Artículo Original

Gestión de la contaminación ambiental mediante técnicas de minería de datos

Ambient pollution manage through data mining tools

Dianelis Portal-Castillo [1] [*]

[1]Centro Meteorológico Provincial, Sancti Spíritus, Cuba.

http://opn.to/a/l3ytc

RESUMEN: Las emisiones de sustancias contaminantes hacia la atmósfera y la magnitud de ellas, son la causa de muchos problemas ambientales en la actualidad. El presente trabajo se enfocó en la contaminación atmosférica de la provincia de Sancti Spíritus, donde se buscó encontrar relaciones entre variables climatológicas y las emisiones contaminantes por medio de técnicas de minería de datos que se puede definir como el proceso de extraer conocimiento válido, útil y comprensible que se encuentra en grandes conjuntos de datos.

Palabras clave: Fuentes fijas, contaminación, emisiones, minería de datos.

ABSTRACT: The emissions of polluting substances into the atmosphere and the magnitude of them are the cause of many environmental problems at present. The present work focused on air pollution in the province of Sancti Spíritus, where it was sought to find relationships between climatic variables and pollutant emissions through data mining techniques that can be defined as the process of extracting valid, useful knowledge and understandable that it is found in large data sets.

Keywords: Fixed sources, pollution, emissions, data mining.

[*]Autor para correspondencia: *Dianelis Portal-Castillo*: E-mail: dianelis.portal@ssp.insmet.cu

Recibido: 18/07/2018

Aceptado: 07/11/2018

INTRODUCCIÓN

La calidad del aire constituye uno de los problemas ambientales más graves que enfrenta la humanidad. Las emisiones de contaminantes a la atmósfera no sólo tienen efectos a nivel global, como el cambio climático y la reducción del espesor de la capa de ozono estratosférico (PNUMA, 2002), sino también a nivel regional. Cuba no escapa de esta problemática, en el caso específico de la provincia de Sancti Spíritus, el constante crecimiento poblacional en la ciudad de Sancti Spíritus, unido a la actividad industrial, la cual presenta tecnología deficiente para llevar a cabo el proceso industrial en la mayoría de los casos vienen provocando deterioro a la calidad del aire, aumentado la contaminación a la atmósfera. Por las razones antes mencionadas fue necesario y urgente obtener el inventario de fuentes que contaminan la provincia de Sancti Spíritus, con la finalidad de contar con una información que nos caracterice las fuentes fijas, su localización y tipo de contaminante emitido, para que sirva de base al estudio de la contaminación atmosférica.

La Minería de datos (o "Data mining", como se le conoce en inglés) es una técnica de análisis y estudio de datos que está surgiendo con fuerza en los últimos años, gracias a la posibilidad de aplicar el potencial de procesamiento de datos de los actuales ordenadores. En sí misma, la minería de datos no es más que una técnica para el análisis y procesado de grandes volúmenes de datos con el objeto de extraer información útil y patrones de fácil comprensión, que sería imposible conseguir por los medios y herramientas estadísticas tradicionales. La calidad del aire es uno de esos campos en los que más útil e interesante podría resultar aplicar estas técnicas de análisis, por lo cual la siguiente investigación tiene como objetivo analizar mediante el uso de la herramienta Openair la emisión de contaminantes atmosféricos a la atmósfera en la ciudad de Sancti Spíritus.

MATERIALES Y MÉTODOS

Características generales de la zona de estudio

La provincia de Sancti Spíritus se encuentra ubicada en la parte central de la isla entre los 21°56′02″ de latitud norte y los 79°26′38″ de longitud oeste, limitando al oeste con Villa Clara y Cienfuegos, al norte con el Estrecho de la Florida, al este con Ciego de Ávila y al sur con el Mar Caribe. Tiene un área de 6777 km² y una población de 466 251 habitantes. (ONE, 2016) Se caracteriza por un relieve variado, con aproximadamente el 80 % de llanuras y el resto de montañas; tiene unos 237 km de costa, de ellas 60 km en la norte y 171 km en la sur.

