Observing the correlation between the yearly average temperature of Innsbruck and the pollution emissions of NMVOCs from transport in Austria

A Data Management Plan created using DMPonline

Creator: Michael Aigner

Affiliation: Other

Template: European Commission (Horizon 2020)

ORCID iD: https://orcid.org/0000-0002-4872-9154

Project abstract:

In this experiment, the question if the emissions of non-methane volatile organic compounds from transport in Austria have an influence on the yearly average temperature in Innsbruck will be treated. This is done by comparing the trends of transport emissions and average temperature, as well as obtaining the correlation between those two variables.

Last modified: 21-04-2019

Observing the correlation between the yearly average temperature of Innsbruck and the pollution emissions of NMVOCs from transport in Austria - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The purpose of the data collection is get data that contains information neccessary for calculating the average yearly temperature of Innsbruck, as well as an emission indicator from transport regarding the emission of non-methane volatile organic compounds

Data generated is used for having a combined view on the average yearly temperature and on the emission indicator for each year in a time interval from 1990 to 2016.

Explain the relation to the objectives of the project

The relation of data collection and generation to the objectives is to answer the question if the emissions of non-methane volatile organic compounds from transport in Austria have an influence on the yearly average temperature in Innsbruck will be treated. This is done by comparing the trends of transport emissions and average temperature, as well as obtaining the correlation between those two variables.

Specify the types and formats of data generated/collected

Reusable data collected

Two different data contents, the average yearly temperature of Innsbruck and the pollution emission indicator of non-methane volatile organic compounds from transport are in use.

Temperature

Data description: For the yearly average temperature of Innsbruck, data provided on the Open Data Austria (https://www.data.gv.at/) will be used. The data set is available under https://www.data.gv.at/katalog/dataset/5eb8278a-4ecf-41e2-a1f8-03383f31af7d, acessed on 20.04.2019. The content is the monthly and yearly average temperature measured in Innsbruck from the years 1971 to 2016.

Data format: The data is stored as a comma-separated values (.csv) file, ensuring reusability since it it a non-proprietary format.

Transport

Data description: For the transport pollution emission indicator, data provided by the EU Open Data Portal (http://data.europa.eu/euodp/en/home) will be used. The data set is available under http://data.europa.eu/euodp/en/data/dataset/gZmNXFTZrjPyK3EHPykmpg, accessed on 20.04.2019). This data contains the indicator of the European Union for emissions produced by transport, the main contributor to air pollution. It includes emissions from nitrogen oxides (NOx), non-methane volatile organic compounds (NMVOCs) and particulate matter (PM10) and is an index to the year 2000 (indicator = 100 for this year). The values are reported under the UNECE Convention on Long-Range Transboundary Air Pollution (CLRTAP). The data contains the three emission measurements for each EU country for the years 1990 to 2016. In the experiment, we will just focus on the Austrian data for non-methane volatile organic compounds (NMVOCs).

Data format: The data provided is stored tab-separated values (.tsv) file, also ensuring convinient reproducibility in a non-proprietary format like using .csv files.

Data created

Analysis data

Data description: This output contains the combined data from the average temperature and per year from Inssbruck and the EU pollution index for NMVOCs of Austria. Each row contains comma-separated values (year;avg yearly temperature;pollution indicator for NMVOCs). The temperature is measured in Celsius degrees. The pollution indicator shows the value of the EU pollution index for NMVOCs (based on year 200, where indicator=100).
Data format: The data will be stored as a comma-separated values (.csv) file. The format is chosen since it is a non-proprietary, open format for storing data. Also, the resulting file size is smaller than for e.g. XML-files.

Figures

Data description: Figures produced by the experiment are a heatmap showing the correlation of year, average temperature and pollution indicator, a line chart showing standardized values of temperature and pollution and another one showing the trend of those values.
Data format: The data will be stored as JPEG files, to make it easier to work with them on different systems and also save disc space.

