FICHA TÉCNICA EXTRACTOR DE DATOS DE CALIDAD DEL AIRE

1. NOMBRE DEL PROYECTO: Observatorio de datos para descubrimientos de patrones Sociales-EspacioTemporales en Salud, Movilidad y Calidad del Aire.

2. Número: 7051.

3. OBJETIVO: Definir las bases de interoperabilidad para homologar la información de las bases de datos de salud, y obtener datos precisos y confiables, que permitan tener series históricas y generar la trazabilidad de los pacientes, en especial de aquellos que padecen enfermedades crónicas no transmisibles, así como la construcción de indicadores de salud y su relación con la exposición a contaminantes atmosféricos.

4. OBJETIVO GENERAL DEL PROYECTO
   Observatorio de datos para descubrimientos de patrones Sociales-EspacioTemporales en Salud, Movilidad y Calidad del Aire .

5. PROPÓSITO DEL RECURSO
   Instalación de Python a través del IDE Anaconda para procesar los datos obtenidos por las estaciones meteorológicas y obtener la correlación existente entre la concentración de estos contaminantes y el número de defunciones ocurridas por cáncer en la Ciudad de México y el Área Metropolitana. De igual manera se utiliza el programa QGIS como sistema de Información Geográfica para el análisis geoestadístico de las defunciones por cáncer en relación a su ocurrencia geográfica y el nivel de concentración de contaminantes en ese mismo punto. Es importante mencionar que la información georreferenciada de las defunciones ocurridas por cáncer fue modificada de sus coordenadas originales por cuestiones de confidencialidad.

6. INVESTIGADOR PRINCIPAL A CARGO
   DR. Miguel Félix Mata.

7. RECOLECCIÓN / OBTENCIÓN DE LOS DATOS

   La información recopilada por el Extractor de Datos de Calidad del Aire desde la página Web de la SEDEMA, se puede realizar para cualquier archivo CSV contenido en esa Web desde la fecha más antigua hasta la fecha más reciente.

8. PERIODO DE RECOLECCIÓN / OBTENCIÓN DE DATOS

El Extractor de Datos de Calidad del Aire puede descargar todos los archivos CSV con las mediciones obtenidas por las estaciones meteorológicas. Por cuestiones de tamaño de los archivos, para este recurso se utilizan sólo las mediciones de los años 2019 y 2020.

9. VARIABLES INCLUIDAS

Archivos descargados por el extractor en formato CSV.

1. cat_estacion – Contiene información relacionada con todas las estaciones meteorológicas a cargo de la Secretaría del Medio Ambiente de la Ciudad de México. Sus columnas son:
   a) id_station - Campo con el acrónimo de la estación meteorológica. Sirve para su identificación en otras tablas.
   b) nom_estac – Campo con el nombre completo de la estación meteorológica.
   c) longitud – Contiene información para su ubicación espacial en grados decimales.
   d) latitud – Contiene información para su ubicación espacial en grados decimales.
   e) alt – Información para su ubicación espacial, altitud en metros sobre el nivel del mar.
   f) obs_estac – Contiene información que se considere relevante en relación a la estación meteorológica.

2. cat_parametros – Contiene información relacionada con todos los elementos tanto contaminantes como variables meteorológicas que pueden ser medidos por las estaciones meteorológicas. Sus columnas son:
   a) id_parameter – Campo con numeración consecutiva.
   b) cve_param – Campo con el acrónimo de la estación meteorológica. Sirve para su identificación con otras tablas.
   c) nom_param – Campo con el nombre completo del elemento a medir.
   d) unidades_param – Campo con valor numérico para relacionar la unidad de medida del parámetro con el archivo 'cat_unidades'.

3. cat_unidades – Contiene información relacionada con las unidades de medida de cada uno de los elementos que pueden ser medidos por las estaciones meteorológicas. Sus columnas son:
   e) id_unidad – Campo con numeración consecutiva.
   f) Clave_undid – Acrónimo o símbolo científico utilizado para su identificación.
   g) nombre_unidad – Campo con el nombre completo para identificar la unidad de medición.

4. contaminantes_AAAA - Contiene el listado de las mediciones hechas por las distintas estaciones meteorológicas pertenecientes a la Red Automática del Medio Ambiente para el periodo AAAA (año de la medición). Sus columnas son:

a) date- Contiene la fecha de captura de la medición.

b) id_station – Campo con el acrónimo de la estación meteorológica. Sirve para su identificación.
c) id_parameter - Campo con el acrónimo del elemento medido. Sirve para su identificación.
d) value – Campo con el valor de la medición.
e) unit – Campo para identificar la unidad de medida.

Archivos utilizados y generados en Python.

01_web_crawler.ipynb – Código en Python para descargar de manera automática todos los archivos CSV vía Web los cuales son generados por las mediciones hechas por las estaciones meteorológicas a cargo de la Secretaría del Medio Ambiente de la CDMX.

