PROPUESTA ANÁLISIS EXPLORATORIO DE DATOS GEORREFERENCIABLES Y TEMPORALES, CASO DE ESTUDIO: PRODUCCIÓN NACIONAL DEL ORO 2012 A 2020.

APPROACH EXPLORATORY ANALYSIS OF GEOREFERENTIAL AND TEMPORARY DATA, CASE STUDY: GOLD NATIONAL PRODUCTION 2012 TO 2020.

Horacio Miguel Marrugo Petro
hmmarrugop@libertadores.edu.co
Jorge Andrés Melo Mayorga
jamelom01@libertadores.edu.co
Fundación Universitaria Los Libertadores

RESUMEN

El escrito plantea una ruta en la cual el lector pueda realizar análisis previos, en datos que cumplan con las condiciones de georreferenciación y temporalidad, por ello, se indaga sobre la base teórica del: análisis exploratorio de datos creada por Jhon W. Tukey, metodología Box Jenkins, normatividad y manuales referidos a uso del marco Geoestadistico Nacional (MGN), datos abiertos referentes a producción nacional del oro de minerales publicados por la agencia nacional de minería. Como segundo pilar del proyecto se utiliza el lenguaje R implementado en el software libre Rstudio y QGis, y así, generar una serie de script que permitan realizar la medición y descripción estadística de datos, seguido de su comparación entre los diferentes años, es así qué, utilizando herramientas informáticas se busca una mayor agilidad, practicidad y confiabilidad para el analista de datos. Finalmente se describen algunas consideraciones sobre el procedimiento, los datos analizados y las perspectivas sobre análisis futuros.

Palabras clave: Análisis Exploratorio, Datos Abiertos, Georreferenciación, Minería, Oro, Series de tiempo.

ABSTRACT

The text proposes a route that the reader can realize previous analyzes on data that comply with temporality and georeferential conditions, that is why it investigate about the theory base of analysis: exploratory analysis of data created by Jhon W. Turkey, methodology box Jenkins, normativity and manuals referrals a used of national geostatic framework (MGN in Spanish) open data referring to national production of gold or minerals published by the national mining agency. As the second pillar or project the language R is implemented in the free software Rstudio y QGis, and so generate a set of script that allow realize the measurement and statical description of dates, and its comparison between the different years, that is why using the informatics tools for create more agility, phaticity and reliability for the analysis of data. Finally, the text describes some considerations about the process of data, analyzing and perspectives about future analysis.

Keywords: Exploratory Analysis, Open data, Georeferencing, mining, Gold, Time Series.

INTRODUCCIÓN

"La efectiva reutilización de Datos Abiertos puede generar beneficios desde las perspectivas económicas y sociales" (MINTIC, 2019); es por ello la propuesta se fundamenta como una herramienta para que el usuario de los datos abiertos mejore o incorpore en los procesos cartográficos el análisis exploratorio en datos georrefenciados o susceptibles a serlo, a través de la incorporación de técnicas estadísticas, resaltando que la ubicación de la información proporciona elementos adicionales al análisis, base determinante para el ordenamiento territorial, geomarketing, administración de recursos naturales, etc.

La propuesta es de análisis exploratorio de datos georreferenciables y temporales, empleara software gis y estadístico y para ello utilizara el caso de estudio de la producción nacional de minerales Oro 2012 a 2020, proponiendo tres objetivos: Primero, identificar las teorías del

análisis exploratorio de datos temporales y el fundamento normativo y practico en la utilización del marco geoestadistico nacional emitido por el DANE como herramienta vinculante en la reutilización de datos abiertos dispuestos en el portal colombiano. Segundo, Analizar espacialmente la información para una mayor comprensión de los datos. Tercero, utilizar métodos estadísticos mediante herramientas de programación en software libre, actividad para su incorporación semi automatizada de la propuesta, en los procesos geográficos/cartográficos, verificando si existen anomalías en su estructura y valores atípicos (Generalmente no son significativos y que tienden a distorsionar el comportamiento, analizar variables en simultánea, identificando la incidencia entre las mismas; describir el comportamiento en el tiempo. Cuarto, aplicar la propuesta con datos abiertos de la página web datos.gov.co, publicada por la Agencia Nacional de Minería (Grupo de Regalías y Contraprestaciones Económicas), teniendo como titulo Producción Nacional de Minerales y Contraprestaciones Económicas Trimestral, información detallada referida la cantidad de mineral extraído en el territorio nacional, asociando a la contraprestación económica [1] generada por municipio productor desde la vigencia 2012 a 2012-2, filtrada por mineral Oro. (Agencia Nacional de Mineria ANM, 2020) Siendo esta base de datos un ejemplo el cual puede ser sustituido por el lector para la implementación de la propuesta exploratoria de datos, siempre en cuanto esta pueda relacionarse geográficamente a través de algún atributo dela misma (código departamento, municipio, topónimo, zona, etc.) y presente datos de temporalidad. razón por la se abarcarán los siguientes temas, que mediante el engranaje y sinergia entre los mismos se garantizara un efectivo análisis exploratorio.

Ilustración 1 - Pilares del Análisis.

Fuente: elaboración propia

Es así como, la propuesta pretende mitigar las deficiencias y/o inexistencia de análisis exploratorios en datos abiertos geográficos o susceptibles a ser georreferenciados y con

[1] La compensación es una contraprestación económica que se fija contractualmente y que no corresponde a una producción de mineral.

información temporal, esto debido a la baja incorporación de técnicas estadísticas como la identificación de medidas de tendencia central, correlación y temporalidad de los datos para su tratamiento, por los profesionales de ciencias de la tierra; evidenciados en la inexistencia de productos cartográficos que reflejen su implementación. Son muchas las ventajas que trae su incorporación al procedimiento, entre las que se encuentran: a) La necesidad de identificar casos anómalos en las bases de datos geográficos que afectan la calidad de la publicación cartográfica; b) el mejoramiento de clasificación para que sea concordante con las medidas de tendencia central y/o distribución espacial; c) Un manejo eficiente de la información temporal de la bases de datos geográfica, todo ello vinculado con la incorporación de algunas funciones del software estadístico en los sistemas de información geográfica y acompañado de una sensibilización al analista geográfico sobre la facilidad y beneficios de la propuesta en el quehacer geocientífico, ya que somos unos convencidos, que es a través de la sinergia entre ambas diciplinas se lograra el mejoramiento de los procesos, procedimientos y por consiguiente nuevas posibilidades en el uso de la información georreferenciable. (Talaya, 2018, pág. 7)

La importancia de la implementación de dichas técnicas estadística por los usuarios de los datos abiertos, radica en la necesidad y esfuerzos del gobierno nacional para disponerlos, prueba de esto es la adopción en 2016 de la carta internacional de datos abiertos como instrumento orientador en la generación y uso de datos en Colombia, reconociendo así que las entidades estatales poseen grandes cantidades de información que pueden ser de interés para los ciudadanos, en consecuencia se acordó seguir los principios que sientan las bases en el acceso de los datos para su publicación y uso, promoviendo: 1) Datos abiertos por defecto, 2) Comparables e interoperables, 3) Oportunos y completos, 4) Mejorar la gobernanza y participación ciudadana, 5) Sean accesibles y utilizables, 6) Apoyar el desarrollo incluyente y la innovación (MINTIC, 2019, pág. 7).

