Calculation of nice („sunny“) days based on small privately owned solar thermal collector systems

A Data Management Plan created using DMPonline

Creator: Roland Wallner

Affiliation: Other

Funder: European Commission (Horizon 2020)

Template: Horizon 2020 DMP

ORCID iD: 0000-0002-2932-9892

Project abstract:

The aim of the project is to develop an algorithm that is capable to calculate the amount of sunny („nice“) hours of a day based on data collected by solar thermal collectors used to heat up homes. To evaluate the algorithm we will use the data collected by an owner of a solar thermal collector system based in Vienna, Austria. The output of this evaluation should be a graph that tells us how many hours of the day could be considered as sunny with the plotted data in csv files to allow future analysis.

Last modified: 14-04-2020

Calculation of nice („sunny“) days based on small privately owned solar thermal collector systems - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The purpose of the experiment is to figure out how sensor data of solar thermal collectors can be used to gather fine-granular information about the amount of sunny hours in a specific region. This is highly relevant for farmers with land nearby. They can use the resulting data to manage their farm land better.

Explain the relation to the objectives of the project

Farming land is highly dependent on the weather. We use data provided by components that are capable to give information about how the sun is heating up water to infer the amount of sunny hours of a day. The plotted information should result in easier analysis.

Specify the types and formats of data generated/collected

The experiment produces aggregated data based on the input data, which is described in the next section. This aggregated data will be exported as csv files containing two columns (date and hours of sunny hours).

These files are also used to generate a graph in the PNG format. It plots the absolute number of hours of sun per day in the timespan that is given as input data.

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

The data used in this experiment is present in the form of sensor data that is stored in csv files aggregated by month. The available columns of the dataset found at https://github.com/ralf-saenger/sonnenkollektordaten are Datum (german for: date), Uhrzeit (german for: time of day), S1/2 - Solar (the temperature measured at a solar thermal panel) and S7/2 - Temp. Aussen (the air temperature measured in the area of the solar thermal panel).

The air temperature is not used in the experiment, but is contained in the referenced dataset.

The specific commit hash on github that is used for this experiment is: 2c13133776dec218fcd2d4c7999408d0a293f930

Accessed on: 14.04.2020

Specify the origin of the data

The data is generated by the control unit (type UVR1611) that is controlling a small solar thermal collector system. This system is located in Rodaun, Vienna.

The recorded data was exported and uploaded to github.com where it is released under the Creative Commons Zero v1.0 Univeral License. Because of this license we are allowed to use the data.

State the expected size of the data (if known)

The generated output csvs contain a row per day that is provided in the input data. If monthly files are used each csv output file will contain at most 32 rows (including the header) and 2 columns.

The resulting file size of each of these csv files will only be a few kilobytes.

Outline the data utility: to whom will it be useful

The data that is used and generated experiment can be used for two distinct purposes:

- Firstly, we can use the data to evaluate the algorithm used to transform it.
- Secondly, farmers might make decisions based on the result of this experiment. The data used in this experiment was recorded in Rodaun, Vienna. Therefore the results will also only be relevant to people in this area.

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

Outline the discoverability of data (metadata provision)

The data that is produced by this software should always contain the information where the input data was recorded and which time range is looked at. This allows other researchers to find data for a specific region and timespan more easily.

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?

The results of this experiment are available under: https://doi.org/10.5281/zenodo.3751578
The software that is used to generate the data is released under: https://doi.org/10.5281/zenodo.3755912

Outline naming conventions used

The software project does not follow any explicit convention, but common best-practices are applied. This is mainly resulting because of the size of the project, as the main contribution are two small python scripts.

These python scripts try to follow the PEP 8 coding standard, which is commonly used for python projects, as close as possible.

The dependencies are stored in a requirements.txt file that follows a common structure.

Furthermore the folder structure is relatively intuitive to use, as there is a "data" folder that can be used for storing data during execution of the program and a "docs" folder that contains documentation material referenced in the README.md

Outline the approach towards search keyword

No clear approach for using keywords was used during the experiment. There are two artifacts produced by this experiment. These have been annotated with keywords with the best intentions.

Outline the approach for clear versioning

The software uses semantic versioning and reached the version 1.0.0.

The software itself is agnostic to the versioning of the data. We only use one version of the input data that is also not supposed to change (its recorded data), so it is omitted in our result description. It is possible to use the software in a way that allows to work with versioned data by using appropriate filenames for the output data.

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

The input data is assigned to a specific timeframe and an area where it was recorded. This information should be added to the resulting data as metadata to allow others to find the relevant data more easily.

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

The data produced by the software during this experiment is openly published at https://doi.org/10.5281/zenodo.3751578 under the Creative Commons Attribution 4.0 International license. By releasing the data under this license everyone is allowed to reuse the resulting data and it is publicly available.

The software itself is hosted on github and archived at https://doi.org/10.5281/zenodo.3755912. The software is released under the MIT license which also completely unrestricted reuse of the software.

Data that is not disclosed is data used in the planning steps of the projects as well as project management information as it is not considered useful for external persons and might contain sensitive data.

Specify how the data will be made available

The code and resulting data of this experiment is hosted on zenodo.

For further development of the software it is also hosted on github. (https://github.com/roland-wallner/dss2020-ex1)

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)?

The resulting data can be downloaded with a normal web browser. To reproduce the results generated during this experiment you can follow the steps provided in the README.md file.

The dependencies that are required are:

- An operating system that is capable to run Python 3.8. (The documentation is written for Ubuntu 19.10, but should be usable for Windows environments with minor modifications)
- The software is written and tested in Python 3.8, but might work in other versions of Python as well.

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

The resulting data is hosted on zenodo.

The documentation and source code can be found on Github.

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

At this stage no restrictions are in place.

If the github repository management system would not host the source code the files will be uploaded to a similar service, such as bitbucket or gitlab.

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.

The generated data is stored as csv files, which is an easy to understand file format that tries to be as reusable as possible. The output graph is stored as an png file, which is also readable by all common operating systems and programs. The encoded data in the image should only be used for interpretation by humans. For further programmatic analysis the csv files should be used.

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?

The dataset does not use any complex data types besides dates. These are stored in the format YYYY-MM-DD which is also considered to be a standard.

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

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

Each published artifact generated by this experiment is released under an Open Access Licence.

For the result data Creative Commons Attribution 4.0 International is used.
For the source code the MIT license is used.

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

Published data is available right after the upload. No temporal restrictions are necessary.

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

The data is useable by third parties as it is. It is well documented and will be existent after the experiment is over.

Describe data quality assurance processes

Because of the small size of the software, no quality assurance process was defined.

The software was tested by manual evaluation.

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

The data will remain reusable until it is deleted by the currently hosting repositories and it is not possible to provide it on another platform.

3. Allocation of resources

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

The experiment is performed by a student that does not receive any money or other recompensation for his time. There is no additional cost for hosting the data. Therefore this will not cost anything.

Clearly identify responsibilities for data management in your project

This project is performed by a single person, so all responsibilities have to be taken on by the author of this document.

Describe costs and potential value of long term preservation

Currently the repositories used to publish the data and code can be done without cost.

If zenodo or github is not available as free hosters anymore and alternatives are also not cost free, these costs will arise. The chance of that happening is very low.

4. Data security

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

The data that is used and published in this project is not sensitive. Accessing and uploading the data is done with secure tunnels. By doing this we can ensure that the data gets to the repositories safely and unmodified.
Zenodo and Github have to be trusted providers.

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

No ethical issues will arise by this experiment or the resulting data.
Only open data with consideration of the respective licenses has been used, so no legal issues should arise either.

6. Other

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

No other procedures are in place.