Using weather and pollutant time series to analyze respiratory hospital admissions trends in South America

A Data Management Plan created using DMPTool

Creators: Eduardo Germani, Eduardo Germani

Affiliation: University of São Paulo (USP)

Template: Digital Curation Centre (DCC)

ORCID iD: 0000-0003-2548-9685

Project abstract:

Weather and pollutant concentration variations may have direct impact on human health. Subtle changes in weather or atmospheric conditions may increase hospital admissions for treatment on respiratory diseases like pneumonia or asthma, especially in risk groups such as elderly and children. This article proposes to analyze weather and pollutant concentration changes in the last ten years and add data from hospital admissions to identify patterns and trends. Time series data on air quality and pollutant concentration from government records on public hospitals are analyzed with big data technologies support and geographically visualized to provide researchers new insights. Also, regression models are proposed to predict new respiratory hospital admissions outcomes based on climate patterns

Last modified: 12-10-2018

Copyright information:

Using weather and pollutant time series to analyze respiratory hospital admissions trends in South America

Data Collection

Consolidation of 2 datasets

- CETESB - Brazilian Government Institution which provides data on air quality from a major city in Brazil, Guarulhos/Sao Paulo State.
    - Data Format: CSV - Needs consolidation of multiple pollutant into single table
- SIH-SUS - Public system specialized on hospital admissions that provides data such as diseases (ICD codes), and demographic data such as date of admission, date of death, age from pacients and hospital information.
    - Data Format: DBC - Needs convertion to csv and consolidation with CETESB data

Software used

- Apache Hadoop - for data consolidation and summarization
- MS Excel - For computing moving averages on consolidated file

- Each file imported from government institution websites will be holded in separated folders identifying institution. Besides, each file will identify institution and searched period.
- For versioning purposes, files will be appended with version number. A cloud repository such as GitHub may be used to control versioning.
- CETESB - Institutional web site - https://qualar.cetesb.sp.gov.br
    - Files will be renamed to concatenate institution, observational period, stationName and pollutant type and collection date timeStamp
        - Example:
            - CETESB_GuarulhosPaco_O3_JanDez2017_v1
            - CETESB_GuarulhosPaco_MP10_JanDez2017_v2
- SIH-SUS - Institutional web site - http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=25
    - SIH-SUS files are montly based, so files are provided without full information.
    - Files will be renamed to concatenate institution, month and year. Exemples:
        - SIHSUS_0117_v1
        - SIHSUS_0217_v2
- Quality Assurance Measures
    - Revision of data to deal with missing and noise
    - Maintain separated files for individual pollutant measures and cross-checking with consolidated file

Documentation and Metadata

To understand the data, information on pollutant type, pollutant concentration, pollutant unity and date and hour of measurement is needed. This data is already presented in the columns on csv.

The Ecological Metadata Language may be used and metadata can be generated with a metadata tool such as Morpho.

Ethics and Legal Compliance

- Access to data: data is already public and anonymized, so no legal rights are needed. There are no individual identification on medical records other than date of birth, death, ICD code and hospital.

- IP/IPR - As a principal investigator, i own the data. The data is free to use but citation on the original data is required.
- Data will be postponed to article publication acceptance.

Storage and Backup

Data will be stored in 2 sites.

- Local computer
- Company Servers with redundancy and backup routines
- The backup is already made in a daily manner for each sql server node.
- In case of incident, data can be recovered through restore from old files or tape disks.

- Data Security - Access is public due to be a public datasets.

Selection and Preservation

No data should be destroyed due to all data be public.

All data will be preserved for future comparisons and other countries benchmark.

Future research may include time series comparisons in different countries.

Data wil be retained in company data center, so it should be retained for long-term with no additional costs.

Alternatively, data may be sent to external repository such as datadryad.org with costs to the company.

Data Sharing

A global repository may be used such as datadryad.org. This worldwide website includes datasets from different researches. Data will be able to be referenced via DOI number only after article publication.

Exclusive access is needed only for six months. Then data will be shared for any research purposes. A citation to original article and dataset is mandatory.

Responsibilities and Resources

As a principal investigator with no partners, i will be responsible for implementing dmp and data management.

No external expertise are needed. All the hardware and software are already licensed and no renewal is necessary. Some charges may be applied to public repositories access such as datadryad.org.