02_quitar_nulos.ipynb – Código en Python el cual contiene una función de nombre "mediasPorMes(archivo)" la cual recibe el nombre de un archivo csv y con el uso de la librería de pandas se realiza la manipulación y análisis del contenido del archivo enviado como parámetro. Esta función agrupa las filas contenidas en el archivo por fecha, estación y elemento medido para asignarle la media de las mediciones a aquellos valores faltantes en el archivo. El resultado es otro archivo CSV con el mismo nombre del archivo de entrada pero con el prefijo "noNulls". Se debe ejecutar esta función por cada archivo que queramos procesar.

03_geolocalizar_estaciones_y_valores.ipynb – Código en Python con el cual relacionamos el archivo del código anterior en el que se obtuvo el archivo concentrado de las mediciones por el rango de años que se haya establecido, con el archivo 'cat_estacion.csv' para poder asignar las coordenadas geográficas de cada estación las cuales están contenidas en el catálogo de estaciones ('cat_estacion.csv). De esta manera se obtiene un archivo 'estacsMedia_Geo_pivoted_NO_Nulls.csv' el cual contiene el listado de estaciones con los valores promedio por cada contaminante que midió en el rango de tiempo elegido así como la ubicación geográfica (latitud y longitud) de cada estación meteorológica.

04_Correlation – Código en Python con el cual podemos leer el archivo 'salu_nd_pob_def_cancer_ratio_voron_galindo_zm_a' el cual se obtuvo después de procesar la información de contaminantes y defunciones por cáncer en QGIS para que podamos encontrar la correlación entre la concentración de contaminantes y la densidad de mortalidad en el mismo espacio geográfico.

Archivos generados en QGIS.

1. salu_nd_def_cancer_galindo_zm_p.gpkg – archivo de puntos georreferenciados con referencia espacial 4326 WGS84 que contiene la ubicación simulada de defunciones por cáncer del periodo 2015-2020. Sus columnas son:

a) fid – campo numérico consecutivo para identificación del punto.
b) lat_RES – campo de tipo real con la información espacial de latitud en formato decimal.

c) Lon_RES – campo de tipo real con la información espacial de longitud en formato decimal.

2. salu_nd_estac_meteo_galindo_zm_p gpkg – archivo de puntos georreferenciados con referencia espacial 4326 WGS84 que contiene la ubicación de las estaciones meteorológicas pertenecientes a la Secretaría del Medio Ambiente de la Ciudad de México. Sus columnas son:

a) fid – campo numérico consecutivo para identificación del punto.
b) id_station – campo de texto de 3 caracteres con el acrónimo del nombre de estación meteorológica.
c) nom_estac – campo de tipo texto con el nombre completo de la estación meteorológica.
d) latitud – campo de tipo real con la información espacial de latitud en formato decimal.
e) longitud – campo de tipo real con la información espacial de longitud en formato decimal.
f) CO – Campo de tipo real con los valores de concentración media del contaminante CO (monóxido de carbono) para cada una de las estaciones meteorológicas.
g) NO – Campo de tipo real con los valores de concentración media del contaminante NO (monóxido de nitrógeno) para cada una de las estaciones meteorológicas.
h) NO2 – Campo de tipo real con los valores de concentración media del contaminante NO2 (dióxido de nitrógeno) para cada una de las estaciones meteorológicas.
i) NOX - Campo de tipo real con los valores de concentración media del contaminante NOX (óxidos de nitrógeno) para cada una de las estaciones meteorológicas.
j) O3 – Campo de tipo real con los valores de concentración media del contaminante O3 (ozono) para cada una de las estaciones meteorológicas.
k) PM10 – Campo de tipo real con los valores de concentración media del contaminante PM10 (partículas menores a 10 micras) para cada una de las estaciones meteorológicas.
l) PM25 – Campo de tipo real con los valores de concentración media del contaminante PM25 (partículas menores a 2.5 micras) para cada una de las estaciones meteorológicas.
m) PMCO – Campo de tipo real con los valores de concentración media del contaminante PMCO (partículas coarse) para cada una de las estaciones meteorológicas.
n) SO2 – Campo de tipo real con los valores de concentración media del contaminante SO2 (dióxido de azufre) para cada una de las estaciones meteorológicas.

