Desarrollo de ETL para limpieza y trasformación de datos crudos de PM10 de la estación de monitoreo de calidad del aire de Nogales, Sonora

ETL for cleaning and transforming raw PM10 data from the air quality monitoring station in Nogales, Sonora

GARCÍA-ALVA, Sigifredo†*, MUÑOZ-ZAMORA, Guillermina, CRUZ–RENTERÍA, Jesus y NUÑEZ-SILVA, Oscar

Instituto Tecnológico de Nogales, Ave. Instituto Tecnológico # 911 Col. Granja CP. 84065, Nogales Sonora, Mexico. División de Estudios de Posgrado e Investigación (DEPI)

ID 1º Autor: *Sigifredo García-Alva* /**ORC ID:** 0000-0001-7559-1421, **Researcher ID Thomson:** F-6909-2018, **arXiv ID:** Sigifredo#1

ID 1º Coautor: *Guillermina Muñoz-Zamora* /**ORC ID:** 0000-0001-7480-8174, **Researcher ID Thomson:** F-4285-2018, **arXiv ID:** guillermina##

ID 2º Coautor: *Jesús Raúl Cruz–Rentería* /**ORC ID:** 0000-0002-9406-3154, **Researcher ID Thomson:** F-7988-2018, **arXiv ID:** raulcruzrenteria

ID 3º Coautor: *Oscar Ruben Nuñez-Silva* /**ORC ID:** 0000-0002-4292-8910, **Researcher ID Thomson:** F-7774-2018, **arXiv ID:** Oscar-nunez

Recibido: 9 de Enero, 2018; Aceptado 22 de Marzo, 2018

Resumen

En el 2012 el 11.5% de las muertes a nivel mundial estuvo relacionado con la contaminación del aire. Este trabajo tiene como objetivo aplicar un procedimiento desarrollado en SQL para la Extracción, Trasformación y Carga (conocido en inglés como Extract, Transform and Load (ETL)) para asegurar la limpieza, verificación y validación de los datos crudos obtenidos por la estación de monitoreo del aire instalada en el Instituto Tecnológico de Nogales en la ciudad de Nogales, Sonora, para contaminantes particulados de PM10. La etapa de trasformación se subdivide en limpieza y transformación de datos. La metodología aplicada en la etapa de limpieza de los datos permite verificar que cada dato meteorológico y contaminante que está registrado sea válido. La etapa de trasformación consiste en agrupar y trasformar la información por rangos de hora y día aplicando en cada uno de ellos su propia etapa de limpieza específicamente para cada grupo de datos.

ETL, PM_{10}, Limpieza, Validación

Abstract

In 2012, 11.5% of deaths worldwide were related to air pollution. The objective of this paper is to apply a procedure developed in SQL for Extraction, Transformation and Loading (ETL) to ensure the cleaning, verification and validation of the raw data obtained by the air monitoring station installed at the Instituto Tecnológico de Nogales in the city of Nogales, Sonora, for particulate pollutants of PM_{10}. The transformation step divided itself in cleaning and transforming data. The methodology used in the cleaning data step verifies that each weather and pollutant data registered is valid. The transformation step is about grouping and transforming the information by time and day applied in each their own cleaning step specifically for each data group.

ETL, PM10, Cleaning Data, Data Validation

Citación: GARCÍA-ALVA, Sigifredo, MUÑOZ-ZAMORA, Guillermina, CRUZ–RENTERÍA, Jesus y NUÑEZ-SILVA, Oscar. Desarrollo de ETL para limpieza y trasformación de datos crudos de PM10 de la estación de monitoreo de calidad del aire de Nogales, Sonora. Revista de Tecnología e Innovación 2018, 5-14: 25-29.

* Correspondencia al Autor (Correo electrónico: sga0097@gmail.com)
†Investigador contribuyendo como primerAutor.

1. Introducción

Las fuentes emisoras de contaminación atmosférica están compuestas por gases y partículas las cuales se integran por las que genera el hombre también conocida como antropogénica y que se subdividen en fijas (industria), móviles (parque vehicular) y de área (comercios y servicios) y la natural o biogenética generada por suelo, vegetación y meteorología. De acuerdo con un informe en Ginebra fechado el 26 de septiembre del 2016 la Organización Mundial de la Salud (OMS) publica que el 92% de la población a nivel mundial vive en lugares que exceden los límites de calidad del aire recomendados y estimó que en el 2012 estuvieron relacionadas con la contaminación del aire tanto de interiores como de exteriores 6.5 millones de muertes a nivel mundial que representa el 11.5% del total de ellas (OMS, 2016).

