README_targets

This is an R targets (https://github.com/ropensci/targets) pipeline for environmental health impact assessment using air pollution as a case study. It has been developed on R 4.1.2 “Bird Hippie” and RStudio 2021.09.2 “Ghost Orchid”. It requires R >= 4.0.0 and access to CARDAT’s Environment_General data storage folder on Cloudstor (https://cloudstor.aarnet.edu.au/).

The structure and syntax of an R targets pipeline may be unfamiliar to you depending on your level of coding experience. Depending on your intended usage, some or all of the following may guide your understanding of the workflow. Links to further useful examples and documentation are provided in the references.

Background

Health Impact Assessment (HIA) of ambient air pollution can quantify the impacts on human health using current and historical air pollution data and point directions for sustainable transitions that could promote policies, programmes, or projects to reduce air population. HIAs can be used to make recommendations for decision-makers and stakeholders, aiming to maximise a proposal’s positive health effects and minimise its negative effects. (1) A HIA also provides a way to engage with the public sphere by producing meaningful numbers to quantify the health effects of air pollution. The Scientific Workflow System (SWS) R targets workflow is a tool for quantifying the impact on health for given air pollution policy intervention scenarios, illustrated by a WHO guideline case study.

Further background information on the Epidemiological concepts underpinning HIAs and the methods utilised in this workflow can be found in Appendix B.

Initial setup and run-through:

To run the Air Health SWS for the first time:

1. Download and unzip or clone the air-health-sws-r-targets (https://github.com/cardat/air-health-sws-r-targets) repository from the Code dropdown button.
2. Load the R project. Open the `_targets.R` script.
    - `_targets.R` script is where you customise the workflow to suit your study needs.
    - Define global variables – line 21
        - Set your analysis year and states (location) to set the study coverage – note with the initial parameters in the pipeline, years are limited to 2010-2015.
        - Set your directory pathway to CARDAT data
            - Set the `download_data` boolean and `dir_cardat` to the correct path (parent directory of Environment_General).
    - Line 86 – data extraction and derivation
        - Provide counterfactual scenario(s) and calculate the delta
        - Counterfactual being the alternative air pollution exposure level
        - This can either be an absolute number (e.g. a guideline value such as the WHO guidelines)
        - Or a derived value – e.g. the minimum value from monitoring data.
    - Line 126 – Analysis

        - Input risk estimate:
            - Relative risk, including 95% confidence intervals, uses 1.062 (1.041, 1.084) is the default – as per WHO guidelines
3. Open the `main.R` script. (This is not integral to the targets pipeline but is a place to keep all the useful commands for visualising, running and exploring the pipeline outside of the pipeline itself.) Begin running the script line-by-line from the top.
    - `renv` should automatically install and activate. Install the packages using `renv::restore()` or try the alternative custom installation function `install_pkgs()` (installs the latest version of the library if it is not already available). This step may take some time.
    - If you have set `download_data <- FALSE` in `_targets.R`, uncomment and run the lines at the top of the *Run pipeline* section to authenticate your cloudstoR package's access to Cloudstor. You should not need to authenticate again unless your credentials have changed.
    - Continue to visualise and run the pipeline.
        - See appendix A for known error codes and troubleshooting.
4. See the results of the desired target with `tar_read(target_name)`.

WHO Use Case Example:

Estimate the mortality burden due to annual exposure to ambient fine particulate matter <2.5 µg (PM2.5) above the current (8 µg) WHO annual PM2.5 guidelines in Western Australia during 2013-2014.

1. Set Global variables – line 21
    - `years <- 2013:2014`
    - `states <- c("WA")`
2. Provide counterfactual scenario – line 104
    - In this use case, a scenario is an absolute number. However, it can be a derived value.
    - Provide counterfactual scenario(s) and calculate delta
        - abs = absolute value
        - min = derives from state monitoring data

```
tar_target(
  combined_exposures,
    do_env_counterfactual(data_env_exposure_pm25,
                          "abs",
                          8),
    pattern = map(data_env_exposure_pm25)
  )
```