3. salu_nd_pob_def_cancer_ratio_voron_galindo_zm_a.gpkg - archivo de polígnos georreferenciados con referencia espacial 4326 WGS84 que contiene polígonos de Thiessen correspondientes al área de influencia de cada una de las estaciones meteorológicas pertenecientes a la Secretaría del Medio Ambiente de la Ciudad de México. Cada polígono contiene información de las concentraciones de contaminantes dentro de su área de influencia, así como la población estimada dentro de cada polígono, el número de defunciones ocurridas por cáncer dentro de ese polígono y su proporción en relación a la población. Sus columnas son:

a) fid – campo numérico consecutivo para identificación del punto.
b) id_station – campo de texto de 3 caracteres con el acrónimo del nombre de estación meteorológica.

c) nom_estac – campo de tipo texto con el nombre completo de la estación meteorológica.
d) latitud – campo de tipo real con la información espacial de latitud en formato decimal.
e) longitud – campo de tipo real con la información espacial de longitud en formato decimal.
f) CO – Campo de tipo real con los valores de concentración media del contaminante CO (monóxido de carbono) para cada una de las estaciones meteorológicas.
g) NO – Campo de tipo real con los valores de concentración media del contaminante NO (monóxido de nitrógeno) para cada una de las estaciones meteorológicas.
h) NO2 – Campo de tipo real con los valores de concentración media del contaminante NO2 (dióxido de nitrógeno) para cada una de las estaciones meteorológicas.
i) NOX - Campo de tipo real con los valores de concentración media del contaminante NOX (óxidos de nitrógeno) para cada una de las estaciones meteorológicas.
j) O3 – Campo de tipo real con los valores de concentración media del contaminante O3 (ozono) para cada una de las estaciones meteorológicas.
k) PM10 – Campo de tipo real con los valores de concentración media del contaminante PM10 (partículas menores a 10 micras) para cada una de las estaciones meteorológicas.
l) PM25 – Campo de tipo real con los valores de concentración media del contaminante PM25 (partículas menores a 2.5 micras) para cada una de las estaciones meteorológicas.
m) PMCO – Campo de tipo real con los valores de concentración media del contaminante PMCO (partículas coarse) para cada una de las estaciones meteorológicas.
n) SO2 – Campo de tipo real con los valores de concentración media del contaminante SO2 (dióxido de azufre) para cada una de las estaciones meteorológicas.
o) población – campo de tipo entero con el número de habitantes en el polígono.
p) Defunciones – campo de tipo real con el número de defunciones ocurridas dentro de cada polígono.
q) ratio – campo de tipo real con la proporción de defunciones con respecto a la población de cada polígono.

4. salu_nd_pob_def_cancer_voron_galindo_zm_a.gpkg - archivo de polígonos georreferenciados con referencia espacial 4326 WGS84 que contiene polígonos de Thiessen correspondientes al área de influencia de cada una de las estaciones meteorológicas pertenecientes a la Secretaría del Medio Ambiente de la Ciudad de México. Cada polígono contiene información de las concentraciones de contaminantes dentro de su área de influencia, así como la población estimada dentro de cada polígono. Sus columnas son:

a) fid – campo numérico consecutivo para identificación del punto.
b) id_station – campo de texto de 3 caracteres con el acrónimo del nombre de estación meteorológica.
c) nom_estac – campo de tipo texto con el nombre completo de la estación meteorológica.
d) latitud – campo de tipo real con la información espacial de latitud en formato decimal.
e) longitud – campo de tipo real con la información espacial de longitud en formato decimal.

f) CO – Campo de tipo real con los valores de concentración media del contaminante CO (monóxido de carbono) para cada una de las estaciones meteorológicas.
g) NO – Campo de tipo real con los valores de concentración media del contaminante NO (monóxido de nitrógeno) para cada una de las estaciones meteorológicas.
h) NO2 – Campo de tipo real con los valores de concentración media del contaminante NO2 (dióxido de nitrógeno) para cada una de las estaciones meteorológicas.
i) NOX - Campo de tipo real con los valores de concentración media del contaminante NOX (óxidos de nitrógeno) para cada una de las estaciones meteorológicas.
j) O3 – Campo de tipo real con los valores de concentración media del contaminante O3 (ozono) para cada una de las estaciones meteorológicas.
k) PM10 – Campo de tipo real con los valores de concentración media del contaminante PM10 (partículas menores a 10 micras) para cada una de las estaciones meteorológicas.
l) PM25 – Campo de tipo real con los valores de concentración media del contaminante PM25 (partículas menores a 2.5 micras) para cada una de las estaciones meteorológicas.
m) PMCO – Campo de tipo real con los valores de concentración media del contaminante PMCO (partículas coarse) para cada una de las estaciones meteorológicas.
n) SO2 – Campo de tipo real con los valores de concentración media del contaminante SO2 (dióxido de azufre) para cada una de las estaciones meteorológicas.
o) población – campo de tipo entero con el número de habitantes en el polígono.

