CALL FOR CODE SPOT CHALLENGE FOR WILDFIRES

Abstract

This document introduces the 5 datasets that are part of the Call for Code Spot Challenge for Wildfires where wildfires will be predicted for Australia in February 2021.

The datasets provided are wildfires, historical weather, historical weather forecast, vegetation index, and land classes.

Challenge: http://ibm.biz/cfcsc-wildfires
Data: https://github.com/Call-for-Code/Spot-Challenge-Wildfires

Hendrik Hamann, hendrikh@us.ibm.com

Wildfire.csv

Description

This wildfire data contains data on fire activities in Australia starting from 2005. It is based on MCD14DL (shortname), which is the MODIS/Aqua+Terra Thermal Anomalies/Fire locations 1km FIRMS V006 NRT data product. Additional information can be found here:

- https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/c6-mcd14dl
- DOI:10.5067/FIRMS/MODIS/MCD14DL.NRT.006

MCD14DL is using swath satellite data (MOD14/MYD14). The thermal anomalies / active fire represents the center of an approximately 1km^2 pixel that is flagged by a Thermal Anomalies algorithm (Giglio 2003) as containing one or more fires within the pixel.

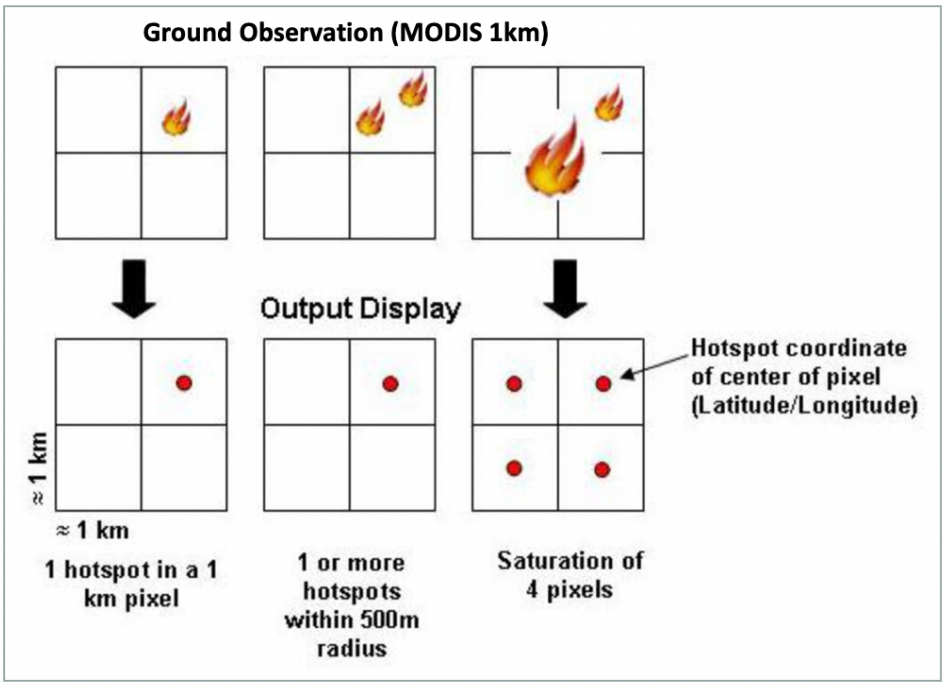

Image from: https://cdn.earthdata.nasa.gov/conduit/upload/12068/MODIS_fire_ground_observation.png

For this competition, all MCD14DL data was further processed by using IBM PAIRS Geoscope (https://ibmpairs.mybluemix.net/).