La incorporación de la estadística en el análisis geográfico no es nuevo, el Instituto Nacional de Estadística de España utiliza este tipo de análisis conjugado, evidenciado en el artículo titulado “Si en el momento de la creación de las instituciones ya la estadística y la geografía estaban íntimamente relacionadas, parece que esta relación aumenta día a día”, presentando un ejemplo sobre su utilidad y resaltando que la publicación en tablas y gráficos puede ser poco atractivas para el usuario no avanzado, pero si la publicación se encuentra acompañada

de técnicas cartográficas, se podrían representar a través de mapas, por lo que bastaría un simple vistazo para que el usuario final podría entender fácilmente los resultados. (Maldonado Cecilia, 2020) En el caso colombiano el Departamento Administrativo Nacional de Estadística establece a través de la resolución 2222 de 2018 el Marco Geoestadístico Nacional, siendo el sistema para referenciar la información estadística a su localización geográfica, siendo el resultado de comprender que el análisis del espacio y la localización, han sido variables inquietantes para la toma de decisiones en los sectores públicos y privados, y la puesta en marcha de planes de manejo territorial (Corso Sicilia & Pinilla Rivera, 2017, pág. 93), entre sus publicaciones se resalta el Atlas Estadístico de Colombia el cual brinda a través de la cartografía temática, los textos, la semiología gráfica y las ayudas visuales complementarias, un perfil de las principales configuraciones poblacionales y territoriales de Colombia sobre aspectos demográficos, sociales y económicos. (Departamento Administrativo Nacional de Estadística DANE, 2020)Es por ello que al querer incorporar técnicas de estadísticas a datos georreferenciables para profesionales no estadísticos, se despliega una oportunidad de formular una propuesta de análisis exploratorio, que permia presentar una identificación descriptiva organizada y teóricamente soportada, buscando obtener como resultado un flujo de trabajo a seguir, que conlleve al mejoramiento del análisis y por consiguiente de la calidad del producto geográfico.

En el análisis exploratorio de datos se utiliza un resumen numérico y visual en búsqueda de la medición ,descripción y comparación y así explorar los datos en busca de patrones no anticipados; autores clásicos como John Tukey, Frederick Hartwig y Brian Dearing lo catalogan como un “estado mental” ante el conocimiento. Es así como el primero lo define como una actitud, una flexibilidad y “algunas hojas con gráficos” (o transparencias, o ambos). Esto último como un reconocimiento de que el ojo que mira al horizonte es el mejor instrumento que tenemos para observar, de manera completa, lo no Anticipado. Por su parte, Hartwig y Dearing argumentan que el investigador debe aprender todo lo posible acerca de una variable o conjunto de variables antes de utilizar los datos para probar hipótesis o teorías acerca de las relaciones sociales. Más recientemente, Eugene Horber y Dominique Ladiray plantean que el “razonamiento” exploratorio es un esquema de análisis que enriquece las posibilidades del investigador para hallar nuevas respuestas a los problemas que se plantee. (Parra Olivares, 2002)

REFERENTES TEORICOS

Antecedentes

A continuación, se describe las investigaciones relacionadas al proyecto, base teórica y conceptual del proyecto, entre las que encontramos:

a) "Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas, Universidad del Bosque, Año 2017", se expone de manera sintetizada la exploración de datos espaciales, analizando técnicas gráficas y estadísticas utilizando el software GeoDa, analizando datos univariantes y multivariantes, identificando características de distribución y agrupamiento espacial. Necesidad la cual surge por la necesidad de identificar técnicas que permitan una coherente descripción y visualización de distribuciones espaciales que den valides al modelo, obtenido como resultado mapas que representen: tendencia central, cuantiles, histograma de frecuencias, desviación típica, valores atípicos, diagrama de caja y bigotes, cartograma, diagramas de dispersión espacio temporal, diagrama de coordenadas paralelas, matrices de interacciones espaciales, test I de Moran, contraste de dependencia espacial local univariante, Mapas LISA Indicadores de Asociación Espacial. (Corso Sicilia & Pinilla Rivera, 2017).

b) En el "Análisis exploratorio de datos espaciales como herramienta de estudios geográficos", (Meza & Arias, 2018), tuvo como fin el identificar las potencialidades que tienen los análisis exploratorios de datos espaciales y advertir la utilidad de diferentes técnicas estadísticas para la obtención de resultados confiables en la investigación geográfica, aplicando técnicas estadísticas descriptivas (medidas de tendencia central, medidas de posición no centrales y las medidas de dispersión), medidas de distribución espacial (centro medio, distancia estándar y elipse de desviación estándar), asociación entre variables y autocorrelación espacial, a partir de un software SIG; destacando que los análisis exploratorios de datos espaciales permiten encontrar patrones de comportamientos o anomalías en la estructura de los datos, tener una visión general de la localización de las variables y comprobar los supuestos necesarios para la aplicación de test estadísticos; y desde el punto de vista geográfico este análisis no es

suficiente, porque se lo debe contextualizar con información del espacio en estudio para así lograr una mejor interpretación de la realidad y hacer más factible la toma de decisiones y/o intervenciones.

El análisis de datos que conducen a la confirmación de teorías o hipótesis, en contraste con un reducido interés en las herramientas estadísticas que orientan a la exploración de datos, caracterizado por el uso de herramientas o técnicas con mucha carga visual o gráfica, con énfasis en revelar información vital sobre la data examinada. El arsenal correspondiente está compuesto, entre otros, por instrumentos como: Diagrama de caja y bigotes, Diagrama de tallo y hojas, Diagrama de dispersión; el análisis exploratorio de los datos requiere una actitud dispuesta y paciente para el "rastreo" del comportamiento de las variables. El análisis confirmatorio utiliza estadísticos numéricos de resumen generados a partir del empleo de un modelo, definido a priori, para confirmar o no una hipótesis, se caracteriza por el empleo de indicadores como la media, la varianza y los coeficientes de correlación y regresión, así como las pruebas de hipótesis. (Parra Olivares, 2002)