Specify if existing data is being re-used (if any)

The data for the average yearly temperature in Innsbruck and the pollution emission indicator of non-methane volatile organic compounds from transport are already existing and obtained from the data providers Open Data Austria and the EU Open Data Portal, as already described in the last section.

Both data sets are available under the CC-BY licence, so the reuseage of the data is permitted.

Specify the origin of the data

The emmissions from transport indicator data is officially reported under the UNECE Convention on Long-Range Transboundary Air Pollution (CLRTAP) .

The data about the temperature of Innsbruck was published by the city of Innsbruck.

State the expected size of the data (if known)

The size of the two source data sets is around 20KB.
The output data needs around 60KB of free disc space.
Due to this small size, no additional costs or challenges are expected and therefore no further solutions dealing with the data volume need to be adressed.

Outline the data utility: to whom will it be useful

The resulting data of the experiment aims to show a possible relation between transport emissions and the average temperature per year, useful for decision makers for transport regulations.

New, forseeable research uses for the data are reusing the provided source code and data extended with newer versions of the collected data sets, to obtain changes in the influence of emissions and the change of the average temperature over future time periods. Also, one can think of observing correlations between temperature and other kinds of emissions, which are already provided in the input data and would not need any significant changes in the source code.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

The data collected contains meta data attached in the described origin of the resources.

For an easier interpretation of the data in the future, a description of the data produced is provided under /documentation/description.txt in the repository. Every output file, as well as the used measurements and units, plus the scales of the figures produces are described there. Units and data used have already been described in the previous section.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

As a persisent identifier of the data, a Digital Object Identifier (DOI) will be used so that people can find the data, as well data citation and reuse can be tracked.
A DOI will be added for the source code and every of the resulting output files. The DOIs are added to the repository for this experiment.

Combined data output: https://doi.org/10.5281/zenodo.2648172

Source code: https://doi.org/10.5281/zenodo.2648176

Outline naming conventions used

The names of all produced files just contain alphanumerical lowercase characters and underscores. E.g. figures are named temparture_pollution_*.jpg, where * stands for the type of plot shown in the figure.

Outline the approach towards search keyword

The repository containing the data can be searched over keywords as already integrated in GitHub.

Outline the approach for clear versioning

For clear versioning, GitHub will be in use. This is achieved by describing every change made in a new version with appropriate comments and making new releases for versions with major changes.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Meta data is stored under /documentation/metadata.xml and contains the following information:

- Title of the experiment
- Creator
- Subject
- A more detailed description of the experiment
- Date when the experiment was executed
- Type/purpose of the application

Format
- Sources used
- Language
- Years covered by the experiment dtaa
- Rights for using the data / sourcecode (we provide free access here)

As a standard for writing the meta data, the Dublin Core has been in use, since it is a basic, domain-agnostic standard which can be easily understood and implemented. The meta data as well as the description file will be created by hand.

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

Every part of the project, the source code, input data as well as outputs created are freely available for everybody.

Specify how the data will be made available

The data will be held in a GitHub repository, publically accessible overhttps://github.com/MBAigner/Correlation-Temperature-TransportPollution/, since it is wide spread. It is chosen because the site ensures a free use, until a memory usage of 1GB, that will not be reached with the data of this project, so additional charges will not be expected within the
next years.

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

For accessing the data, just a simple text editor and image viewer are needed (since the output s are in the .csv and .jpg format).

For a reproduction of the experiment, R is needed. For the experiment the R version 3.5.1 was used.

Most of the functions used are already provided in R but additionally, the following R libraries have been in use and need to be installed if not already present in a system:

- ggcorrplot version 0.1.2
- ggplot2 version 2.2.1
- dplyr version 0.7.7
- tidyr 0.8.2

All used libraries are loaded in the source code, however if they are not installed this step needs to be done before executing the code.