Processing

1. The data was spatially averaged to the following 7 regions/states in Australia
   - NSW=New South Wales
   - NT=Northern Territory
   - QL=Queensland
   - SA=Australia

- TA=Tasmania
- VI=Victoria
- WA=Western Australia

The corresponding polygons for each region can be queried/obtained from IBM PAIRS using the following polygon ids, respectively:

[200003,200004,200005,200006,200007,39027,200008]

2. In addition to spatial aggregation, all data was aggregated by day starting 1/1/2005. Multiple fires might have been observed in each region at different timestamps during a single day. The numbers of flagged pixels for each day are reported in the count column.
3. Furthermore, only fires are/were considered which were identified by the algorithms with high confidence (>75%).
4. In addition, only fires are/were considered which were identified as inferred hotspot type = 0 meaning a presumed vegetation fire. More than 98% of all detected fires are presumed vegetation fires.
5. The fire area is estimated by multiplying the along scan pixel size [in km] by the along track pixel size [in km].
6. The brightness is estimated by averaging both the brightness temperature 21 (obtained from channel 21/22) and brightness temperature 31 (obtained from channel 31).

Known issues

Please note that there is some data missing. 2007 has some small missing data from mid-August and data is missing for part of 21 April 2009 and missing for 22 April 2009. There might also be some erroneous data present in the data set.

Columns

1. **Region**: The respective regions as outlined above for which the data was/is aggregated.
2. **Date**: Day of acquisition of the data. All dates are in UTC and provide the data for 24 hours ahead.
3. **Estimated_fire_area**: Daily sum of estimated fire area for presumed vegetation fires with a confidence level of larger than 75% for a given region. To obtain this estimated area the along scan pixel size was multiplied by the along track pixel size. The nominal unit for the area is in km^2
4. **Mean_estimated_fire_brightness**: Daily mean (by flagged fire pixels(=count)) of estimated fire brightness for presumed vegetation fires with a confidence level of larger than 75% for a given region. The data was obtained by averaging the means of both the brightness temperature 21 (obtained from channel 21/22) and brightness temperature 31 (obtained from channel 31). The units are in Kelvin.
5. **Mean_estimated_fire_radiative_power**: Daily mean (by flagged fire pixels(=count)) of estimated radiative power for presumed vegetation fires with a confidence level of larger than 75% for a given region. The units are in megawatts.
6. **Mean_confidence**: Daily mean of confidence for presumed vegetation fires with a confidence level of larger than 75% for a given region. This value is based on a collection of intermediate algorithm quantities used in the detection process. It is intended to help users gauge the quality of individual hotspot/fire pixels. Confidence estimates range between 0 and 100%

7. **Std_confidence**: Standard deviation of estimated fire radiative power if available. The units are in megawatts.
8. **Var_confidence**: Variance of estimated fire radiative power if available. The units are in megawatts.
9. **Count**: Daily numbers of pixels for presumed vegetation fires with a confidence level of larger than 75% for a given region.
10. **Replaced**: Indicates with an "R" whether the data has been replaced with higher quality data when available (usually with a 2-3 month lag). Replaced data has slightly higher quality but it is expected to be of very minor impact in this contest. Please note that most corrections in the replaced data are associated with the precise geolocation and the fire classification. Most notably there is no classification for the type of fire, which means that all fires are assumed to be vegetation fires. Comparing the data from 2005 to 2020 more than 98% of the fires were eventually classified to be presumed vegetation fires.

Attribution

Please refer to the disclaimer and citation requirements below.

CITATION

We acknowledge the use of data and imagery from LANCE FIRMS operated by NASA's Earth Science Data and Information System (ESDIS) with funding provided by NASA Headquarters.

https://earthdata.nasa.gov/earth-observation-data/near-real-time/citation#ed-firms- citation

DISCLAIMER

The LANCE system is operated by the NASA/GSFC Earth Science Data and Information System (ESDIS). The information presented through LANCE, Rapid Response, GIBS, Worldview, and FIRMS are provided "as is" and users bear all responsibility and liability for their use of data, and for any loss of business or profits, or for any indirect, incidental or consequential damages arising out of any use of, or inability to use, the data, even if NASA or ESDIS were previously advised of the possibility of such damages, or for any other claim by you or any other person. ESDIS makes no representations or warranties of any kind, express or implied, including implied warranties of fitness for a particular purpose or merchantability, or with respect to the accuracy of or the absence or the presence or defects or errors in data, databases of other information. The designations employed in the data do not imply the expression of any opinion whatsoever on the part of ESDIS concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

