GUIDE

author: Petr Kolar   
email: kolda.efk@gmail.com

This is a brief guide to preparing the backward trajectory (BWT) data for use by the *openair* package in R, which offers a wide array of functions for BWT analysis. Essentially, *openair* functions expect a single dataframe containing all BWTs with properly named columns (see [*openair* documentation](https://davidcarslaw.com/files/openairmanual.pdf) for the details). Therefore, the preparation of BWT data for analysis in the *openair* requires two steps. First, the raw BWT data (obtained typically from the [*HYSPLIT Web*](https://www.ready.noaa.gov/HYSPLIT.php)) need to be combined into a single dataframe with the appropriate columns*.* Second, this BWT dataframe must be merged with the dataframe containing air pollution data if we wish to use functions for analysis of the origin of air pollution, such as potential source contribution function (PSCF) or concentration-weighted trajectories (CWT). This guide is best used together with the provided R script.

# Data description

The folder **BWT\_120h** contains daily BWTs for the months of March, April, and July of 2020, arriving at Chiang Mai in Northern Thailand. The BWTs arrive at 500 m above ground level, and their duration is 120 hours (5 days). The raw BWT data were obtained from HYSPLIT Web and are stored in individual text files (one file per trajectory) (Figure 1). Each file contains a header plus 121 rows of BWT data - one row per hour (1 starting hour + 120 hours back in time).

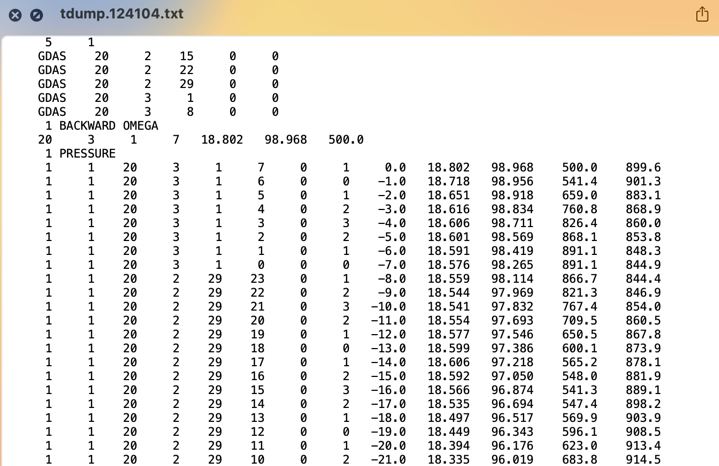


Figure 1. A part of the raw file for a single BWT as downloaded from HYSPLIT Web.

After the individual BWT files are processed within the first part of the **BWT\_in\_R.R** script, a single dataframe containing all BWTs is created – **BWT\_500m\_120h.csv**.

The file **PM\_daily.xlsx** contains daily PM2.5 data for the months of March, April, May, June, and July of 2020. The PM2.5 concentrations were obtained from a PCD Monitoring station 36T in Chiang Mai.

The data were provided by the Environmental Science Research Center at Chiang Mai University, Thailand.

# BWT\_in\_R.R

The first section of the **BWT\_in\_R***.* script (*BWT data preparation)* defines two functions. First, *combine\_files(folder\_path, traj\_length)* allows combining of all viable raw BWT files from the specified folder – *folder\_path*. The function first reads all the rows from the BWT text file (Figure 1) except the header. It uses the length of the BWTs – *traj\_length –* to achieve so. Specifically, it reads the last *traj\_length* + 1 rows. The file is then combined with the other BWT files; if there is an error (such as too few rows), the file is ignored, and a message is displayed on the console log. Finally, when all files in the folder have been processed, the function returns a single dataframe of all trajectories combined and a message showing the total number of BWTs.

The second function, *prep\_data(bwt\_file)* assigns column names to the data frame that are expected by the *openair* package. It drops the columns that are not needed and adds two columns: *date* and *date2*. Importantly, the argument of this function - *bwt\_file* - is expected to be a dataframe outputted by the *combine\_files* function. The function returns a dataframe that is ready to be used by *openair* and can be exported as a .csv or .xlsx file (**BWT\_500m\_120h.csv)** (Figure 2).

