repurpose Documentation

Release

Christoph Paulik

Mar 29, 2018

Contents

This package provides routines for the conversion of image formats to time series and vice versa. It is part of the poets° project and works best with the readers and writers supported there. The main use case is for data that is sampled irregularly in space or time. If you have data that is sampled in regular intervals then there are alternatives to this package which might be better for your use case. See *Alternatives* for more detail.

The readers and writers have to conform to the API specifications of the base classes defined in pygeobase to work without adpation.

CHAPTER 1

Citation

If you use the software in a publication then please cite it using the Zenodo DOI. Be aware that this badge links to the latest package version.

Please select your specific version at https://doi.org/10.5281/zenodo.593577 to get the DOI of that version. You should normally always use the DOI for the specific version of your record in citations. This is to ensure that other researchers can access the exact research artefact you used for reproducibility.

You can find additional information regarding DOI versioning at http://help.zenodo.org/#versioning

CHAPTER 2

Installation

This package should be installable through pip:

```
pip install repurpose
```

CHAPTER 3

Modules

It includes two main modules:

- `img2ts` for image/swath to time series conversion, including support for spatial resampling.
- `ts2img` for time series to image conversion, including support for temporal resampling. This module is very experimental at the moment.

CHAPTER 4

Alternatives

If you have data that can be represented as a 3D datacube then these projects might be better suited to your needs.

- PyReshaper is a package that works with NetCDF input and output and converts time slices into a time series representation.
- Climate Data Operators (CDO) can work with several input formats, stack them and change the chunking to allow time series optimized access. It assumes regular sampling in space and time as far as we know.
- netCDF Operators (NCO) are similar to CDO with a stronger focus on netCDF.

CHAPTER 5

Contribute

We are happy if you want to contribute. Please raise an issue explaining what is missing or if you find a bug. We will also gladly accept pull requests against our master branch for new features or bug fixes.

5.1 Development setup

For Development we recommend a `conda` environment

5.2 Guidelines

If you want to contribute please follow these steps:

- Fork the repurpose repository to your account
- make a new feature branch from the repurpose master branch
- Add your feature
- Please include tests for your contributions in one of the test directories. We use py.test so a simple function called test_my_feature is enough
- submit a pull request to our master branch

CHAPTER 6

Note

This project has been set up using PyScaffold 2.4.4. For details and usage information on PyScaffold see http://pyscaffold.readthedocs.org/.

CHAPTER 7

Contents

7.1 ts2img

7.1.1 Introduction

Conversion of time series data to images (ts2img) is a general problem that we have to deal with often for all kinds of datasets. Although most of the external datasets we use come in image format most of them are converted into a time series format for analysis. This is fairly straightforward for image stacks but gets more complicated for orbit data.

Orbit data mostly comes as several data values accompanied by latitude, longitude information and must often be resampled to fit on a grid that lends itself for time series analysis. A more or less general solution for this already exists in the img2ts module.

7.1.2 Possible steps involved in the conversion

The steps that a ts2img program might have to perform are:

1. Read time series in a geographic region - constrained by memory
2. Aggregate time series in time
    - methods for doing this aggregation can vary for each dataset so the method is best specified by the user
    - resolution in time has to be chosen and is probably also best specified by the user
    - after aggregation every point in time and in space must have a value which can of course be NaN
    - time series might have to be split into separate images during conversion, e.g. ASCAT time series are routinely split into images for ascending and descending satellite overpasses. **This means that we can not assume that the output dataset has the same number or names of variables as the input dataset.**
3. Put the now uniform time series of equal length into a 2D array per variable
4. A resampling step could be performed here but since only a part of the dataset is available edge cases would not be resolved correctly. A better solution would be to develop a good resampling tool which might already exist in pyresample and pytesmo functions that use it.

5. write this data into a file

    - this can be a netCDF file with dimensions of the grid into which the data is written
    - this could be any other file format, the interface to this format just has to make sure that in the end a consistent image dataset is built out of the parts that are written.