Modelación con Aermod

En 2005 la Agencia de Protección Ambiental de los Estados Unidos (Environmental Protection Agency, EPA) estableció el AERMOD como el modelo de uso recomendado para la dispersión de contaminantes a escala local, en sustitución del ISCST3, hasta ese momento usado. Ha sido demostrado y documentado, tanto por evidencias científicas como por estudios de validación, que el AERMOD representa un sólido y significativo avance respecto al ISCST3. La formulación del AERMOD ha sido sometida a una revisión profunda e independiente, lo que permite concluir que las bases científicas del AERMOD están al nivel del estado del arte de la ciencia (Cimorelli, et. al., 2005). AERMOD representa una técnica de dispersión que incorpora las técnicas más avanzadas de parametrización de la capa límite planetaria, dispersión convectiva, formulación de la elevación de la pluma e interacciones complejas del terreno con la pluma razones por las cuáles fue escogido para el estudio. Mediante el uso del modelo AERMOD fueron modelados datos de la emisiones contaminantes de 2 fuentes emisoras de la ciudad de Sancti Spíritus.

Openair una herramienta para el análisis de calidad del aire

R es en sí mismo un lenguaje de programación creado en un entorno pensado para el análisis estadístico y gráfico de datos, siendo un software libre que se distribuye bajo licencia GNU GPL. R como entorno de programación se desarrolla mediante librerías (también llamadas en R como paquetes) que lo que hacen es completar el lenguaje con nuevos desarrollos previstos para distintas áreas del análisis estadístico y gráfico de los datos(R Developement Core Team 2012). A efectos prácticos R consiste básicamente en un lenguaje de programación para el estudio

estadístico de datos que presenta subconjuntos de lenguaje desarrollados para áreas específicas del análisis de datos, como en nuestro caso la calidad del aire. Openair es una de esas librerías desarrolladas por David Carslaw del King's Collegue London y Karl Ropkins de University of Leeds, que permite comunicar los resultados de forma fácil y sobre todo, es fácil de usar, específicamente diseñada para tratar datos de calidad del aire (Carslaw, Ropkins, 2012).

Por qué R y Openair?:

Son herramientas gratuitas, lo cual es un motivo de enorme peso.

R es un lenguaje abierto, que mejora día a día, y que cuenta con la aportación de expertos en minería de datos de muy diversos campos en todo el mundo.

Dispone de módulos específicos para el tratamiento de datos de la calidad del aire, y entre ellos uno en concreto, denominado Openair, que presenta herramientas de minería sencillas y muy atractivas.

La instalación y manejo del sistema es "bastante" sencillo, funciona en casi cualquier ordenador, y bajo múltiples sistemas operativos.

El procesado de información es extraordinariamente rápido, aun cuando el volumen de datos sea gigantesco(Pliego, 2012).

RESULTADOS Y DISCUSIÓN

Los datos analizados corresponden al Grupo Electrógeno y a la Planta de Asfalto de la ciudad de Sancti Spíritus en el año 2014. Luego de obtener el inventario de emisiones de las fuentes antes mencionadas se procedió a la modelación de la emisión con el modelo AERMOD.

Con los datos arrojados por dicha modelación comprobaremos los beneficios que nos ofrecen las herramientas R y Openair.

calendarPlot: La función gráfica calendarPlot aporta al técnico una herramienta muy práctica a la hora de elaborar gráficas que resulten más asequibles al público en general, y que sirvan para detectar más detalladamente la evolución de las medidas diarias de un contaminante, que quedan ahora recogidas de una forma más comprensible (Carslaw, 2015), tal y como se puede observar en la siguiente gráfica para la evolución del SO2 en nuestro conjunto de datos durante el año 2014.

Esta función es muy gráfica, y permite ver los días con diferentes promedios diarios de contaminación, SO2 en este caso. Quedando representados los valores más críticos de concentración en color mas oscuro.

polarPlot: representa un parámetro seleccionado en función de las distintas combinaciones de dirección y velocidad del viento. La utilidad de este tipo de gráficas es evidente puesto que son capaces de identificar con bastante detalle fuentes potenciales de origen de la contaminación y su influencia respecto a los niveles globales de contaminación detectados por una estación.