The columns of the data frame are:

* **traj** – the trajectory id number
* **year, month, day, hour** – time of the BWT point
* **hour.inc** – hours back in time of the BWT point
* **lat, lon** – coordinates of the BWT point
* **date** – the time of the arrival of the BWT at the receptor
* **date2** – the time of the BWT point (year, month, day, hour combined)
* A screenshot of a table

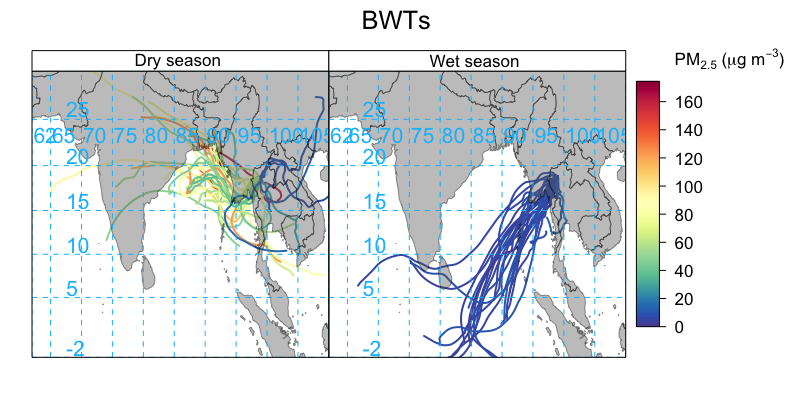
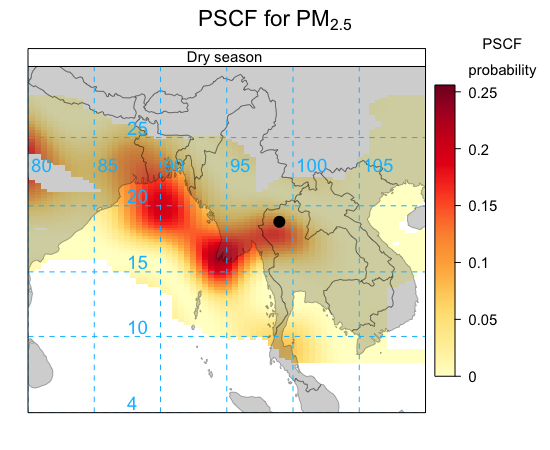
  Description automatically generated**height** – the height above ground level of the BWT point

Figure 2. The combined dataframe.

In the second section of the **BWT\_in\_R.R** script (*DATA IMPORT AND MERGE*) the previously created BWT dataframe **BWT\_500m\_120h.csv** is imported (note that if running the entire script, this dataframe will already be loaded in the working environment). Then the PM2.5 dataframe **PM\_daily.xlsx** is imported. Caution must be taken to have the same date format in both dataframes, including the time zone! The two dataframes are then merged using the date column (by default, as the column name is the same in both dataframes). Lastly, the column representing the respective season – seas - in Chiang Mai (dry/wet) is added.

The third section of **BWT\_in\_R.R** script demonstrates a few examples of how the pollution data and BWTs can be analysed together. Specifically, the functions *trajPlot, trajCluster* and *trajLevel* are used (Figure 3)*.* For the descriptions of these functions, refer to the *openair* documentation.

Figure 3. Examples of analysis methods – backward trajectory plot and potential source contribution function - produced by the openair package.



References:

Carslaw, D.C. (2019). The openair manual — open-source tools for analysing air pollution data. Manual for version 2.6-6, University of York (<https://davidcarslaw.com/files/openairmanual.pdf>).

Citation Draxler, R.R. and Rolph, G.D., 2010. HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) Model access via NOAA ARL READY Website ([<https://www.ready.noaa.gov/HYSPLIT.php>](https://www.ready.noaa.gov/HYSPLIT.php)). NOAA Air Resources Laboratory, College Park, MD.