HistoricalWeather.csv

Description

The file HistoricalWeather.csv contains daily aggregates computed from the hourly ERA5 climate reanalysis. Please refer to the links below for further information

- https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview
- https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

Processing

Raw ERA5 data was processed using IBM PAIRS Geoscope. For each parameter, temporal aggregation, unit conversion and other transformations (such as calculation of relative humidity) is applied first. Spatial aggregation is applied to the output of that first step. Although the raw data uses pixels in a latitude/longitude grid, the weights used in the spatial aggregation correspond to the physical area (in km^2) of each pixel. Example: Given a row with Date 2020-01-01 and Parameter Temperature, the value under max() is the maximum daily average temperature across the region. Similarly, the variance is the second moment of the daily average temperatures across the region. For a row with date YYYY-mm-dd, each parameter is processed as follows:

1. **Precipitation**: Derived from total precipitation (tp). Hourly, raw data is converted from m/hour to mm/hour. Then, values from YYYY-mm-ddT01:00:00Z to YYYY-mm-(dd+1)T00:00:00Z are added up. (Precipitation in ERA5 denotes accumulated precipitation over the past hour.)
2. **Relative humidity**: Derived from temperature and dewpoint (t2, d2). For each hourly timestamp, relative humidity is computed using the equation:

   100*{EXP[(17.625*TD)/(243.04+TD)]/EXP[(17.625*T)/(243.04+T)]}

   Here, TD is the dewpoint in degrees Celsius and T the temperature in the same unit. Both values are valid at 2 meters above ground. Subsequently, the same temporal aggregation as for temperature is used. (All 25 hourly timestamps from YYYY-mm-ddT00:00:00Z to YYYY-mm-(dd+1)T00:00:00Z. 0:00 UTC values are given half weight, all other values given full weight.)

3. **Soil water content**: Soil water content 0 - 7 cm below the surface (swvl1). Same temporal aggregation as for temperature. (All 25 hourly timestamps from YYYY-mm-ddT00:00:00Z to YYYY-mm-(dd+1)T00:00:00Z. 0:00 UTC values are given half weight, all other values given full weight.)
4. **Solar radiation**: The raw data is also known as Surface Solar Radiation Downwards (ssrd). Unit conversion from J/h to MJ/h. Same temporal aggregation as for precipitation. (I.e. values from YYYY-mm-ddT01:00:00Z to YYYY-mm-(dd+1)T00:00:00Z are added up.)
5. **Temperature**: Uses all 25 hourly timestamps from YYYY-mm-ddT00:00:00Z to YYYY-mm-(dd+1)T00:00:00Z. 0:00 UTC values are given half weight, all other values given full weight.
6. **Wind speed**: Derived form Easterly and Northerly wind speeds (u10 and v10). For each hourly timestamp, the wind speed is calculated from the Easterly and Northerly 10 meter wind components commonly known as u10 and v10 using (u10**2 + v10^2)^1/2. Subsequently, the same temporal aggregation as for temperature is used. (All 25 hourly timestamps from YYYY-mm-

ddT00:00:00Z to YYYY-mm-(dd+1)T00:00:00Z. 0:00 UTC values are given half weight, all other values given full weight.)

Columns

1. **Date**: Timestamp of the data. If the data is YYYY-mm-dd, the data in the row was created by aggregating the hourly ERA5 data from YYYY-mm-ddT00:00:00Z to YYYY-mm-(dd+1)T00:00:00Z.
2. **Region**: Denotes a region in Australia. Raw ERA5 data comes in raster form on a grid of 0.25 x 0.25 degrees resolution. Following the temporal aggregation, the data was spatially aggregated.
3. **Parameter**: Precipitation [mm/day], Relative humidity [%], Soil water content [m**3 m**-3], Solar radiation [MJ/day], Temperature [C], Wind speed [m/s]. Note that raw ERA5 data does not contain relative humidity. Instead, relative humidity data was computed from the temperature and dewpoint values provided by ERA5.
4. **count()[unit: km^2]**: Area of the Region. In km**2.
5. **min()**: Minimum value of the spatial aggregation.
6. **max()**: Maximum value of the spatial aggregation.
7. **mean()**: Average of the spatial aggregation.
8. **variance**: variance or 2nd moment of the spatial aggregation.