Los contaminantes atmosféricos más relevantes para la salud son material particulado (PM_{10}) con un diámetro de 10 micrómetros o menos, son tan pequeñas que pueden llegar a los pulmones, lo que puede causar graves problemas de salud (EPA, 2017). Para obtener datos útiles sobre la contaminación del aire se utiliza un procedimiento que se conoce como Extract, Transform and Load (ETL) el cual sirve para extraer, trasformar y cargar los datos de la estación de monitoreo. Este procedimiento también asegura la limpieza, verificación y validación de los datos por medio de banderas. Dentro de todo el proceso ningún dato se debe perder solo debe ser etiquetado como válido o no válido, en esto estriba la diferencia con un ETL para almacén (Data Warehouse) o minería de datos (Data Mining) en los cuales los datos no válidos son eliminados. (Connolly Tomas, 2015).

La ciudad de Nogales, Sonora, México cuenta con una estación de monitoreo de calidad del aire para el monitoreo de PM_{10} con equipos de muestreo como de la Figura 1. Debido a que la ciudad tiene una topografía donde predominan las cañadas y los arroyos, donde las vialidades se organizan sin un orden geométrico definido, con una topografía accidentada, con un estimado *65%* de calles sin pavimento, con un parque vehicular prácticamente de 1 por cada 2 habitantes, cuenta con más de 220,000 habitantes.

Finalmente la industria maquiladora generó casi 60 mil empleos de acuerdo a INEGI en el 2010, representan poco más del 45% entre los empleos relacionados directamente con la industria y los servicios que ésta demanda, es de suma importancia que los datos obtenidos por la estación de monitoreo de calidad del aire se puedan procesar para obtener los indicadores por hora y día para conocer el nivel de contaminación a la que se expone la poblacion (LT Consoulting, 2016).

Figura 1 Equipo de Muestreo de estación de monitoreo de calidad del aire

2. Metodología de desarrollo

Los datos analizados de la estación de monitoreo de la ciudad de Nogales, Sonora, México, comprenden el periodo del 1 de julio del 2015 al 30 de junio del 2016 (se tomó como año bisiesto ya que febrero tiene 29 días). Para cumplir las normas vigentes mexicanas las estaciones requieren de un procedimiento robusto de calibración y sus datos por minuto pasen por un proceso de verificación y validación asegurando su calidad para que sean generados los indicadores por hora y día. Para realizar este proceso se requirió del uso de un ETL en SQL para el procesamiento de los datos (DOF, 2012) (DOF, 2014).

El ETL se divide en 3 etapas que son la de extracción, transformación y carga, en la etapa de extracción se deben extraer datos de la estación de monitoreo y descargarlos en una base de datos como datos crudos.

ISSN-2410-3993

En la etapa de trasformación se pasan los datos por una serie de filtros y se marcan los datos no válidos y datos válidos, estos últimos se utilizan en el proceso de limpieza, verificación y validación para transformarlos a datos por minuto, hora y día como lo marca el Instituto Nacional de Ecología y Cambio Climático (INECC) a través del Sistema Nacional de Información de la Calidad del Aire (SINAICA) (INECC, 2016) (SINAICA, 2010).

Los datos limpios y validados por minuto, hora y día son cargados en tablas separadas de datos limpios. El proceso de ETL para asegurar la calidad de datos meteorológicos y de PM_{10} de la estación de monitoreo consta de los siguientes pasos:

1. Se extraen los datos crudos de las calibraciones mensuales de la estación de monitoreo en archivos extensión CSV como el que se muestra en la tabla 1, los cuales tienen como característica que sus valores están separados por comas.
2. Se cargan los datos crudos a una base de datos y se les coloca una bandera a cada uno de los datos meteorológicos y de PM_{10}.
3. Se revisa si los datos crudos obtenidos pasaron el proceso de calibración mensual, en caso contrario se marcan todos como datos no válidos.
4. Si los datos tienen banderas de prueba, de alarma o mantenimiento, faltan datos o hay números igualados a cero, se etiquetan como no válidos.
5. Se revisan si los datos están fuera de rango para flujo o contamínante, o si son constantes por más de 3 horas se etiquetan como no válidos. Todos los pasos de validación del paso 3 al 5 son en forma automática por medio de SQL.
6. Este paso de validación se hace de manera manual y se revisa la bitácora para buscar eventos que no fueron propios de la ciudad, pero incidieron en la estación de monitoreo, actividades donde se tuvieron que apagar los equipos de muestreo de PM_{10} para realizarles algún tipo de mantenimiento o hubo falla en la energía eléctrica debido a que estos equipos requieren de tiempo para estabilizarse y estos datos deben marcarse como no válidos.
7. El resto de los datos son etiquetados por una bandera como válidos y son copiados a otra tabla de datos limpios por minuto.
8. Para generar el indicador por hora se crea otra tabla de datos por hora para datos meteorológicos y de PM_{10}, se promedian los datos de la tabla por minuto en la tabla de horas y se le coloca una bandera con la cantidad de minutos encontrados en una hora.
9. Se verifica que cada hora cumpla con la condición de que exista el 75% de los datos o más para que los datos de la hora sean marcados como válidos, en caso contrario serán etiquetados como no válidos.
10. Para generar el indicador por día se crea otra tabla de datos por 24 horas para datos meteorológicos y de PM_{10}, se promedian los datos de la tabla por hora en la tabla de 24 horas y se le coloca una bandera para la cantidad de datos encontrados por 24 horas.
11. Se verifica que cada 24 horas se cumpla con la condición de que exista el 75% de los datos o más para que los datos de 24 horas sean marcados como válidos, en caso contrario se marcan como no válidos.

