colores, cada observación se encuentra vinculada un valor en una tabla de atributo (el valor de la variable a representar) y una geometría por lo general un polígono, las unidades espaciales.

Mostrar valores cuantitativos en mapas coropléticos envuelve la tarea de representar los datos de manera correcta y al mismo tiempo poder ordenar los mismos de forma que podamos detectar patrones en ellos. Las clase van a agrupar las entidades con valores similares, estas se van a representar con un color. Los límites de clases, van a definir como las entidades quedaran agrupadas, según como establezcamos estos límites, vamos a obtener mapas distintos. La efectividad de un mapa coroplético dependerá de los esquemas de clasificación junto a las estrategias de color y simbología a adoptar. En términos generales los esquemas de clasificación define el número de clases, así como las reglas de asignación, mientras la forma de simbolizar debería transmitir información sobre la diferenciación de valor entre las clases.

Existen dos criticas básicas que se le puede atribuir a este tipo de representación según Buzai y Baxandale, (2012). El primero de ellos radica en el tamaño relativo o absoluto de las unidades espaciales, ya que cada valor se representa de forma homogénea en su interior cuanto mayor superficie ocupa más alejado estará de la realidad. El segundo aspecto a resolver es la cantidad de intervalos de clase (K) a utilizar pues aquí se producen ciertos inconvenientes entre el máximo de categorías y el nivel de lectura, debido a que la mayor cantidad de categorías brindaría la mejor variación en la distribución espacial, sin embargo esto afectaría la comprensión del mapa.

* 1. ***Software y datasets***

Python es un lenguaje de programación de alto nivel utilizado en ciencia de datos. Python es un lenguaje que se caracteriza por su facilidad al implementarlo y de alta performance. Este lenguaje es ampliamente utilizado en todos los ámbitos de la ciencia, la NASA en el control de satélites hasta en muchos ámbitos de la industria en el control de procesos. Existen comunidades científicas en todo el mundo para respaldar y mejorar este lenguaje a través de la biblioteca Scintific Python scipy y la organización PyData. Dentro del dominio geográfico es el lenguaje utilizado para la creación de scripts en sistemas de información geográfica como QGIS Y ArcGIS.

Python se compila sobre la marcha sin necesidad de compilarlo. Lo que lo diferencia de otro tipo de lenguaje, que requieren un paso adicional no interpretativo en el que el programa se convierte en un archivo binario, que luego se ejecuta. En su lugar podemos escribir simplemente el código, ejecutarlo, solucionar cualquier problema y volver a ejecutarlo. Esto convierte a Python en una herramienta realmente muy buena y productiva para la ciencia, ya que podemos crear código prototipo, modificarlo, mejorarlo y hasta reutilizar.

El poder de Python como herramienta de análisis de datos proviene realmente de paquetes adicionales, código adicional llamadas bibliotecas que agrega funcionalidad al lenguaje en sí. Python posee un amplio ecosistema de bibliotecas para ciencia de datos y en especial para el análisis espacial, siendo la principal GeoPandas. La carga de paquetes Python se denomina importación. GeoPandas amplia a la biblioteca Pandas de ciencia de datos al agregar soporte para datos geoespaciales.

La estructura de datos central en GeoPandas es geopandas.GeoDataFrame una subclase pandas.DataFrame capas almacenar columnas geométricas y realizar operaciones espaciales. Un archivo que contiene datos como geometrías con extensión GeoPackage, GeoJSON o Shapefile puede ser leído por esta biblioteca.

* 1. ***Preguntas a respondes***

* 1. ***Clasificación de los datos***

En la elección de los esquemas de clasificación se debe tener en cuenta la distribución estadística de los datos. Para ver esta cuestión se analizaran los datos de muertes por COVID-19 en los 16 partidos que integran la Cuenca del río Luján. La variable que utilizaremos será la tasa de mortalidad (Tasa\_Mortalidad).

Como se puede observar, la distribución esta algo sesgado hacia la derecha, es decir esta sesgada positivamente. En otras palabras, la media supera a la mediana (50%, como se puede ver en la siguiente tabla).

El método seaborn.distplot usa la función del histograma de matplotlib para determinar los limites de clase y el conteo de las observaciones en cada clase. En nuestro histograma tenemos cinco clases.

El objeto conteo entrega cuantas observaciones tiene cada categoría en la clasificación:

El objeto bin almacena estos puntos de ruptura que nos interesa cuando consideramos esquemas de clasificación.

Podemos observar cinco conteos, el conteo que tiene el mayor número de observaciones comprende como límite inferior de 181,4 y un límite superior 210,8 con 5 observaciones. Para la determinación del ancho del intervalo y el número de conteos seaborn se basa en la regla de Freedman Diaconis.

* 1. ***Métodos para la determinación de intervalos de clase***

***1.5.1 Mapa de cortes naturales***

Es una representación que se basa en la estructura que forman los datos en la búsqueda de regularidades interna. Se debe encontrar los saltos naturales que pueden ser tomados como limites de los intervalos.

Cuando el procedimiento se realiza de forma visual la cantidad de intervalos puede resultar evidente, pero si el objetivo es realizar una serie cartográfica para diferentes variables en una misma área de estudio esta metodología puede dar como resultado mapas con diferentes cantidades de intervalos. La desventaja es que no permite hacer comparaciones, espaciales y/o temporales, dado que la clasificación se basa en el comportamiento intrínseco de los datos.

*Geographic Data Science or Spatial Data Science?*

En muchos ámbitos académicos la ciencia de datos ha impactado de lleno, la informática y los nuevos algoritmos de procesamiento de Machine learning y Deep Learnig entraron en las ciencias sociales y la geografía no es una acepción. A que nos referimos realmente con ciencia de datos,

## *\*What Is Spatial Data Science?*

"Spatial data science can be viewed as a subset of generic “data science” that focuses on the special characteristics of spatial data, i.e., the importance of “where.” Data science is often referred to as the science of extracting meaningful information from data. In this context, it is useful to stress the difference between standard (i.e., non-spatial) data science applied to spatial data on the one hand and spatial data science on the other. The former treats spatial information, such as the latitude and longitude of data points as simply an additional variable, but otherwise does not adjust analytical methods or software tools. In contrast, “true” spatial data science treats location, distance, and spatial interaction as core aspects of the data and employs specialized methods and software to store, retrieve, explore, analyze, visualize and learn from such data. In this sense, spatial data science relates to data science as spatial statistics to statistics, spatial databases to databases, and geocomputation to computation." (Luc Anselin, 2019, "Spatial Data Science" in *The International Encyclopedia of Geography: People, the Earth, Environment, and Technology*.)