Known issues

There are a very few missing data points. These become apparent when pivoting the data with index Date and columns (Parameter, Region). For all regions except New South Wales, the min() column for soil water content shows values of order 10^{-6} m^3 m^{-3}. The reason for this is that the underlying model defines ocean pixels as having 0 soil water content. Coastal pixels may then have very negligible values due to interpolation. It is enough for one of these to enter the area covered by the spatial aggregate to bring the min() value down to this scale.

Attribution

CITATION

Generated using Copernicus Climate Change Service information.

HistoricalWeatherForecast.csv

Description

This file contains daily aggregates computed from the output of the Global Forecast System. For further information please refer to:

- https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs

Processing

Please refer to the processing description for the HistoricalWeather.csv and here we only refer to date and lead time.

In opposite to observations of the weather, a weather forecast generally comes with two timestamps: The time the forecast was made and the time it is for. The table reflects this in the columns "Date" and "Lead time". For concreteness, let us consider a row with Date YYYY-mm-dd and Lead time s. Then, the forecast in question was released on YYYY-mm-(dd-s) and predicts the weather between YYYY-mm-ddT00:00:00Z and YYYY-mm-(dd+1)T00:00:00Z. In other words, the Date is the valid date of the forecast, the lead time is measured in days, the issue time can be obtained from the formula issue time = valid time - lead time and given a date YYYY-mm-dd, the values in the table correspond to the 24 hours following YYYY-mm-ddT00:00:00Z.

Columns

1. **Date**: Timestamp of the data. See Comments on Date and Lead time.
2. **Region**: Denotes a region in Australia. Raw ERA5 data comes in raster form on a grid of 0.5 x 0.5 degrees resolution. Following the temporal aggregation, the data was spatially aggregated.
3. **Parameter**: Precipitation [mm/day], Relative humidity [%], Solar radiation [MJ/day], Temperature [C], Wind speed [m/s].
4. **Lead time [days]**: Difference between the time the forecast is for ("valid time") and the time the forecast was made ("issue time" or "data time"). See "Comments on Date and Lead time".
5. **count()[unit: km^2]**: Area of the Region. In km**2.
6. **min():** Minimum value of the spatial aggregation.
7. **max()**: Maximum value of the spatial aggregation.
8. **mean()**: Average of the spatial aggregation.
9. **variance()**: 2nd moment of the spatial aggregation.

Known issues

1. The Lead time=15 forecast for Precipitation for 2017-10-06 is clearly an outlier. We have left it in the data, yet most users will want to remove it.
2. Different combinations of Parameter and Lead time are available for different time ranges. Precipitation is generally available from July 2015 onwards. With small differences in the exact start date depending on Lead time. For the other paramters, the Lead time=5 forecasts are available form January 2014 with the remainder being available form July 2015 onwards. Again, there are small differences depending on the Lead time.

Attribution

CITATION

Derived from NOAA datasets

VegetationIndex.csv

Description

The vegetation index data set reports the normalized differential vegetation index (NDVI) starting in 2005 for Australia monthly. The data set is based on the observations of the MODIS, Terra 13 satellite at 250 m resolution at 16 days intervals. Generally, the data can be assumed to be cloud free. Additional information can be found:

- https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod13q1_v006
- https://lpdaac.usgs.gov/sites/default/files/public/product_documentation/mod13_user_guide.pdf
- https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index
- https://modis.gsfc.nasa.gov/about/specifications.php

The normalized vegetation index is a well-established concept to assess the “greenness” of vegetation,

- https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index

In general, a lower vegetation index indicates a “drier” state. The vegetation index can be between -1 and 1. -1 would generally correspond to water bodies but note that this data set is only for land. -0.2 is something like bare rock.

For this competition, all MODIS Terra 13 satellite data was further processed by IBM PAIRS Geoscope (https://ibmpairs.mybluemix.net/).