La función polarPlot permite al usuario obtener una idea sobre los focos de origen de la contaminación al mostrar sobre una rosa de los vientos la velocidad del viento y la concentración esperada de un determinado contaminante. En el expuesto en la Figura 2 podemos observar que las mayores concentraciones de SO2 se detectan en dirección Norte-Noreste y con vientos medios que estarían entre los 5 m/s y los 7m/s.

timeVariation: analiza la variación que presentan uno o más parámetros con respecto a distintos períodos temporales de carácter cíclico (días, semanas, años), estableciendo así patrones de evolución en el comportamiento de dichos parámetros en la zona de estudio es, evidentemente, una herramienta de enorme utilidad para el análisis de la calidad del aire.

En el ejemplo de la Figura 3 la primera gráfica nos permitirá ver los días de la semana en los que se producen las mayores concentraciones contaminantes y las horas del día en que ocurren las mismas observando que en el caso de estudio ocurren los martes al final de la tarde y los jueves en horas de la mañana. Luego tenemos una gráfica resumen donde se establece la evolución del parámetro en las horas del día, por lo se puede observar en qué horas ocurren las mayores concentraciones. La tercera gráfica resume donde se establece la evolución del parámetro en función de los meses del año, con la que se apreciará la evolución a lo largo del año y comprobar, por ejemplo, si existe algún tipo de estacionalidad en el dato viendo en el ejemplo graficado que las mayores concentraciones corresponden a los meses Noviembre y Diciembre del período seco. El último gráfico