Specify where the data and associated metadata, documentation and code are deposited

For an easier interpretation of the data in the future, a description of the data produced is provided under /documentation/description.txt in the repository. Every output file, as well as the used measurements and units, plus the scales of the figures produces are described there. Also, a graphical illustration of the workflow is stored under /documentation/architecture.txt.

The project has the following structure:

- src: Contains functions defined for loading and preprocessing the temperature and pollution data that are called in the main file analysis.R.
- data: Contains all input sources, more detailed descriped in the next section.
- output: Contains the final preprocessing version of the data we will work with.
- figures: Contains all plots produced by this experiment.
- analysis.R: This is the main file of this application, that executes every step needed for this experiment.

Specify how access will be provided in case there are any restrictions

The data will be held in a GitHub repository, publically accessible overhttps://github.com/MBAigner/Correlation-Temperature-TransportPollution/ without any restrictions.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

Data saved in non-proprietary formats like .csv, .tsv and .jpg in our case ensures interoperability for most systems. Also, with the Dublin Core for metadata a widespread vocabulary was used.

With the usage of R for data processing, an open source solution that also works on most common used systems was provided.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

A standard vocabulary for all data types present in the data sets is used. For meta data, the Dublin Core is used, as described above. Because of that, no further mapping is provided.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The data is licenced under the MIT licence that permits

- Commercial use,
- modification,
- distribution and
- private use,

but is limited regarding liability and warranty.

That means that there are no restrictions for reuse, planned that way because the experiment should be freely available for everybody. Also, the sharing of the data will not be postponed or restriced in anyway.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

The data is available for reuse since 20.04.2019. No embargo was needed, but no intermediate results and just final versions have been published to ensure correctness of the results.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

Both, the data produced and used are useable by third parties and published under the MIT licence. Reuse of data has no kind of any restriction.

Describe data quality assurance processes

The quality of the data is ensured by comparing them to measurements provided in other data sets for the same topic. Additionally, it was checked if the data contains any unrealistic outliers.
The data is documented by description files and meta data, stored under /documentation in the repository.

Specify the length of time for which the data will remain re-usable

The data will also be publically available in the GitHub repository for at least 5 years. Since the file sizes are small compared to other Data Science projects, no additional charges for the data repository are expected.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

No costs are expected for making the data FAIR, because just open source software was in use and because this is a not founded research project.

Clearly identify responsibilities for data management in your project

For every task, data capture, metadata production, data quality, storage and backup, data archiving and data sharing just the developer of this experiment (myself, Michael Benedikt Aigner https://orcid.org/0000-0002-4872-9154) will be responsible.

Describe costs and potential value of long term preservation

No additional costs will be expected for preparing the data for sharing and preservations, because just open source products are used for this research project. The data will also be publically available in the GitHub repository for at least 5 years. Since the file sizes are small compared to other Data Science projects, no additional charges for the data repository are expected. GitHub site ensures a free use, until a memory usage of 1GB, that will not be reached with the data of this project, so additional charges will not be expected within the next years.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

The chosen repository (GitHub) ensures a secure storing of the data by the following aspects:

- Two-factor Authentication (2FA) (SMS, TOTP)
- Git over Secure Shell (SSH) and HTTPS
- GPG commit-signing verification
- Security audit log

For versioning of the data, a connection over SSH is established beforehand to safely transfer the data.

Additionally to a local storing of the data, it is backed up by frequently commits to the repository done by the researcher.

For data recovery, it is possible to roll back the data to earlier save stages / versions.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Because the data collected and created contains no information about certain individuals, just averaged temperature values and the emission from transport indicator without reference to any human contributors or institutions causing the emissions, the data does not need to be anonymized. Therefore also no security issues of transferring and storing the data publically are addressed here.
Permissions for sharing and publishing the data are ensured since the data collected was published under the CC-BY licence.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

There are no such national/funder/sectorial/departmental procedures for data management that I am using .