Report - Site Nogales Sonora Mexico : Time Beginning					
Date&Time	WD	Temp	RH	PM_{10}	Flow
	Grados	ºC	%	Ug/m3	Lt/min
4/13/2016 12:00	215.2	13.5	52.7	89.4	16.7
4/13/2016 12:01	207.5	13.4	52.6	88.5	16.7
4/13/2016 12:02	209	13.3	52.7	87.4	16.7
4/13/2016 12:03	248.8	13.5	52.8	86.3	16.7
4/13/2016 12:04	236.4	13.4	52.9	85.6	16.7
4/13/2016 12:05	218.6	13.5	53.1	84.6	16.7
4/13/2016 12:06	198.2	13.5	53.4	83.2	16.7
4/13/2016 12:07	192.6	13.5	53.3	81.9	16.7
4/13/2016 12:08	196.3	13.5	53	80.3	16.7
4/13/2016 12:09	197.2	13.4	53.3	78.6	16.7

Tabla 1 Muestra de datos de archivo CSV

Con los datos obtenidos por 24 horas (indicador diario) para los datos de PM_{10} se logró definir los siguientes datos de calidad del aire por día de la ciudad de Nogales, Sonora, México, para el rango del 1 de julio del 2015 al 30 de junio del 2016 de acuerdo al rango de tiempo de mínimo un año y los niveles de contaminación de acuerdo a la NOM de salud ambiental vigente (DOF, 2014).

ISSN-2410-3993

- Los días con buena calidad del aire: se dan cuando el dato diario obtenido se ubica entre cero y menos de 37.5 ug/m^3.
- Días con calidad del aire regular: son cuando el dato diario obtenido se ubica en el intervalo 37.5 ug/m^3 y 75 ug/m^3.
- Días con mala calidad del aire: se presentan cuando el dato diario obtenido rebasa el límite especificado de 75 ug/m^3.
- Así mismo con los datos obtenidos y dado que se cumplió con la condición de tener más del 75% de los datos diarios en un lapso de un año se logró conocer el promedio anual y fue comparado con el promedio anual de la NOM de salud ambiental vigente (DOF, 2014).

3. Resultados

Con la aplicación del ETL se obtuvieron los siguientes resultados de la estación de monitoreo de Nogales, Sonora, México: un total de 526,980 registros de datos crudos por minuto, de los cuales solo fueron válidos 502,024 y no fueron válidos 24,956.

Para el indicador por hora se realizó el agrupamiento por hora. Del total de 8,784 horas que tiene el año y las horas válidas de los datos monitoreados fueron 7844 quedando un total de 940 que no fueron validos por contar con menos de 45 registros en cada hora.

Para el indicador por día se realizó el agrupamiento por 24 horas o día. Del total de 366 días se validaron 329 días, quedando solo 37 días que no cumplieron con tener menos 18 horas válidas en el día.

Con los datos diarios se obtuvo la siguiente distribución de los días en el año con calidad de aire buena, regular y mala de acuerdo con la norma mexicana vigente tal y como se muestran en la Gráfico 1:

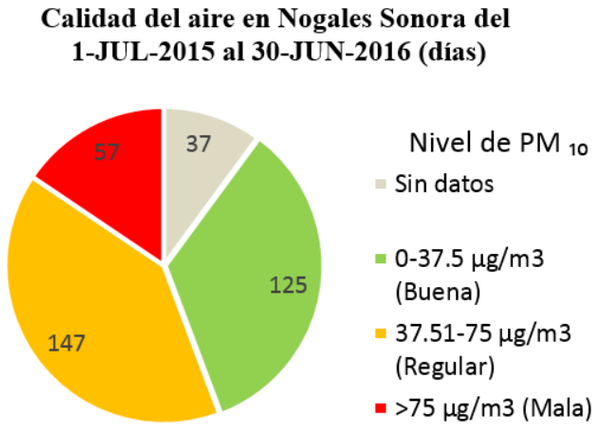

Gráfico 1 Indicador por dia

Por último se encontró que la ciudad de Nogales Sonora rebasa el promedio de contaminación anual de acuerdo a la norma mexicana vigente que es de 40 ug/m^3 resultando un promedio anual de 50.75 ug/m^3 del 1 de julio del 2015 al 30 de junio del 2016 como se muestra en la Gráfico 2.