C) En el proyecto "Lattice Data: Plugin de QGIS que implementa análisis estadístico exploratorio de datos lattice para la identificación de correlación espacial", realiza un análisis exploratorio de datos espaciales (AEDE) de tipo polígono, comúnmente denominados lattice, para estimar, visualizar e interpretar de manera sencilla el grado de autocorrelación espacial (AE) que estos datos presentan en el área de estudio comúnmente analizada mediante herramientas de sistemas de información geográfica. Para la determinación de la AE se usó el índice estadístico I. de Morán Global, a partir de un análisis de vecindad o contigüidad, el cual ofrece medidas resumen que indican la intensidad y el tipo de la relación espacial presente en los datos. (Castillo Giraldo, Rodríguez Álvarez, & Carrillo García, 2015)

Datos Abiertos en Colombia

El (MINTIC, 2019) define a los datos abiertos como información pública dispuesta en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos en el numeral sexto como "todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones. En el Ciclo de Datos Abiertos se presenta en cuatro momentos que a la vez se desagrega en 3 actividades, como los son: 1) Establecer el plan de apertura para identificar, Analizar, Priorizar; 2) Estructurar y publicar datos: documentar, estructurar y publicar; 3) comunicar y promover el uso de los datos: Consolidar y posicionar, vincular actores, dar a conocer; 4) Monitorear la calidad y el uso: Medir impacto, usos y calidad, dispuestos en formatos que permiten su uso, reutilización y aprovechamiento sin restricciones legales y bajo licencia abierta, algunos de los formatos de datos abiertos más comunes son: CSV, XML, RDF, RSS, JSON, ODF, WMS, WFS. Base de este esfuerzo gubernamental es el establecido en La Constitución Política de Colombia en sus articulo 20 y 74 establece el accesos a información pública como derecho fundamental, más tarde en la ley 1712 de 2014 establece el procedimiento para garantizar este derecho instando así a las entidades a liberar o publicar sus datos, adicionalmente en los decretos 1078 y 1081 de 2015 se desarrolla la política de gobierno digital con el propósito de toma de decisiones; en la resolución 3564 de 2015 el Min Tic incluye las condiciones técnicas de apertura de datos abiertos; el gobierno nacional aprobó el documento Conpes 3920 de 2018 Política Nacional de Explotación de datos (Big Data) y con ello garantiza y aumenta el aprovechamiento de datos públicos para generar valor social y económico al país.

Marco Geoestadístico Nacional

El MGN está constituido por áreas geoestadísticas (departamentos, municipios, cabeceras municipales, centros poblados, rural disperso, entre otras), delimitadas principalmente por

accidentes naturales y culturales, y que son identificables en terreno. Comprende: 1101 municipios del país, 20 áreas no municipalizadas y la isla de San Andrés, los límites de los departamentos y municipios que conforman la vigencia del MGN se basan en los límites oficiales suministrados por el IGAC. El código de la Divipola está compuesto por 22 posiciones permite relacionar fácilmente los niveles del MGN con los datos estadísticos; es jerárquico en la medida que parte desde el nivel más amplio (código del departamento) hasta el nivel más detallado (código de la manzana censal en el área urbana y código de la sección rural en el área rural del municipio). Entonces la posición 1 y 2 se refiere al Departamento, Posición 3 a 5: Municipio, posición 6: Clase (Urbano = 1, Centro poblado = 2 y Área rural dispersa= 3), posición 7 a 9: Sector rural, posición 10 y 11: Sección rural, posición 12 a 14: Código del centro poblado, posición 15 a 18: Sector urbano, Posición 19 y 20: Sección urbana y Posición 21 y 22: Manzana censal. Los niveles geográficos del MGN se pueden integrar con otros niveles geográficos como: uso del suelo, catastro, divisiones administrativas, amenazas naturales, distritos de salud, entre otras; también con variables ambientales, sociales y económicas, tal que se cuente con las variables requeridas para el diseño de una operación estadística (DANE, 2018)

Métodos Estadísticos para Análisis de Datos Exploratorios

El análisis exploratorio de datos espaciales (AEDE), se define como el grupo de técnicas que describen y visualizan las distribuciones espaciales, identificando localizaciones atípicas, descubriendo esquemas de asociación (autocorrelación espacial) y sugiriendo estructuras en el espacio geográfico (heterogeneidad espacial); por consiguiente es más una técnica descriptiva que confirmatoria (Corso Sicilia & Pinilla Rivera, 2017, pág. 93). El estadístico y matemático norteamericano Jhon W. Tukey (1915-2000), fue el creador y figura central en la aplicación de la metodología del Análisis Exploratorio de Datos en su clásico trabajo de 1977 “Exploratory Data Analysis”. El análisis exploratorio de datos utiliza el resumen numérico y visual para explorar datos en busca de patrones no anticipados y así identificar las variables del dato para su posterior utilización, los instrumentos utilizados para el fin son: a) Diagrama de Caja y Bigotes el cual permite identifica si uno o ambos extremos contiene valores inusualmente grandes o pequeños; b) Diagrama de tallos y hojas se asemeja a un

histograma; c) diagrama de dispersión sirve para identificar la relación (dirección, fuerza y forma) entre las variables. (Parra Olivares, 2002).

A continuación, se describirá los gráficos mencionados anteriormente de acuerdo a las definiciones propuestas por (Pérez Medinilla & Crespo Borges, 2018)

- El gráfico de tallo y hojas, ideado por Tukey (1972, 1977) y con un precedente en Dudley (1946), es una especie de híbrido entre histograma y tabla de distribución de frecuencias en el que las líneas o barras se construyen con los propios datos. Frente al histograma presenta la ventaja de que los datos originales no se pierden.
- El gráfico de hojas y tallos es adecuado para muestras de cualquier tamaño, es un método muy flexible, se puede elaborar de diferentes formas que posibilitan que el investigador adapte el resultado a sus intereses de información, posibilita observar características de la distribución como: forma de la simetría de la distribución, la dispersión que presentan los datos del conjunto, la presencia y cantidad de valores extremos o atípicos, la concentración de datos en determinados puntos de la distribución, y la existencia y situación de agujeros en el conjunto de datos.
- El Boxplot comunica información acerca de cinco características de la distribución de un grupo de datos: localización del valor central (media o mediana), dispersión central de los valores (distancia entre el cuarto superior e inferior), simetría de la distribución, longitud de las colas y valores extremos o fuera de ámbito, la longitud de la caja, es decir, la diferencia entre los dos cuartos permite detectar posibles asimetrías de la masa central de los datos.
- El gráfico de dispersión muestra la relación entre variables, es útil para examinar la dirección, fuerza y forma de la relación, estos diagramas deben usarse cuando tenemos un análisis estadístico bivariable, o sea, una tabla de datos de doble entrada, la ventaja que tiene es que se puede graficar de una forma sencilla una distribución bivariante conjunta, permite analizar proximidades entre individuos y/o poblaciones y localizar outliers. Otra forma de representar relaciones entre más de dos variables es dibujando pares de variables mediante diagramas de dispersión, que son ordenados en una matriz de diagramas de dispersión.