3. Input relative risk estimate used for case scenario (default RR as defined by WHO guidelines) – line 128
    - RR input as c(RR, lower 95% CI bound, upper 95% CI bound)

```
analysis <- list(
# construct a function given relative risks and theoretical minimum risk
# the argument exposure_response_func takes a three element numeric vector, representing the
relative risk, lower confidence interval and upper confidence interval (in that order)

tar_target(health_impact_function,
             do_health_impact_function(
               case_definition = 'crd',
               exposure_response_func = c(1.062, 1.041, 1.084),
               theoretical_minimum_risk = 0
             )
  )
```

4. Run pipeline from `Main.R`
5. Create an HIA report by running `HIA 2013 2014 WA 8ug.Rmd`

Appendix A

Known errors

- Requires many packages to be installed. Depending on your existing setup, it could take up to an hour to install everything
- Packages `visNetwork` and `rgdal` may need to be installed manually for all code to work
- The `CloudstoR` package, used to download datasets from CloudStor, may fail during download for an unknown reason

Error:

```
Error installing package 'ps':
==============================
* installing to library 'C:/Users/djor8013/OneDrive - The University of Sydney (Staff)/En Hea
lth/air-health-sws-r-targets-main/renv/staging/1'
* installing *source* package 'ps' ...
** package 'ps' successfully unpacked and MD5 sums checked
** using staged installation
Warning in system("sh ./configure.win") : 'sh' not found
ERROR: configuration failed for package 'ps'
* removing 'C:/Users/djor8013/OneDrive - The University of Sydney (Staff)/En Health/air-healt
h-sws-r-targets-main/renv/staging/1/ps'
Error: install of package 'ps' failed [error code 1]
```

Solution: Uncomment out help download function:

```
source("R/func_helpers/helper_install_pkgs.R")
install_pkgs(repos = getOption("repos"))
```

Error: Unable to authenticate access to CloudStor.

Solution: Need to create app password in CloudStor to link data to R. View this support page for instructions on how to do this [here].(https://support.aarnet.edu.au/hc/en-us/articles/236034707-How-do-I-manage-change-my-passwords- (https://support.aarnet.edu.au/hc/en-us/articles/236034707-How-do-I-manage-change-my-passwords-))

Error:

```
! Error running targets::tar_make()
  Target errors: targets::tar_meta(fields = error, complete_only = TRUE)
  Tips: https://books.ropensci.org/targets/debugging.html
  Last error: The download destination specified is likely used by a sync client. Please choo
se another destination.
```

Solution:

- Create destination folder for CARDAT mirroring that is not a sync client (e.g. on local computer)
- Save `_targets.R` after doing this

Error:

```
! Error running targets::tar_make()
  Target errors: targets::tar_meta(fields = error, complete_only = TRUE)
  Tips: https://books.ropensci.org/targets/debugging.html
  Last error: Failed to open file C:/Users/djor8013/OneDrive - The University of Sydney (Staf
f)/Desktop/air-health-sws-r-targets-main/data_provided/Environment_General/Air_pollution_mode
l_
GlobalGWR_PM25/GlobalGWR_PM25_V4GL02/data_derived/
GlobalGWR_PM25_GL_201301_201312-RH35-NoNegs_AUS_20180618.tif.curltmp,
```

Solution:

- File path name too long
- Shorten `dir_cardat` pathname in `_targets.R`
    - For example: remove “OneDrive - The University of Sydney (Staff)”

Error:

```
! Error running targets::tar_make()
  Target errors: targets::tar_meta(fields = error, complete_only = TRUE)
  Tips: https://books.ropensci.org/targets/debugging.html
  Last error: error in evaluating the argument 'x' in selecting a method for function 'brick
': Cannot create RasterLayer object from this file; perhaps you need to install rgdal first
```

Solution: Install and load package `rgdal`

Error:

```
Error:
! Error running targets::tar_make()
  Target errors: targets::tar_meta(fields = error, complete_only = TRUE)
  Tips: https://books.ropensci.org/targets/debugging.html
  Last error: 'breaks' are not unique
```

Solution:
Attributable number for some scenarios can be very small – colourQuantile in r script viz_leaflet_an cannot handle these small numbers, will have quantile cutoffs with same break numbers.