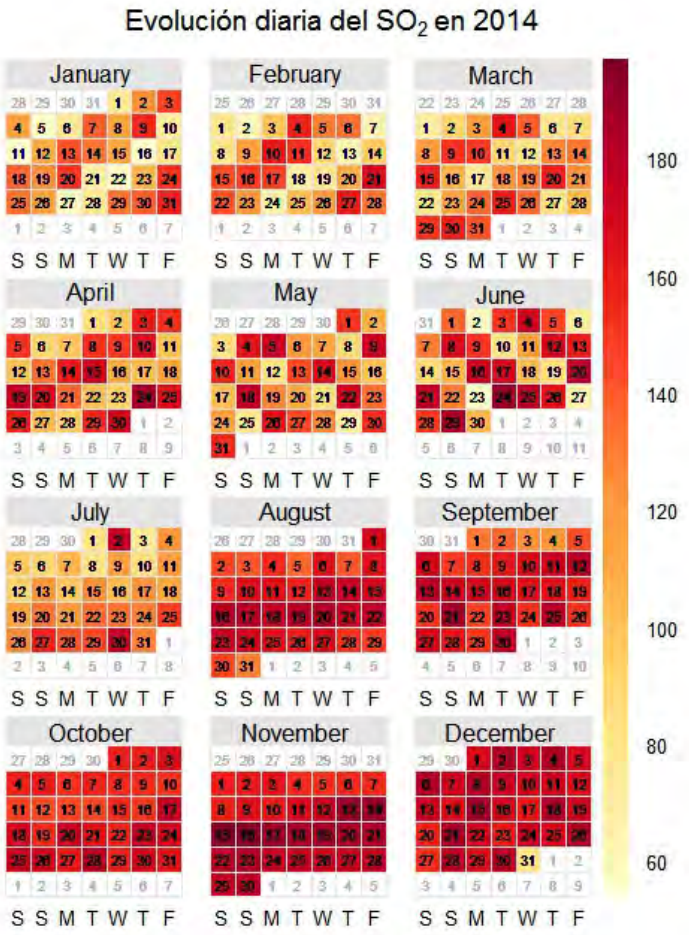

Figura 1. Evolución diaria del SO2 en 2014

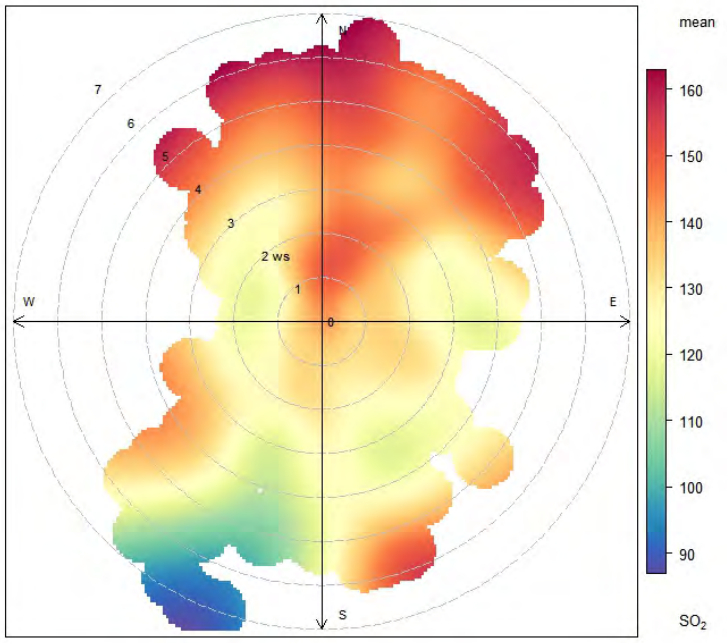

Figura 2. Distribución de medias de SO2.

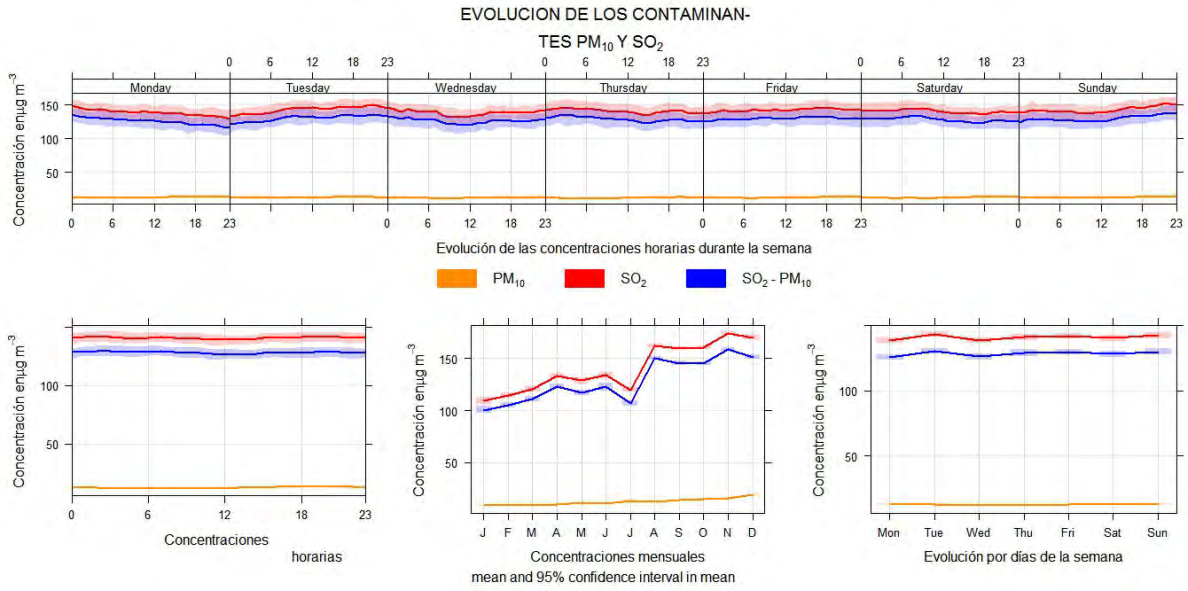

Figura 3. Evolución temporal de los contaminantes SO2 y PM10.

