

---

## Plan Overview

*A Data Management Plan created using DMPTool*

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

**Creator:** Fernando Fernandes

**Affiliation:** Non Partner Institution

**Funder:** Digital Curation Centre (dcc.ac.uk)

**Funding opportunity number:** 36290

**Template:** Digital Curation Centre

### **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:**

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

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

---

## Data Collection

### What data will you collect or create?

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 patients and hospital information.
  - Data Format: DBC - Needs conversion 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

### How will the data be collected or created?

- Each file imported from government institution websites will be held 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 monthly based, so files are provided without full information.
  - Files will be renamed to concatenate institution, month and year. Examples:
    - 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

### What documentation and metadata will accompany the data?

To understand the data, information on pollutant type, pollutant concentration, pollutant unit 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

### How will you manage any ethical issues?

- 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.

### How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

- 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

### How will the data be stored and backed up during the research?

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.

#### **How will you manage access and security?**

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

#### **Selection and Preservation**

##### **Which data are of long-term value and should be retained, shared, and/or preserved?**

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.

##### **What is the long-term preservation plan for the dataset?**

Data will 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**

##### **How will you share the data?**

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.

##### **Are any restrictions on data sharing required?**

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**

##### **Who will be responsible for data management?**

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

##### **What resources will you require to deliver your plan?**

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.