7.1.3 Solution

The chosen first solution uses netCDF as an output format. The output will be a stack of images in netCDF format. This format can easily be converted into substacks or single images if that is needed for a certain user or project.

The chosen solution will **not** do resampling since this is better and easier done using the whole converted dataset. This also means that if the input dataset is e.g. a dataset defined over land only then the resulting "image" will also not contain land points. I think it is best to let this be decided by the input dataset.

The output of the resulting netCDF can have one of two possible "shapes":

- 2D variables with time on one axis and gpi on the other. This is kind of how SWI time series are stored already.
- 3D variables with latitude, longitude and time as the three dimensions.

The decision of which it will be is dependent on the grid on which the input data is stored. If the grid has a 2D shape then the 3D solution will be chosen. If the input grid has only a 1D shape then only the 2D solution is possible.

Time Series aggregation

The chosen solution will use a custom function for each dataset to perform the aggregation if necessary. A simple example of a function that gets a time time series and aggregates it to a monthly time series could look like *agg_tsmonthly*

Simple example of a aggregation function

```python
def agg_tsmonthly(ts, **kwargs):
    """
    Parameters
    ----------
    ts : pandas.DataFrame
        time series of a point
    kwargs : dict
        any additional keyword arguments that are given to the ts2img object
        during initialization

    Returns
    -------
    ts_agg : pandas.DataFrame
        aggregated time series, they all must have the same length
        otherwise it can not work
        each column of this DataFrame will be a layer in the image
    """
    # very simple example
    # aggregate to monthly timestamp
    # should also make sure that the output has a certain length
    return ts.asfreq("M")
```

Time series iteration

The function `agg_tsmonthly` will be called for every time series in the input dataset. The input dataset must have a `iter_ts` iterator that iterates over the grid points in a sensible order.

Interface to the netCDF writer

The netCDF writer will be initialized outside the *ts2img* class with a filename and other attributes it needs. So the *ts2img* class only gets a writer object. This writer object already knows about the start and end date of the time series as well as the target grid and has initialized the correct dimensions in the netCDF file. This object must have a method `write_ts` which takes a array of gpi's and a 2D array containing the time series for these gpis. This should be enough to write the gpi's into the correct position of the netCDF file.

This approach should also work if another output format is supposed to be used.

Implementation of the main ts2img class

The ts2img class will automatically use a the function given in `agg_ts2img` if no custom `agg_ts2img` function is provided. If the tsreader implements a method called `agg_ts2img` this function will be used instead.

```
class Ts2Img(object):

    """
    Takes a time series dataset and converts it
    into an image dataset.
    A custom aggregate function should be given otherwise
    a daily mean will be used

    Parameters
    ----------
    tsreader: object
        object that implements a iter_ts method which iterates over
        pandas time series and has a grid attribute that is a pytesmo
        BasicGrid or CellGrid
    imgwriter: object
        writer object that implements a write_ts method that takes
        a list of grid point indices and a 2D array containing the time series data
    agg_func: function
        function that takes a pandas DataFrame and returns
        an aggregated pandas DataFrame
    ts_buffer: int
        how many time series to read before writing to disk,
        constrained by the working memory the process should use.

    """

    def __init__(self, tsreader, imgwriter,
                 agg_func=None,
                 ts_buffer=1000):

        self.agg_func = agg_func
        if self.agg_func is None:
            try:
                self.agg_func = tsreader.agg_ts2img
            except AttributeError:
                self.agg_func = agg_tsmonthly
        self.tsreader = tsreader
        self.imgwriter = imgwriter
        self.ts_buffer = ts_buffer

    def calc(self, **tsaggkw):
        """
        does the conversion from time series to images
        """
        for gpis, ts in self.tsbulk(**tsaggkw):
            self.imgwriter.write_ts(gpis, ts)

    def tsbulk(self, gpis=None, **tsaggkw):
        """
        iterator over gpi and time series arrays of size self.ts_buffer

        Parameters
        ----------
        gpis: iterable, optional
            if given these gpis will be used, can be practical
            if the gpis are managed by an external class e.g. for parallel
            processing
        tsaggkw: dict
            Keywords to give to the time series aggregation function


        Returns
        -------
        gpi_array: numpy.array
            numpy array of gpis in this batch
        ts_bulk: dict of numpy arrays
            for each variable one numpy array of shape
            (len(gpi_array), len(ts_aggregated))
        """
        # have to use the grid iteration as long as iter_ts only returns
        # data frame and no time series object including relevant metadata
        # of the time series
        i = 0
        gpi_bulk = []
        ts_bulk = {}
        ts_index = None
        if gpis is None:
            gpis, _, _, _ = self.tsreader.grid.grid_points()
        for gpi in gpis:
            gpi_bulk.append(gpi)
            ts = self.tsreader.read_ts(gpi)
            ts_agg = self.agg_func(ts, **tsaggkw)
            for column in ts_agg.columns:
                try:
                    ts_bulk[column].append(ts_agg[column].values)
                except KeyError:
                    ts_bulk[column] = []
                    ts_bulk[column].append(ts_agg[column].values)

            if ts_index is None:
                ts_index = ts_agg.index

            i += 1
            if i >= self.ts_buffer:
                for key in ts_bulk:
                    ts_bulk[key] = np.vstack(ts_bulk[key])
                gpi_array = np.hstack(gpi_bulk)
                yield gpi_array, ts_bulk
                ts_bulk = {}
                gpi_bulk = []
                i = 0
        if i > 0:
            for key in ts_bulk:
                ts_bulk[key] = np.vstack(ts_bulk[key])
            gpi_array = np.hstack(gpi_bulk)
            yield gpi_array, ts_bulk
```