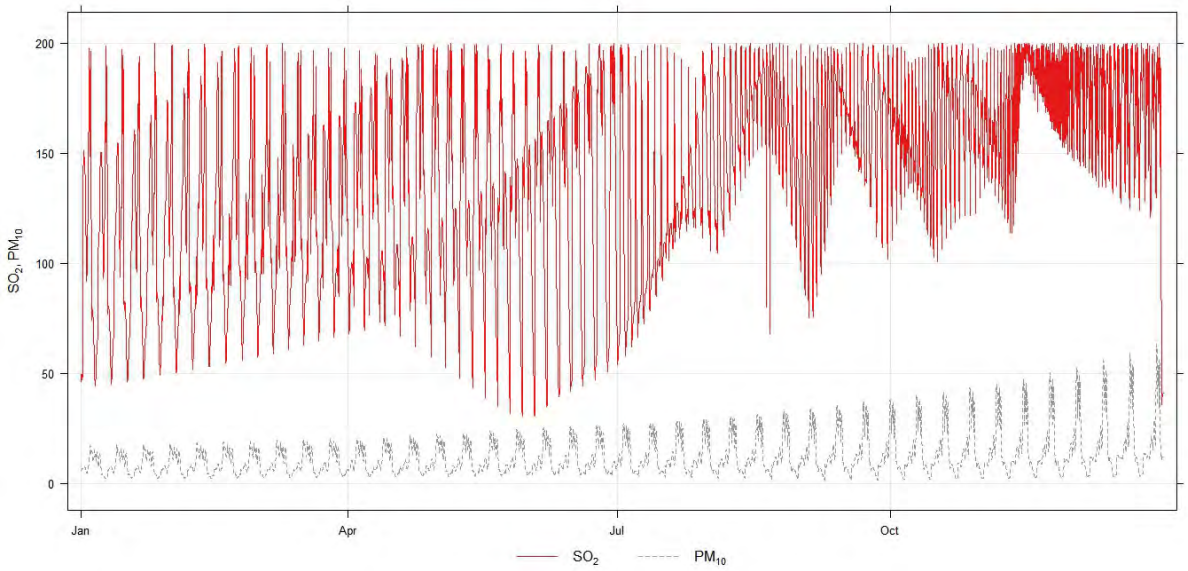

Figura 4. Evolución temporal de los contaminantes SO2 y PM10 en todo el período.

resume la evolución de las concentraciones por días de la semana, mostrando las concentraciones por cada día, pudiendo observarse como evolucionan a lo largo de la semana.

timePlot: La función gráfica de tiempo se corresponde con la función gráfica clásica, similar al Plot que anteriormente hemos visto que utilizaba R, en la que se representa la evolución de las concentraciones de un determinado parámetro, representado en el eje Y, frente al tiempo, representado en el eje X. La función gráfica de tiempo es en sí misma una gráfica típica, ampliamente utilizada en calidad del aire, y disponible a través de muchas de las actuales herramientas de cálculo del mercado. Sin embargo, Openair dota a esta función gráfica de unas peculiaridades y una versatilidad que la acaba convirtiendo en algo único, muy útil y práctico para el estudio de la evolución temporal de la contaminación.

Con el uso de la función **timePlot** se observa el comportamiento de los contaminantes en todo el período estudiado, muestra la información temporal de los contaminantes de una forma más generalizada que la función timeVariation. También es una herramienta útil para el análisis a priori del comportamiento de los contaminantes analizados.