El enfoque Box-Jenkins para el análisis de los datos y la identificación de un modelo apropiado, propone las siguientes etapas: selección de datos, análisis de datos, estabilización de varianza y desestacionalizad, selección preliminar del modelo, resultados, análisis gráfico, identificación del modelo, diagnostico, selección del modelo y estimación de los parámetros, validación. (Jaramillo Ayerbea, González Gómeza, Núñez Cabreraa, & Lucio García, 2007)

Producción Nacional de Minerales y Contraprestaciones Económicas

La minería ha sido una actividad económica central en Colombia desde la época Precolombina, en un comienzo, la actividad minera dio origen al comercio regional caracterizado por el trueque de varios minerales, posteriormente durante la época de la Colonia la minería creció en grandes proporciones, para el periodo de la República, la actividad minera, representada casi en su totalidad por la explotación de oro y piedras preciosas, ya gozaba de una posición aventajada frente a otros sectores básicos como la agricultura. La evolución favorable de la minería en el pasado ha llevado a que la explotación, la producción y la exportación de oro hayan sido catalogadas como las actividades económicas más antiguas y unas de las de mayor importancia para el país, durante buena parte del siglo XIX la exportación de este metal, acompañada de las de la plata y el platino, permitieron equilibrar la balanza comercial y se convirtieron en una importante fuente de atracción de inversión extranjera y hasta los últimos años de este siglo, los metales preciosos permanecieron como los únicos productos significativos de la minería en colombiana. (Fedesarrollo, Cárdenas, & Reina, 2008, págs. 23-24)

La dinámica de crecimiento de la actividad minero-energética en Colombia representa una oportunidad importante para impulsar el crecimiento de la economía y la generación de capacidades sostenibles en otros sectores. Si bien el Gobierno Nacional promueve la actividad minera, también reconoce la necesidad de hacer una explotación sostenible de los recursos naturales y para lograrlo tiene estrategias concretas plasmadas en el Plan Nacional de Desarrollo. (Estupiñán Vargas & Polanía, 2011)

El oro ha desempeñado un rol muy importante en la economía colombiana desde la época colonial, sin embargo, históricamente el valor de la producción y las exportaciones se ha

caracterizado por presentar un comportamiento fluctuante explicado básicamente por una fuerte correlación del precio doméstico con el precio internacional; Antioquia es el mayor productor de oro en Colombia (con aproximadamente el 68% de la producción nacional en 2006) y además es el Departamento que cuenta con el mayor grado de tecnificación para la extracción del mineral; le siguen en importancia Chocó, Bolívar y en menor medida Cauca, Tolima, Santander, Nariño, Valle del Cauca y Risaralda (Fedesarrollo, Cárdenas, & Reina, 2008, págs. 23-24).

Las regalías constituyen una de las contribuciones más importantes de las minerías a las finanzas públicas, especialmente en la medida en que representan un beneficio económico fundamental para algunos departamentos y municipios. En Colombia la ley 2056 de 2020 es el marco normativo el cual regula la organización y el funcionamiento del sistema general de regalías.

METODOLOGÍA

La propuesta es de análisis exploratorio de datos georreferenciables y temporales, empleara software gis y estadístico y para ello utilizara el caso de estudio de la producción nacional de minerales Oro 2012 a 2020 por medio de series de tiempo univariante aplicando metodología de Box-Jenkins (Jaramillo Ayerbea, González Gómeza, Núñez Cabreraa, & Lucio García, 2007);) (Pérez Medinilla & Crespo Borges, 2018), (Corso Sicilia & Pinilla Rivera, 2017); Las fases de las metodológicas propuestas son :

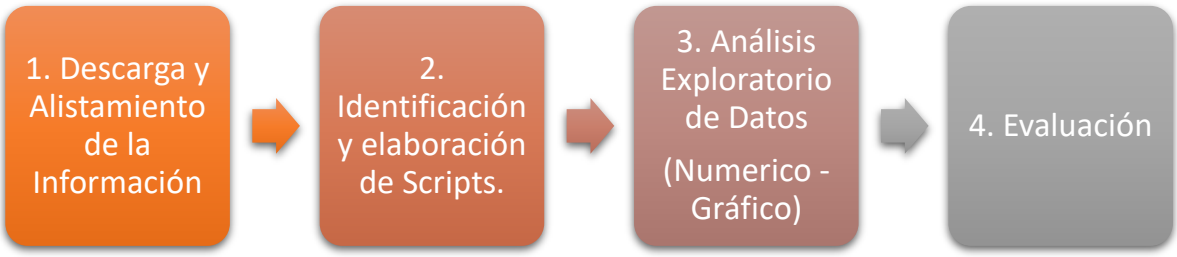

Ilustración 2 – Fases metodológicas.

Fuente: elaboración propia

1. ***Fase 1 - Descarga y Alistamiento***

Etapa en la cual tiene como propósito: a) Descargar la información alfanumérica en la pagina datos abiertos.gov, para este caso corresponde a la Producción Nacional de Minerales y Contraprestaciones Económicas Trimestral, publicado por la Agencia Nacional Minera (ANM)[2]; b) descarga del marco geoestadistico Nacional que para este caso corresponde a los limites municipales de Colombia[3]; c) Proyectar la información geográfica, despliegue Alistamiento de datos el cual corresponde en desplegar la información alfanumérica y organizarla de manera tal que corresponda a la estructura geográfica, correspondiente a un registro (fila) por cada municipio. C) Realizar la unión para incorporar la base alfanumérica de la Agencia Nacional Minera con los datos geográficos del MGN.

2. ***Fase 2 - Identificación y elaboración de script***

Fase en la cual se identifica las necesidades de la información para realizar script e incorporar en el software GIS (QGis) y estadístico (Rstudio), de acuerdo con las necesidades de análisis temporal de los datos.

3. ***Fase 3 - Análisis exploratorio de datos.***

Conforme a los pasos propuestos por la metodología definida por John W. Tukey (E.D.A.: Exploratory data analysis) y Box Jenkins, se realiza. ***A) Descripción***: Paso en el cual se busca identificar la base de datos, realizando el cargue y consultando tipo de variables, números de filas. ***B) Análisis Exploratorio:*** Consulta de valores máximos y mínimos, la media, mediana, varianza, desviación estándar, clase de variable, año y mes de inicio y finalización de la serie,

[2] https://www.datos.gov.co/Minas-y-Energ-a/ANM-Producci-n-Nacional-de-Minerales-y-Contraprest/r85m-vv6c

[3] https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-mgn-marco-geoestadistico-nacional/

cuartiles, entre los métodos gráficos encontramos histograma de frecuencias, Boxplot, grafica de serie, tendencia, grafica de funciones de autocorrelación, verificación de estacionariedad por medio de las pruebas de raíces unitarias Dickey Fuller aumentada. ***B) Análisis cartográfico:***) Graficar información cartográfica resultante, con el fin de realizar un análisis de comportamiento espacial.