Open `viz_leaflet_an.R`
Change:

```
pal1 <- colorQuantile(
    palette = "RdYlBu",
    domain = sf_an$attributable, n = 5, reverse = TRUE
  )
```

To:

```
pal1 <- colorNumeric(
    palette = "RdYlBu",
    domain = sf_an$attributable
  )
```

This creates a continuous pallet rather than a categorical palette.

Appendix B – Additional Epidemiological and Methodological Background information

We have presented additional information about various concepts underpinning epidemiological study design that you may find helpful in understanding health impact assessment methodology.

Epidemiological study designs

Source and sample populations

One of the fundamental concepts underpinning all epidemiological research is the requirement to clearly define the study base or the source population. (2) The source population is the entire group of individuals or objects with a common characteristic or condition that interests the study. This may include all individuals living in a particular geographical area, all individuals of a particular age group, or all individuals with a specific disease or health condition. (3) The source population is the starting point for identifying potential study participants or units.

On the other hand, a sample population is a subset of the source population selected for inclusion in the study. The sample population is usually selected through a sampling process that aims to ensure that the sample is representative of the source population in terms of the characteristics or conditions of interest. (3)

Case definition

In epidemiology, a case definition is a set of standard criteria used to identify whether an individual has a particular disease or health condition of interest. The case definition usually includes specific clinical,

laboratory, or other diagnostic criteria that are used to classify an individual as a case or non-case. (4)

A case definition aims to ensure that all cases are identified consistently and accurately across different settings and by other investigators. This is critical in epidemiological studies, as the case definition's accuracy can affect the study findings' validity and reliability. (4)

Mortality

Mortality is a special type of incidence in epidemiology because it represents the ultimate outcome of disease or health conditions. While incidence refers to the number of new cases of a disease or health condition that occur in a population over a specified period, mortality refers to the number of deaths that occur in a population over the same period.

In epidemiology, incidence and mortality are important measures of disease burden, but they provide different types of information. Incidence data provides information about the number of people newly diagnosed with a disease or health condition. In contrast, mortality data includes information about the number of people who die because of a disease or health condition.

Relative risk, odds ratio and hazard ratio

Relative Risk

Relative risk (RR) measures the strength of association between exposure to a risk factor and the occurrence of an outcome. It is calculated by dividing the incidence rate of the outcome in the exposed group by the incidence rate of the outcome in the unexposed group. An RR of 1 indicates that there is no association between exposure and outcome, an RR greater than 1 indicates a positive association (i.e., the exposed group has a higher risk of experiencing the outcome), and an RR less than 1 indicates a negative association (i.e., the exposed group has a lower risk of experiencing the outcome). (5)

In air pollution studies, RR is expressed as the ratio by which the risk of mortality increases per given increase in air pollution level. RR for a unit change in pollution level is represented by the coefficient β derived from empirical studies. For example, the WHO case study example uses a β coefficient from a pooled RR estimated from a meta-analysis of European and North American studies, as recommended by WHO. (1) That is a RR of 1.062 (95% CI 1.041, 1.084) per 10-g/m3 increment in annual average PM2.5 exposures of people aged ≥30 years. (1)

RR is a function of the difference in pollution levels $(x_1 - x_0)$. For any change in pollution level from $(x_1 - x_0)$, the relative risk is given by the formula:

$$RR(x_1 - x_0) = exp(\beta(x_1 - x_0)$$

The pollution level x_1 may be a target or cutoff level for which a policy or legislation aims, and it is likely to be lower than x_0.

Odds Ratio

The odds ratio expresses the measure of the association between exposure and outcome, often used in case-control studies. It is calculated by dividing the odds of exposure in cases by the odds of exposure in controls. An OR of 1 indicates no association, an OR greater than 1 indicates a positive association and an OR less than 1 indicates a negative association. (5)

Relative risk and odds ratio assume that the exposure precedes the outcome, but they differ in how they account for temporality. Relative risk is calculated using incidence rates, which require follow-up time, and therefore assumes a temporal relationship between exposure and outcome. On the other hand, the odds ratio does not require follow-up time and therefore does not directly account for temporality. (5) However, careful

study design and analysis can still establish the temporal relationship between exposure and outcome

Hazard Ratio

Hazard ratios (HR) measure the strength of association between an exposure and a time-to-event outcome, such as the onset of a disease or death. Hazard ratios are commonly used in epidemiology and survival analysis to compare the risk of an outcome between two or more groups while accounting for differences in follow-up time. This is important because the time at risk for an event may differ between the two groups due to differences in the onset of exposure, the time of diagnosis, or the study duration. By accounting for differences in follow-up time, hazard ratios can provide a more accurate estimate of the risk of the outcome associated with the exposure.