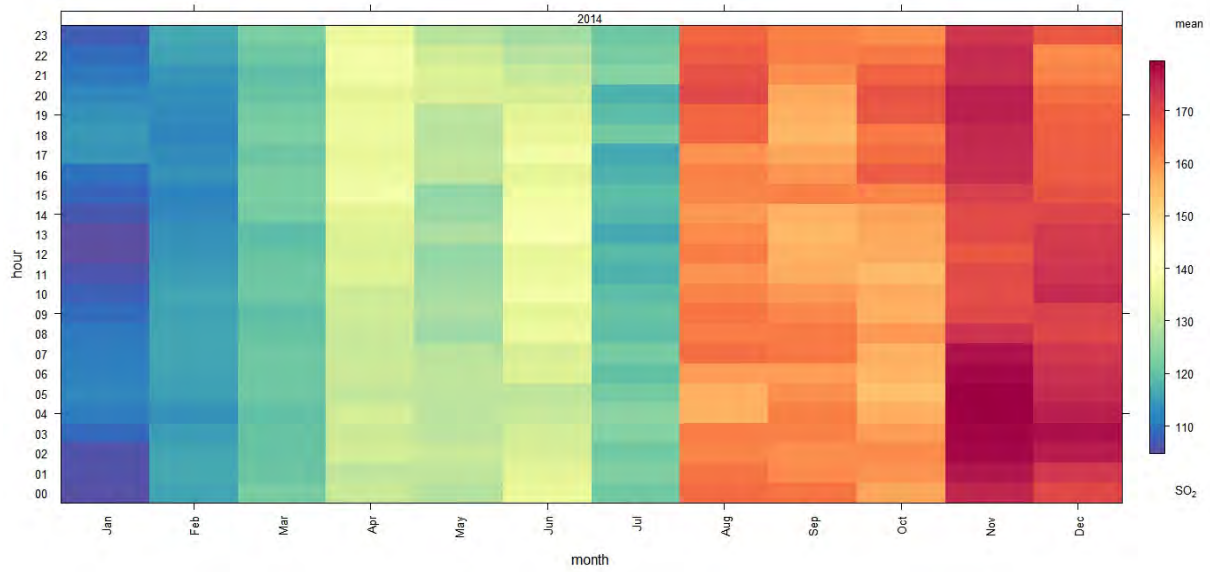

Figura 5. Gráfico de tendencias para el SO2 en el 2014.

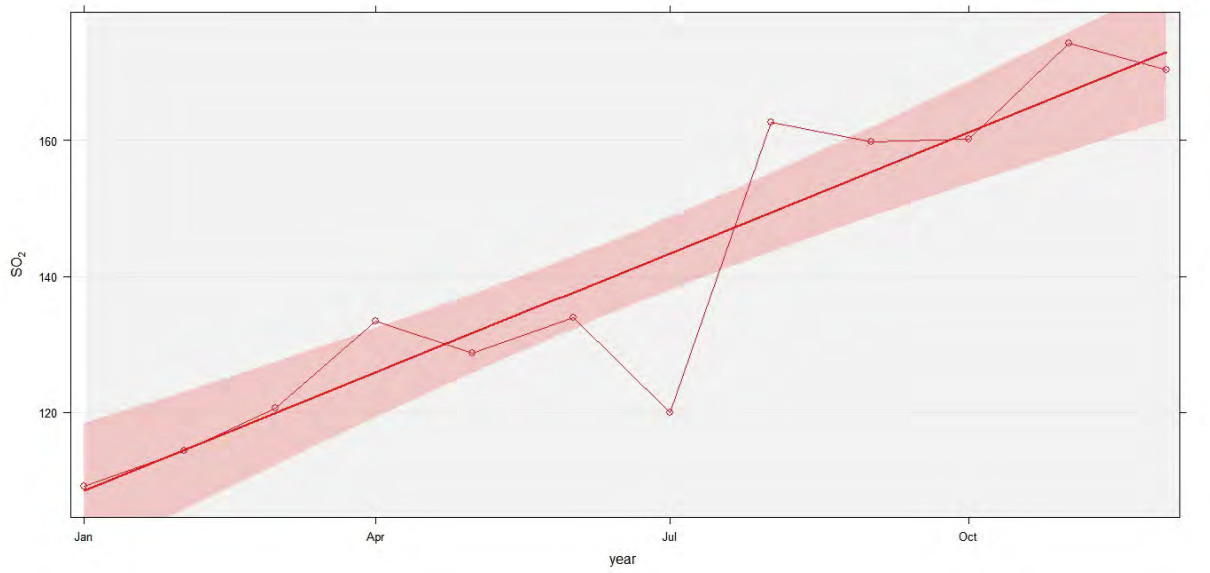

Figura 6. Media mensual de SO2 en el 2014.