4. Fase 4 – Evaluación.

Paso en el cual se proponen modelos y así observar cual de ellos presenta un mayor ajuste e información bayesiana, seguido de la realización de la prueba de autocorrelación de Ljung-Box de normalidad y aleatoriedad, buscando afirmar que los residuales del modelo sean auto correlacionados, normales y aleatorios, cumpliendo así los supuestos del modelo.

RESULTADOS

El análisis exploratorio con herramientas estadísticas ayuda a extraer información adicional a la información georreferenciada, la cual podría no ser evidente mirando simplemente el mapa, ya que se trata de datos cómo distribución de los valores de los atributos, tendencias, patrones, modelos y pronósticos; a diferencia de las funciones de consulta, tales como identificar o selección sobre las entidades individuales como los ofrecen los sistemas de información geográfico básico, el análisis estadístico revela las características de un conjunto de entidades como un todo. (ESRI, 2020). En este contexto se presentarán los resultados más relevantes sobre la data Producción Nacional de Minerales y Contraprestaciones Económicas Trimestral, conteniendo información detallada referida la cantidad de mineral extraído en el territorio nacional, asociando a la contraprestación económica generada por municipio productor desde la vigencia 2012 a 2012-2[4], la cual para el ejercicio es filtrada por el mineral Oro y se realiza consecuente a la metodología planteada, obteniendo los siguientes resultados:

[4] Información publicada por la Agencia Nacional de Minería del Grupo de Regalías y Contraprestaciones Económicas.

En la **Etapa 1 - Descarga y Alistamiento** se descarga los datos referentes a la producción nacional de minerales 2012 a 2020 en la pagina datos.gov[5]; seguido de la descarga en escala del marco geoestadisitico nacional[6]. Posterior a esto se realiza el cargue de datos en software libre SIG QGIS[7] con el fin de asociar (join), la información alfanumerica y geografica, visualizando así los municipios de producción del mineral Oro en Colombia; en Rstudio[8] se carga la base de datos y aplicando una transformación para reducir la escala de valores, visualizando cada en millones de pesos.

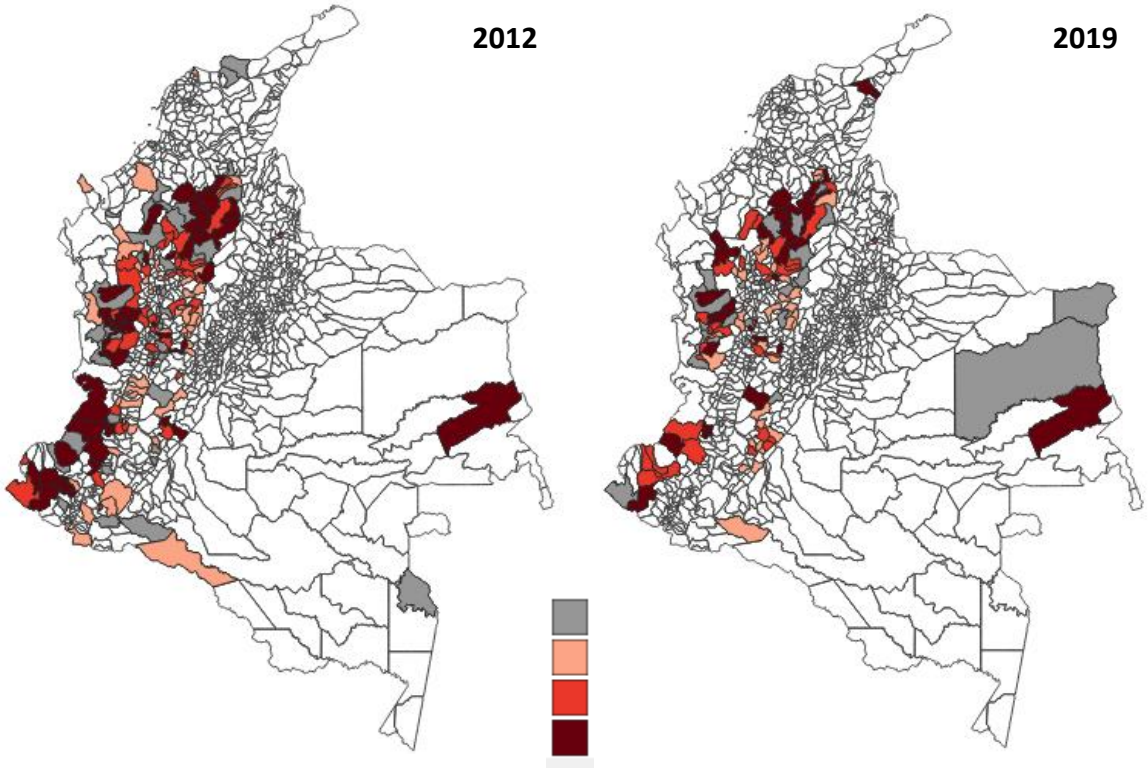

Ilustración 3 – Producción de Oro en Colombia 2012 a 2019. Fuente: Elaboración propia.

[5] Para más información puede consultar: https://www.datos.gov.co/Minas-y-Energ-a/ANM-Producci-n-Nacional-de-Minerales-y-Contraprest/r85m-vv6c

[6] Para más información puede consultar: https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-mgn-marco-geoestadistico-nacional/

[7] Para más información puede consultar: https://www.qgis.org/es/site/

[8] Para más información puede consultar: https://rstudio.com/

La **Etapa 2 - Identificación y elaboración de script** se plantea los script a genrar para realizar un tratamiento, analisis y pronostico de datos.