5. salu_nd_poligons_voron_galindo_zm_a.gpkg - archivo de polígonos georreferenciados con referencia espacial 4326 WGS84 que contiene polígonos de Thiessen correspondientes al área de influencia de cada una de las estaciones meteorológicas pertenecientes a la Secretaría del Medio Ambiente de la Ciudad de México. Cada polígono contiene información de las concentraciones de contaminantes dentro de su área de influencia. Sus columnas son:

a) fid – campo numérico consecutivo para identificación del punto.
b) id_station – campo de texto de 3 caracteres con el acrónimo del nombre de estación meteorológica.
c) nom_estac – campo de tipo texto con el nombre completo de la estación meteorológica.
d) latitud – campo de tipo real con la información espacial de latitud en formato decimal.
e) longitud – campo de tipo real con la información espacial de longitud en formato decimal.
f) CO – Campo de tipo real con los valores de concentración media del contaminante CO (monóxido de carbono) para cada una de las estaciones meteorológicas.
g) NO – Campo de tipo real con los valores de concentración media del contaminante NO (monóxido de nitrógeno) para cada una de las estaciones meteorológicas.
h) NO2 – Campo de tipo real con los valores de concentración media del contaminante NO2 (dióxido de nitrógeno) para cada una de las estaciones meteorológicas.
i) NOX - Campo de tipo real con los valores de concentración media del contaminante NOX (óxidos de nitrógeno) para cada una de las estaciones meteorológicas.

j) O3 – Campo de tipo real con los valores de concentración media del contaminante O3 (ozono) para cada una de las estaciones meteorológicas.
k) PM10 – Campo de tipo real con los valores de concentración media del contaminante PM10 (partículas menores a 10 micras) para cada una de las estaciones meteorológicas.
l) PM25 – Campo de tipo real con los valores de concentración media del contaminante PM25 (partículas menores a 2.5 micras) para cada una de las estaciones meteorológicas.
m) PMCO – Campo de tipo real con los valores de concentración media del contaminante PMCO (partículas coarse) para cada una de las estaciones meteorológicas.
n) SO2 – Campo de tipo real con los valores de concentración media del contaminante SO2 (dióxido de azufre) para cada una de las estaciones meteorológicas.

6. salu_nd_pob_manzana_galindo_zm_a.gpkg - archivo de polígonos georreferenciados con referencia espacial 4326 WGS84 que contiene entidades poligonales con información de la población a nivel de manzana obtenida del SCINCE (Sistema de Consulta de Información Censal) año 2020. Sus columnas son:

a) fid – campo numérico consecutivo para identificación del punto.
b) CVEGEO – campo de tipo cadena de caracteres con la información de la clave geoestadística de cada manzana en la Ciudad de México y zona metropolitana del Valle de México.
c) POB1 – campo de tipo entero con la cantidad de habitantes por cada manzana.

10. ESTRATEGIA DE ASEGURAMIENTO PARA LA PROYECCIÓN DE DATOS SENSIBLES / PERSONALES

De la información proporcionada por la Secretaría de Salud de la Ciudad de México, se eliminó toda información personal por cuestiones de confidencialidad por lo que se utilizó información generada aleatoriamente que simulara las defunciones a nivel de dirección de residencia. Con respecto a la información de contaminantes, ésta se encuentra disponible para su descarga desde la página Web de la Secretaría del Medio Ambiente de la Ciudad de México.

11. FECHA ÚLTIMA DE ACTUALIZACIÓN

La información de defunciones proporcionada por la Secretaría de Salud de la Ciudad de México para el periodo 2015-2020 pero esta fue modificada por cuestiones de confidencialidad. La lectura de contaminantes fue hasta el mes de agosto de 2021 pero se puede ejecutar el Extractor de Datos de Calidad del Aire para obtener las lecturas del día anterior a su ejecución, hasta el registro más antiguo.

12. CONTROLES PARA LA VALIDACIÓN Y VERIFICACIÓN DE LA CAPTURA/REGISTRO DE LOS DATOS.

    La información de la base de datos fue previamente procesada mediante técnicas de procesos ETL (Extracción Información Carga) para validar y verificar el registro de la información.

13. OTRAS PLATAFORMAS DONDE SE ENCUENTRA DISPONIBLE EL RECURSO DE INFORMACIÓN.

    El presente recurso no se encuentra en ninguna otra plataforma.

14. OTRAS FUENTES DE FINANCIAMIENTO

    Sin otras fuentes de financiamiento

15. SEGUIMIENTO DE LA COHORTE EN ESTUDIO.

    Este recurso por su naturaleza es un estudio retrospectivo por lo cual no existe seguimiento de la Cohorte en estudio.

16. PUBLICACIONES

    Sin publicaciones.

17. OTRA INFORMACIÓN

    Es importante mencionar que para poder subir la información a la Plataforma de CONACYT fue necesario reducir lo más posible el número de registros en las bases de datos razón por la cual los resultados presentados no necesariamente serán los mismos que los que se obtuvieron como resultado final de la investigación.