7.2 License

7.3 Developers

- Christoph Paulik <christoph.paulik@geo.tuwien.ac.at>

7.4 repurpose

7.4.1 repurpose package

Submodules

repurpose.img2ts module

repurpose.ts2img module

module for conversion of time series data to image data Created on Mon Apr 20 11:08:58 2015

@author: christoph.paulik@geo.tuwien.ac.at

class `repurpose.ts2img.`**`Ts2Img`**(*tsreader*, *imgwriter*, *agg_func=None*, *ts_buffer=1000*)

Bases: `object`

Takes a time series dataset and converts it into an image dataset. A custom aggregate function should be given otherwise a daily mean will be used

Parameters

- **tsreader** (*object*) – object that implements a iter_ts method which iterates over pandas time series and has a grid attribute that is a pytesmo BasicGrid or CellGrid
- **imgwriter** (*object*) – writer object that implements a write_ts method that takes a list of grid point indices and a 2D array containing the time series data
- **agg_func** (*function*) – function that takes a pandas DataFrame and returns an aggregated pandas DataFrame
- **ts_buffer** (*int*) – how many time series to read before writing to disk, constrained by the working memory the process should use.

calc(***tsaggkw*)

does the conversion from time series to images

tsbulk(*gpis=None*, ***tsaggkw*)

iterator over gpi and time series arrays of size self.ts_buffer

Parameters

- **gpis** (*iterable, optional*) – if given these gpis will be used, can be practical if the gpis are managed by an external class e.g. for parallel processing
- **tsaggkw** (*dict*) – Keywords to give to the time series aggregation function

Returns

- **gpi_array** (*numpy.array*) – numpy array of gpis in this batch
- **ts_bulk** (*dict of numpy arrays*) – for each variable one numpy array of shape (len(gpi_array), len(ts_aggregated))

`repurpose.ts2img.`**`agg_tsmonthly`**(*ts*, ***kwargs*)

Parameters

- **ts** (*pandas.DataFrame*) – time series of a point
- **kwargs** (*dict*) – any additional keyword arguments that are given to the ts2img object during initialization

Returns **ts_agg** – aggregated time series, they all must have the same length otherwise it can not work each column of this DataFrame will be a layer in the image

Return type pandas.DataFrame

Module contents

CHAPTER 8

Indices and tables

- genindex
- modindex
- search

Python Module Index

Index