Within the context of air pollution epidemiological studies, a hazard ratio (HR) is the ratio of hazard rates corresponding to the conditions characterised by two distinct air pollution levels. The hazard rate (H) at pollution level x_1 is derived from those at level x_0 by:

$$H(x_1) = RR(x_1 - x_0) \times h(x_0)$$

The PAF (population attributable fraction)

Population attributable fraction (PAF) estimates the proportion of disease or adverse health outcomes in a population that can be attributed to a specific risk factor or exposure.

PAF is calculated by comparing the incidence of the disease or outcome in the total population to the incidence that would be expected if the population were not exposed to the risk factor or exposure of interest. The difference between these two incidences represents the proportion of cases attributable to the exposure. (6)

Mathematically, PAF can be expressed as:

$$PAF = (P_e \times (RR - 1))/(1 + (P_e \times (RR - 1)))$$

Where:

P_e = proportion of the population exposed to the risk factor or exposure
RR = relative risk (or hazard ratio) associated with the exposure

PAF can be interpreted as the proportion of cases that would be prevented if the exposure was eliminated from the population. A PAF of 0% indicates that the exposure is not associated with the disease or outcome, while a PAF of 100% indicates that all cases of the disease or outcome in the population can be attributed to the exposure.

TMREL – Theoretical minimum risk exposure level

The theoretical minimum risk exposure level (TMREL) is the level of exposure to a pollutant below which no adverse health effects are expected to occur. It is often used in air pollution epidemiology to inform regulatory decision-making and to set air quality standards.

The TMREL is based on a risk assessment of the available evidence on the health effects of exposure to the pollutant of interest. The TMREL is typically set at a level well below the lowest level of exposure associated with adverse health effects in the available studies.

Setting a TMREL involves balancing the need to protect public health and avoid unnecessary economic or social costs associated with reducing pollution levels. The TMREL can be influenced by various factors, including the nature and severity of the health effects associated with exposure, the size and characteristics of the exposed population, and the feasibility and costs of reducing exposure levels.

R targets Workflow - Methods

Population data

The R targets workflow uses age-specific population counts in 5-year age groups for each Statistical Area level 2 (SA2) geographical area (2016 ABS geographical boundaries), freely available from the Australian Bureau of Statistics dataset “Population by Age and Sex, Regions of Australia, Estimated Residential Population 2006–2016” from ABS-TableBuilder (cat. no. 3235.0).

As the PM2.5 and mortality risk function applied is for persons age 30+, the population used analyses is limited to 5-year age groups from 30 - 35 up to 100+.

Health data

Mortality data by 5-year age groups from 30 years up to and including 100 years + is used, freely accessible from the Australian Bureau of Statistics (Cat. No. 3302.0 – Deaths, Australia, available from the ABS.Stat website). Baseline age-specific annual mortality rates are calculated for each year by linking the mortality data with age-specific populations.

Exposure assessment

Accurate and reliable exposure assessment is critical to the validity and generalization of environmental epidemiological studies. Errors in exposure assessment can lead to biased or inaccurate estimates of the association between exposure and health outcomes, potentially leading to erroneous conclusions about the health risks of environmental agents. Exposure assessment of PM2.5 can be challenging, as the pollutant is widespread and varies in concentration over time and space.

The R targets workflow estimates population exposure to PM2.5 by obtaining annual average PM2.5 concentrations from a validated satellite-based land-use regression (LUR) model, as described by Knibbs et al. (7) This model incorporates observed PM2.5 measurements from air-monitoring stations with satellite data, chemical-transport model simulations and land-use data to predict concentrations across the study region by ABS mesh-block (MB) spatial unit. Data are available upon request from the Australian Centre for Air pollution, energy, and health Research (CAR) (https://cloudstor.aarnet.edu.au/plus/f/2454567279).