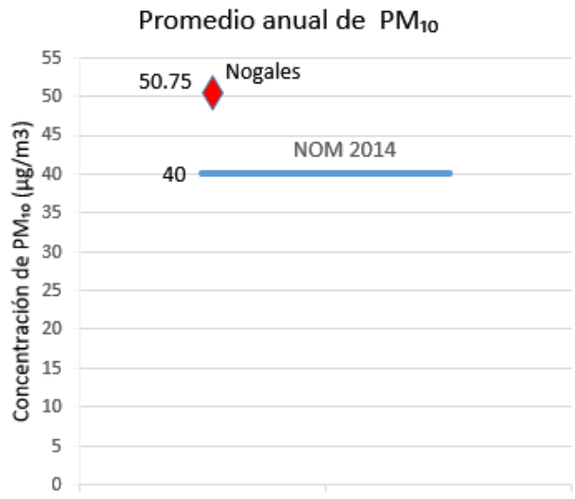

Gráfico 2 Nivel de PM_{10} anual

4. Conclusiones

Se comprobó con datos reales el desempeño de la implementación del ETL para verificar y validar los datos por minuto y su uso para generar los indicadores por hora y día para PM_{10}, de Nogales, Sonora, México. Los datos anuales resultaron ser válidos al obtener el 89.89%, que es más del 75% que marca la NOM mexicana de salud ambiental vigente. Así mismo los días con buena calidad del aire, entre el 1 de julio del 2015 y 30 de junio del 2016 fueron el 34.15%, los de calidad del aire regular fueron el 40.16%, los de mala calidad del aire fueron el 15.57% y los días que no se obtuvieron suficientes datos fueron el 10.10%.

ISSN-2410-3993

Finalmente, los datos que se obtuvieron tambien son la base para el desarrollo de otros 2 proyectos de minería de datos utilizando redes neuronales y correlación entre datos meteorológicos y PM_{10}.

5. Agradecimiento

Se agradece al Tecnológico Nacional de México (TecNM) porque el desarrollo del presente artículo es uno de los productos académicos del proyecto financiado y realizado en el Instituto Tecnológico de Nogales, el cual se tituló "Desarrollo de un ETL para datos de PM10 de la estación de calidad del aire de Nogales Sonora", con número de proyecto 5802.16-P y terminado en junio del 2017. También se agradece a la Comisión de Ecología y Desarrollo Sustentable del Estado de Sonora (CEDES) por compartir los datos de la estación de calidad del aire instalada en Nogales, Sonora, así mismo a Arizona Department of Environmental Quality (ADEQ), especialmente a José M. Rodríguez por su apoyo y a la Diputada por la LXIII Legislatura de Sonora, Lic. Leticia Amparano Gámez, por haber pagado la publicación de este artículo; así como también a la Delegación Sindical D-V-99 del ITN por el apoyo en la gestión del mismo.

6. Referencias

Connolly Tomas, B. C. (2015). *Database Systems A prectical Approach to Design, Implementation and Management.* Inglaterra: Pearson.

DOF. (2012, Julio 16). *Diario Oficial de la Federación.* Obtenido del Diario Oficial de la Federación: http://sinaica.inecc.gob.mx/archivo/noms/NOM-156-SEMARNAT-2012.pdf

DOF. (2014, Agosto 20). *Diario Oficial de la Federación.* Obtenido del Diario Oficial de la Federación: http://www.dof.gob.mx/nota_detalle.php?codigo=5357042&fecha=20/08/2014

EPA. (2017, 01 31). *AirNow*. Obtenido de AirNow: https://cfpub.epa.gov/airnow/index.cfm?action=aqibasics.particle

INECC. (2016, 11 15). *INFORME 2015.* Obtenido del INFORME 2015: http://sinaica.inecc.gob.mx/archivo/informes/Informe2015.pdf

LT Consoulting. (2016). Desarrollo del ProAire. *Avances del diagnóstico y medidas y acciones* (pp. 3-5). Nogales Sonora: ProAire.

OMS. (2016, 09 27). *Organizacion Mundial de la Salud*. Obtenido de la Organizacion Mundial de la Salud: https://cfpub.epa.gov/airnow/index.cfm?action=aqibasics.particle

SINAICA. (2010, 04 11). *Manual 5 Protocolo de manejo de datos de la calidad del aire.* Obtenido del Manual 5 Protocolo de manejo de datos de la calidad del aire: http://sinaica.inecc.gob.mx/archivo/guias/5%20-%20Protocolo%20de%20Manejo%20de%20Datos%20de%20la%20Calidad%20del%20Aire.pdf