Codigo QGIS - R	Codigo RStudio
##Basic statistics=group **##Layer=vector** **##Field1=Field Layer** **Summary_statistics1<-** **…** **>Summary_statistics1**	cat(paste("El número de filas son:",length(pno$VALORNAC)),"\n", paste("El valor mínimo es: ",min(pno$VALORNAC)),"\n", paste("El valor máximo es: ",max(pno$VALORNAC)),"\n", paste("El valor promedio es: ",mean(pno$VALORNAC)),"\n",

En la **Etapa 3 - Análisis exploratorio de datos** es la actividad en la cual se identifico valores como: a) El número de filas son: 34 , b) Valor mínimo: 20809.7268, c) valor máximo es: 58094.917834, d) promedio: 39590.6092553235, e) mediana: 37387.425715, f) varianza: 119580784.558205, g) Desviación estándar: 10935.299929961; h) Clase de la variable: matrix, i) Año y mes de inicio de la serie: 2012 – 1, j) Año y mes de finalización de la serie: 2020 – 2, k) Frecuencia: 4 necesarios para la posterior clasificación del mapa, l) Cuartiles: 0%: 20809.73, 25%: 30365.95, 50%: 37387.43,75%: 50459.76 y100%: 58094.92 y así gráficamente obtenemos, el histograma y el Boxplot:

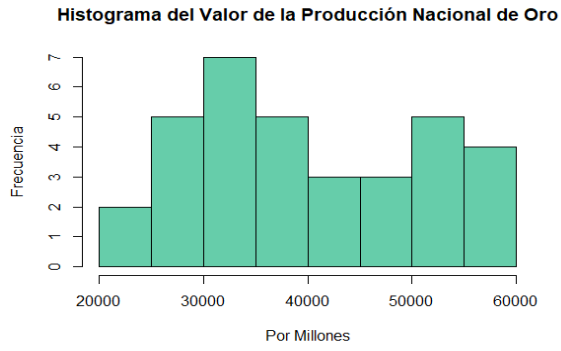

Ilustración 4 - Gráfico Histograma. Fuente: Elaboración propia.

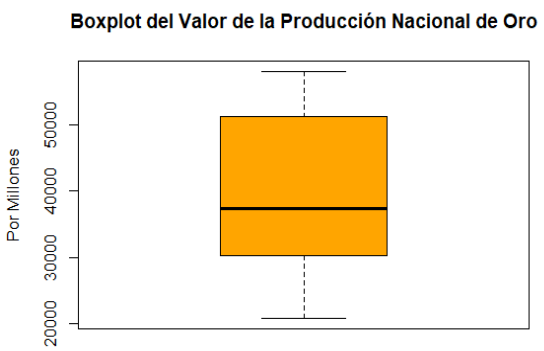

Ilustración 5 - Gráfico Boxplot. Fuente: Elaboración propia.

En el gráfico de la serie correspondientes a las ilustraciones 6 y 7, se analiza el valor de la producción nacional del oro (VALORNAC) muestra una tendencia que decrece hasta cierto punto y luego crece, oscilando en ese comportamiento, esto muestra que no posee media y varianzas constantes; se sospecha que es estacional.

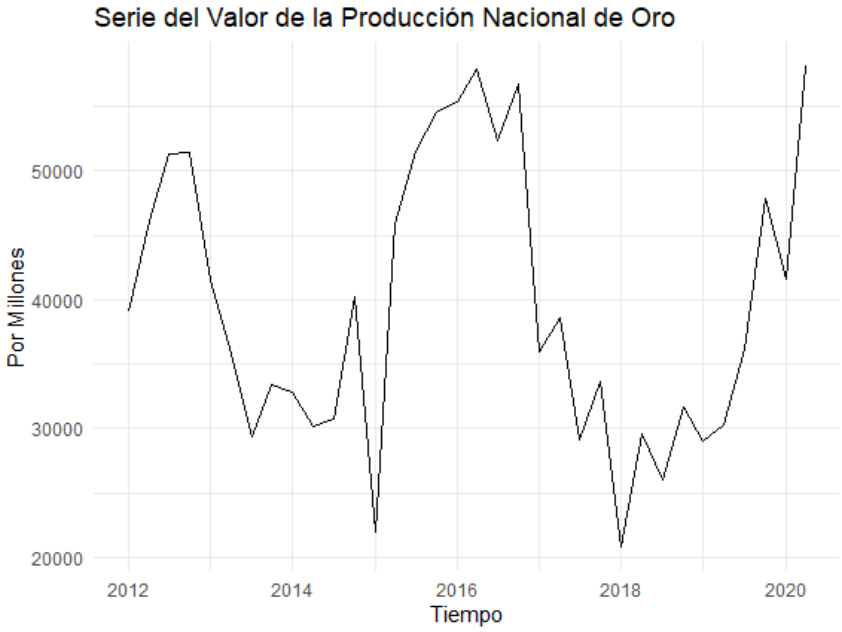

Ilustración 6 - Gráfico de la serie, producción nacional del Oro 2012 a 2020. Fuente: Elaboración propia

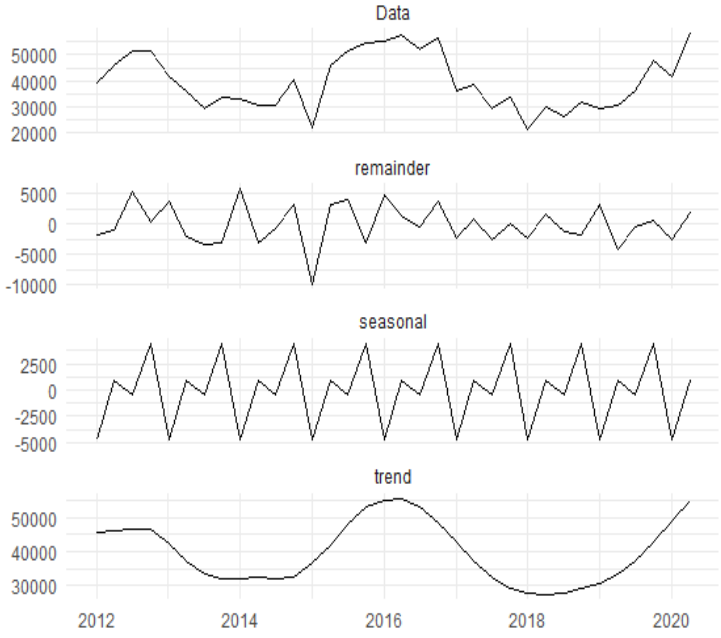

Ilustración 7 - Gráfico de la Serie estacionalidad, producción nacional del Oro 2012 a 2020. Fuente: Elaboración propia

En la ilustración 8 se observan los gráficos de las funciones de autocorrelación y autocorrelación parcial para la serie original y aplicando diferenciación.