Annual average PM2.5 concentrations are calculated for the centroids of Australian Bureau of Statistics (ABS) MBs from the 2011 census geography. MBs are then assigned to SA2s from 2016 to derive population-weighted average exposures.

Attributable number

Australia has a limited number of epidemiological studies of long-term exposure to PM2.5 and mortality, so attributable mortality was calculated by applying a relative risk (RR) function estimated from a meta-analysis of European and North American studies, as recommended by WHO (8). A pooled RR of 1.062 (95% CI 1.041, 1.084) per 10-g/m3 increment in annual average PM2.5 exposures of people aged ≥30 years is recommended for health-impact assessments of PM2.5. (1) That is, for every 10μg/m3 increase in the PM2.5 annual average exposure, the risk of death increases by 6.2% (95% CI 4.1, 8.4%). (8)

This RR was used to calculate the attributable numbers (AN) of deaths associated with PM2.5 exposure in each SA2. AN was calculated based on estimates of baseline PM2.5 compared to the counterfactual and then aggregated to the state using the following equation:

$$AN = \sum(1 - e^{(1 - \beta\triangle_i_j}) \times Expected_i_j$$

Where $Expected_i_j$ is the death count estimated by applying the mortality rate in age-group i by age-specific

population counts within SA2 j, $\beta = log(RR)/10$ and $\triangle X_{ij}$ is the change in annual PM2.5 concentration from baseline concentrations to counterfactuals concentrations in SA2 j. Baseline concentrations were estimated as the population-weighted PM2.5 levels for each SA2 by year.

Counterfactual exposure concentrations

A counterfactual exposure concentration is a hypothetical exposure level that represents what would have happened if an individual or population had been exposed to a different level of an environmental agent than they actually were. It is a crucial concept in assessing the causal relationship between environmental exposures and health outcomes.

An example of a counterfactual exposure value is the WHO PM2.5 annual average guideline value of 5µg/m3. An alternative scenario is using the MB with the lowest annual average PM2.5 value for the study region.

References

1. WHO. Health risks of air pollution in europe—HRAPIE project: Recommendations for concentration-response functions for cost-benefit analysis of particulate matter, ozone and nitrogen dioxide. 2013.
2. Checkoway H, Pearce N, Kriebel D. Selecting appropriate study designs to address specific research questions in occupational epidemiology (https://doi.org/10.1136/oem.2006.029967). Occup Environ Med. 2007;64(9):633–8.
3. Banerjee A, Chaudhury S. Statistics without tears: Populations and samples (https://doi.org/10.4103/0972-6748.77642). Ind Psychiatry J. 2010;19(1):60–5.
4. Sharma SK. Importance of case definition in epidemiological studies (https://doi.org/10.1159/000332609). Neuroepidemiology. 2011;37(2):141–2.
5. Viera AJ. Odds ratios and risk ratios: What's the difference and why does it matter? (https://doi.org/10.1097/SMJ.0b013e31817a7ee4) South Med J. 2008;101(7):730–4.
6. Health AI of, Welfare. Australian burden of disease study 2015: Interactive data on risk factor burden [Internet]. AIHW; 2020. Available from: https://www.aihw.gov.au/reports/burden-of-disease/interactive-data-risk-factor-burden (https://www.aihw.gov.au/reports/burden-of-disease/interactive-data-risk-factor-burden)
7. Knibbs LD, Donkelaar A van, Martin RV, Bechle MJ, Brauer M, Cohen DD, et al. Satellite-based land-use regression for continental-scale long-term ambient PM(2.5) exposure assessment in australia (https://doi.org/10.1021/acs.est.8b02328). Environ Sci Technol. 2018;52(21):12445–55.
8. Hoek G, Krishnan RM, Beelen R, Peters A, Ostro B, Brunekreef B, et al. Long-term air pollution exposure and cardio- respiratory mortality: A review. Environmental Health [Internet]. 2013;12(1):43. Available from: https://doi.org/10.1186/1476-069X-12-43 (https://doi.org/10.1186/1476-069X-12-43)