Processing

1. The data was spatially averaged to the following 7 regions/states in Australia
    - NSW=New South Wales
    - NT=Northern Territory
    - QL=Queensland
    - SA=Australia
    - TA=Tasmania
    - VI=Victoria
    - WA=Western Australia

The corresponding polygons for each region can be queried/obtained from IBM PAIRS using the following polygon ids, respectively but note that MODIS, Terra 11 only reports land based observations

[200003, 200004, 200005, 200006, 200007,39027,200008]

2. In addition to spatial aggregation, all data was aggregated by month starting 1/1/2005. Multiple observations might exist in each region at different timestamps during a single month.

Known issues

There are no known issues.

Columns

1. **Region**: The respective regions as outlined above for which the data was/is aggregated.
2. Date: Month of acquisition of the data. All dates are in UTC and provide the data for the same months. If multiple timestamps are available during the month the mean of the observation is computed first.
3. **Vegetation_index_mean**: The spatial mean of the vegetation index for the given region and month.
4. **Vegetation_index_max**: The spatial maximum value of the vegetation index for the given region and months.
5. **Vegetation_index_min**: The spatial minimum value of the vegetation index for the given region and months.
6. **Vegetation_index_variance**: The spatial variance of the vegetation index for the given region and months.

Attribution

CITATION

Didan, K. (2015). MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA EOSDIS LP DAAC. doi: 10.5067/MODIS/MOD13Q1.006

LandClass.csv

Description

The land class data set was derived from the CGLS land cover data product, which is based on PROBA-V satellite measurements. More specifically, the dataset contains the version 2.0 data that is available for 2015 alone

You can find additional information at:

- https://land.copernicus.eu/global/products/lc
- https://doi.org/10.3390/rs12061044
- https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_PUM_LC100m-V2.0_I2.20.pdf

Processing

1. The data was spatially averaged to the following 7 regions/states in Australia
    - NSW=New South Wales
    - NT=Northern Territory
    - QL=Queensland
    - SA=Australia
    - TA=Tasmania
    - VI=Victoria
    - WA=Western Australia

The corresponding polygons for each region can be queried/obtained from IBM PAIRS using the following polygon ids:

[200003, 200004, 200005, 200006, 200007,39027,200008]

2. In addition to spatial aggregation, all data was normalized and so that the land classes are in % of the entire area. A single land pixel can only be of one class. The respective areas for the polygons can be obtained from the weather data sets.

Known issues

There are no known issues.

Columns

1. **Region**: The respective regions as outlined above for which the data was/is aggregated.
2. **Shrubs**: Percentage [%] of shrubs in the respective region.
3. **Herbaceous vegetation**: Percentage [%] of herbaceous vegetation in the respective region.

4. **Cultivated and managed vegetation/agriculture (cropland)**: Percentage [%] of cultivated and managed vegetation/agriculture (cropland) in the respective region.
5. **Urban / built up**: Percentage [%] of urban / built up in the respective region.
6. **Bare / sparse vegetation**: Percentage [%] of bare / sparse vegetation in the respective region.
7. **Permanent water bodies**: Percentage [%] of permanent water bodies in the respective region.
8. **Herbaceous wetland**: Percentage [%] of herbaceous wetland in the respective region.
9. **Closed forest, evergreen, broad leaf**: Percentage [%] of closed forest, evergreen, broad leaf in the respective region.
10. **Closed forest, deciduous broad leaf**: Percentage [%] of closed forest, deciduous broad leaf in the respective region.
11. **Closed forest, unknown**: Percentage [%] of closed forest, unknown in the respective region.
12. **Open forest, evergreen broad leaf**: Percentage [%] of open forest, evergreen broad leaf in the respective region.
13. **Open forest, deciduous broad leaf**: Percentage [%] of open forest, deciduous broad leaf in the respective region.
14. **Open forest, unknown definitions**: Percentage [%] of open forest, unknown definitions in the respective region.
15. **Open sea**: Percentage [%] of open sea in the respective region.

Attribution