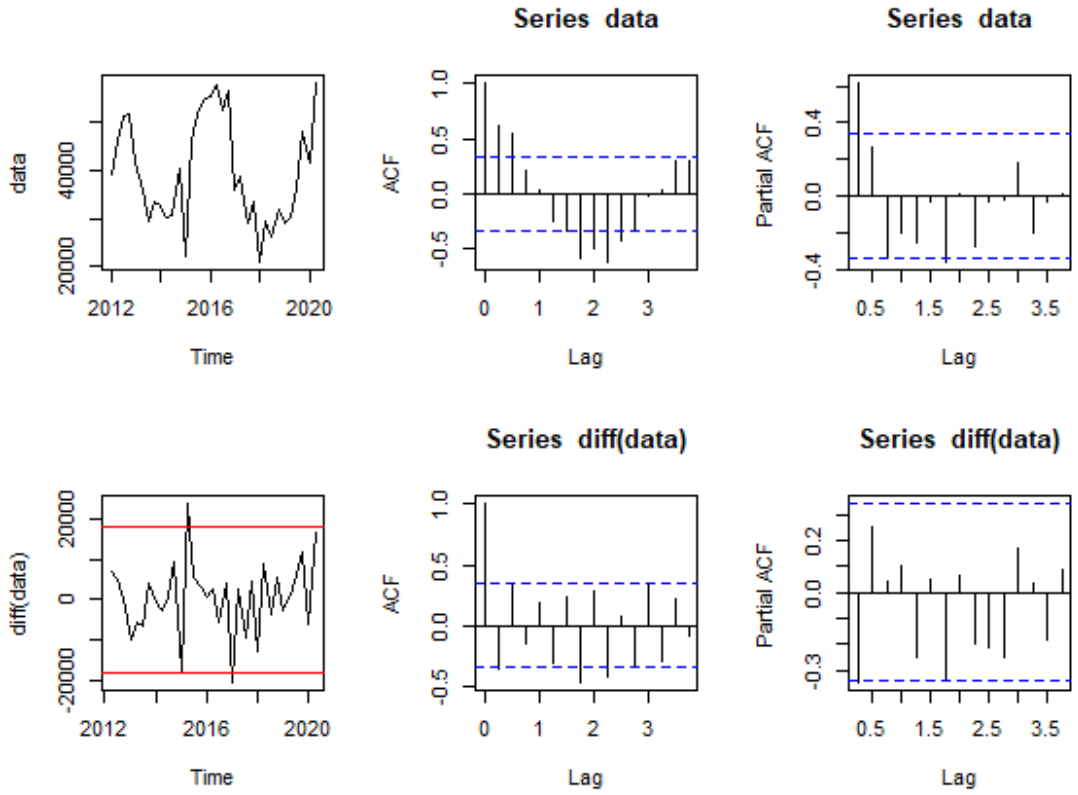

Ilustración 8 - Autocorrelación de la serie original y diferenciada. Fuente: Elaboración propia.

Se observa que la serie diferenciada muestra una tendencia constante y la varianza no es estable; es por ello se comprobara si la serie cumple la hipótesis de estacionariedad a través de las pruebas de raíces unitarias de Dickey-Fuller aumentada, obteneindo como resultado que el p −valor está por debajo del nivel de significancia ($\alpha = 0.05$), por lo cual se rechaza la hipótesis nula de existencia de raíces unitarias, y a la vez se concluye que la serie diferenciada del Valor Nacional de la producción del oro es estacionaria.

Etapa 4 – Selección del modelo y pronostico, conforme a los resultados satisfactorios de la serie diferenciada se tuvieron en cuenta varias modelos que se obtuvieran mejor criterio de información Bayesiana y ajuste, cómo lo fueron 1) $ARIMA(1,1,0)$, 2) $ARIMA(1,1,1)$, 3) $ARIMA(7,1,7)$; los residuales de estos modelos aprueban los test de Autocorrelacion de Ljung-Box, de normalidad y aleatoriedad no rechanzando las hipotesis nulas correspondientes.

```
modelo1<-stats::arima((pnots[,'VALORNAC']),order=c(1,1,0))
modelo2<-stats::arima((pnots[,'VALORNAC']),order=c(1,1,1))
modelo3<-stats::arima((pnots[,'VALORNAC']),order=c(7,1,7))
cbind(BIC(modelo1),BIC(modelo2),BIC(modelo3))

##         [,1]     [,2]     [,3]
## [1,] 697.1496 695.1584 717.6092
```

Pero, el mejor modelo según el Criterio de Información Bayesiana, es el $ARIMA(1,1,1)$, debido a que todos los p −valores, están por encima de nivel de significancia ($\alpha = 0.05$), por lo cual no se rechaza la hipótesis de cada prueba y se puede afirmar que los residuales del modelo son autocorrelacionados, normales y aleatorios; así, cumpliendo el supuesto del modelo, graficando así el pronostico del modelo.