trendLevel: proporciona una manera muy práctica y visual de condensar grandes bloques de datos de calidad del aire en una sola gráfica en la que además podremos comprobar como evolucionan las tendencias globales, diarias e incluso estacionales para los distintos contaminantes. La propiedad fundamental que tiene trendLevel, y que la diferencia del resto de funciones de Openair, es precisamente la capacidad de poder condensar la información al ofrecer diversas combinaciones para calcular el estadístico deseado del contaminante o parámetro que se le indique.

En la figura anterior se muestra la evolución de las concentraciones máximas de SO2 en función de las horas del día, los días de la semana, de forma que se puede observar como las emisiones se producen con independencia del día de la semana y con mayor intensidad en los meses del período seco, donde las condiciones meteorológicas suelen ser más desfavorables para permitir una correcta dispersión de los contaminantes.

smoothTrend: la función de tendencias suavizadas smoothTrend va más allá, y sustituye los parámetros de regresión lineal utilizados para calcular la tendencia por funciones de suavizado

que van adaptando la línea de tendencia a la evolución del parámetro. Desde el punto de vista práctico, y en su configuración más básica, la función smoothTrend lo que hace es generar una gráfica de concentraciones medias mensuales de un parámetro seleccionado, dando lugar al diagrama de dispersión de pares de datos sobre el que construye una línea de tendencia suavizada, y finalmente muestra el intervalo de confianza al 95%. Para hacer esto la función smoothTrend utiliza un modelo de regresión adicitivo denominado GAM (de sus siglas en inglés General Additive Model) proporcionado por otra librería de R especializada en el modelizado aditivo de datos(Pliego, 2012).

En la gráfica de tendencias suavizadas obtenidas para el SO2, se puede ver como las concentraciones medias de este contaminante presentan una inestabilidad durante todo el período analizado.

Estas son sólo alguna de las funciones que se pueden utilizar con R y Openair, entre otras muchas que contempla el propio paquete de Openair o paquetes adicionales de R.

CONCLUSIONES

Con el uso de estas herramientas se obtienen gráficas asequibles y fáciles de entender por el usuario. También permite hacer análisis especializados de los datos que no se logran con herramientas tradicionales.

RECOMENDACIONES

Continuar analizando las bondades que nos brinda el lenguaje R y Openair para así lograr análisis más completos y potentes de los datos que hemos obtenidos con los inventarios de emisiones provinciales.

REFERENCIAS

Carslaw, D.C. and Kart Ropkins, (2012) openair- an R package for air quality data analysis. Environmental Modelling & Software. Volume 27-28, 52-61.

Carslaw, D.C . and K. Ropkins, (2015): openair: Open-Source tools for the analysis of air pollution data. R package version 1.6, http://CRAN.Rproject.org/package=openair.

Carslaw, D.C . (2017). The openair manual - open-source tools for analyzing air pollution data. Manual for version 2.1-6, University of York.

Cimorelli, A. et. al. «AERMOD: A Dispersion Model for Industrial Source Applications. Part I: General Model Formulation and Boundary Layer Characterization». Journal of Applied Meteorology, 44(5): 682-693, 2005.

Oficina Nacional de Estadística e Información 2016b. *Anuario Estadístico de Sancti Spíritus 2015*. La Habana, Cuba: Oficina Nacional de Estadística e Información (ONEI).

Pliego, F. F. 2012. Lenguaje R aplicado al análisis de datos de Calidad del Aire - Manual de uso de R y Openair.

PNUMA. Perspectivas de medio ambiente mundial GEO-3. Grupo Mundi-Prensa. España. 2002.

R Developement Core Team (2012). R: A lenguaje and environment for statistical computing. R Foundation for Statistical Computing,Viena, Austria. ISBN 3-900051-07-0.

Los autores de este trabajo declaran no presentar conflicto de intereses.