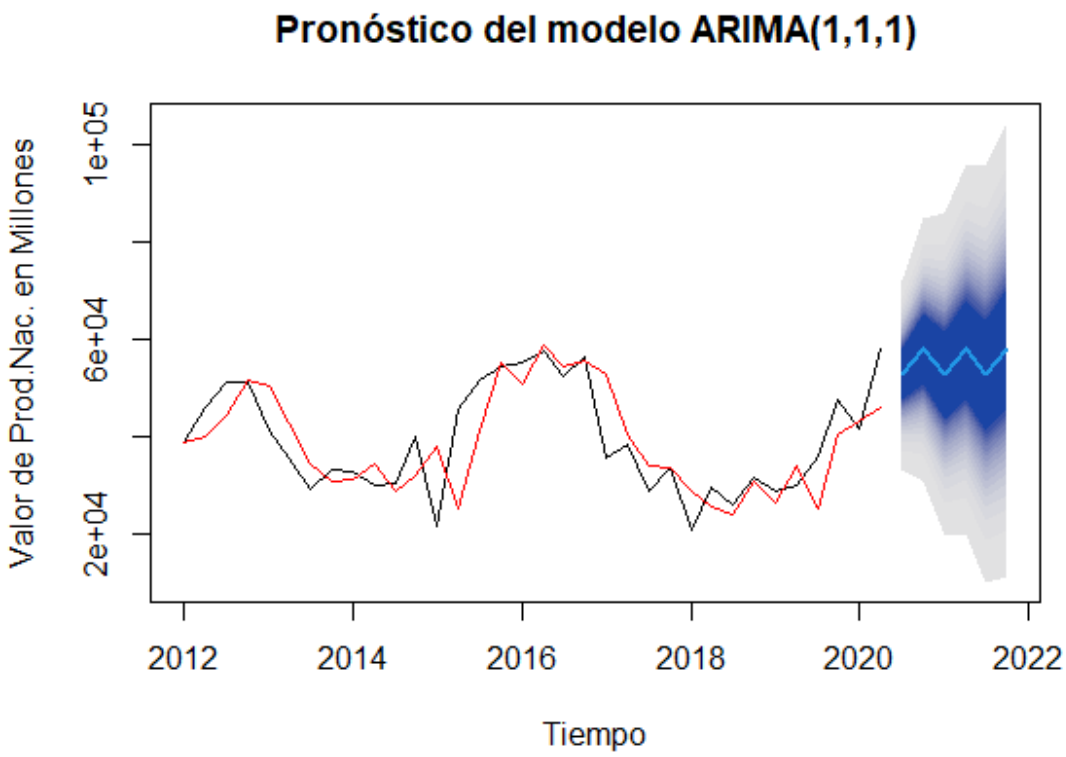

Ilustración 9 - Pronostico del modelo ARIMA. Fuente: Elaboración propia

CONCLUSIONES

Como resultado metodologico se puede indicar que el analisis exploratio de datos abiertos que permiten ser localizados en un espacio geograficos y ademas cuentan con inforación temporal, es necesario para su analisis la sinergia de las ciencias cartograficas y estadisticas para su mayor aprovechamiento, es por esto que la propuesta presentada toma valor para los profesionales de dichas diciplinas o interesados en el aprovechamiento de datos abiertos de este tipo; los resultados son los planteados en el comienzo de la iniciativa con el valor agrado de seguir una ruta conductora planteada por la docente el cual permitio obtener una trazabilidad y consecucion desde la elección del tema, genración del arbol de problemas, objetivos, consulta y lectura de proyectos similares al nuestro, teoria, manuales y demas docuemntos, los cuales soportan la necesidad y pertinencia del mismo.

Por otra parte en los resultado referidos al analisis de datos de la contraprestacion generada por la explotación del oro desde 2012 hasta el segundo trimestre del 2020, se obtiene entonces la obtención de una serie estacionaria aplicando la diferenciación, que a la vez no presenta una media y varianza constante, observando que en el 2016 se incremento la extracción y en 2015 y 2018 una caida de la contraprestación económica; planteando así un modelo ARIMA (1,1,1), conforme a un criterio de información Bayesiana y afirmando que los residuales del modelo son autocorrelacionados, normales y aleatorios.

Por ultimo se plantea la realización de analisis con cada uno de los minerales de la base de datos y realizar un mayor enfoque en la genración de script lenguaje R, para el fortalecimiento del software libre GIS, cómo herramienta fundamental para la incorporación de técnicas estadisticas en los procesos cartograficos, realizados por profesionales de ciencias de la tierra, en busqueda del mejoramiento y atomatización de procesos, procedimeintos, actividades y tareas, resultados que seran reflejados en la calidad del dato en pro de la toma de desiciones.

REFERENCIAS BIBLIOGRÁFICAS

Agencia Nacional de Mineria ANM. (16 de 09 de 2020). *Datos Abiertos*. Obtenido de ANM Producción Nacional de Minerales y Contraprestaciones: https://www.datos.gov.co/Minas-y-Energ-a/ANM-Producci-n-Nacional-de-Minerales-y-Contraprest/r85m-vv6c

Cadena, Á., & Pinzón, W. (2011). Clusters minero energéticos en Colombia: Desarrollo, hallazgos y propuestas. *Revista de Ingeniería, Universidad de los Andes*, 49-60.

Castillo Giraldo, A., Rodríguez Álvarez, D., & Carrillo García, J. (2015). *Lattice Data: Plugin de QGIS que implementa análisis estadístico exploratorio de datos lattice para la identificación de correlación espacial.* Bogotá D.C.: Universidad Distrital.

Corso Sicilia, G., & Pinilla Rivera, M. (2017). Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas. *Cuadernos Latinoamericanos de Administración*, 92-104. doi:https://doi.org/10.18270/cuaderlam.v13i25.2417

DANE. (2018). *Manual de Uso del Marco Geoestadístico Nacional en el Proceso Estadístico.* Bogotá D.C.: DANE. Obtenido de https://www.dane.gov.co/files/sen/lineamientos/manual-uso-marco-geoestadistico-nacional-en-proceso-estadistico.pdf

Departamento Administrativo Nacional de Estadística DANE. (2020). *DANE*. Obtenido de https://www.dane.gov.co/

ESRI. (2020). *ArcGIS Desktop*. Obtenido de https://desktop.arcgis.com/es/arcmap/10.5/analyze/commonly-used-tools/statistical-analysis.htm

Estupiñán Vargas, F., & Polanía, O. (2011). Las locomotoras del desarrollo: Minas, energía e innovación. *Revista de Ingeniería, Universidad de los Andes*, 44-48. Obtenido de https://revistas.uniandes.edu.co/doi/abs/10.16924/revinge.34.8

Fedesarrollo, Cárdenas, M., & Reina, M. (2008). *La Minería en Colombia: Impacto Socieconomico y Fiscal.* Bogota D.C.: La Imprenta Editores Ltda.

Francois, J. (2018). *Analisis Espacial con R: Usar R como un Sistema de Información Geografica.* Kočan, Macedonia: European Scientific Institute. Obtenido de https://eujournal.org/files/journals/1/books/JeanFrancoisMas.pdf

González, J. (2020). La importancia de la georreferenciación y la geolocalización para las empresas. *Índice*, 25-27. Obtenido de http://www.revistaindice.com/numero76/p25.pdf

Jaramillo Ayerbea, M., González Gómeza, D., Núñez Cabreraa, M., & Lucio García, J. (2007). Análisis de series de tiempo univariante aplicando metodología de Box-Jenkins para la predicción de ozono en la ciudad de Cali, Colombia. *Revista Facultad de Ingeniería*, 79-88. Obtenido de https://revistas.udea.edu.co/index.php/ingenieria/article/view/20192/17017

Maldonado Cecilia, J. (2020). Si en el momento de la creación de las instituciones ya la estadística y la geografía estaban íntimamente relacionadas, parece que esta relación aumenta día a día. *Índice*, 4-6.

Meza, J., & Arias, F. (2018). Análisis exploratorio de datos espaciales como herramienta de estudios geográficos. *Geoweb*, 14. Obtenido de

https://hum.unne.edu.ar/revistas/geoweb/Geo26/archivos/congreso%20geografia/Exposiciones/Exposiciones%20Eje%203/Mesa-Arias_EJE3.pdf

MINTIC. (2019). *Guía para el uso y aprovechamiento de Datos Abiertos en Colombia.* Bogotá D.C.: MinTIC.

Parra Olivares, J. (2002). Análisis exploratorio y análisis confirmatorio de datos. *Espacio Abierto. Cuaderno Venezolano de Sociología*, 115-124. Obtenido de https://produccioncientificaluz.org/index.php/espacio/article/view/2023

Pérez Medinilla, Y., & Crespo Borges, T. (2018). Análisis Exploratorio de Datos a Traves de Mapas Conceptuales. *Publicación Latinoamericana y Caribeña de Educación*, 96-105. Obtenido de https://www.researchgate.net/publication/326232106_analisis_exploratorio_de_datos_a_traves_de_mapas_conceptuales

Talaya, J. (2018). Georreferenciación y Datos Estadísticos. *Índice*, 7-8.