USA Wildfires Project

I have chosen this data set as it contains both geospatial and time series data. With over 2 million records, I am interested to see what patters and trends could be uncovered. It is also something which directly impacts communities, wildlife and the environment.

The data set contains information on wildfires which occurred in the USA over 27 years (1992 – 2018). It was the fourth update of a publications with the original purpose of supporting the US Fire Program Analysis (FPA) system. Although the program has since been retired, the purpose was to support the process for strategic fire management planning and budgeting.

The database includes over 2 million records representing a total of 165 million acres burned during the period.

Data Source

The wildfire records were collated from federal, state and local fire organisations making it a more reliable source. Basic error-checking was performed and redundant records were removed, making the data set easier to work with.

It is an open-source data set without additional permissions or fees:
https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0009.5

Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.5

Data Collection

In some states, reporting wildfires is voluntary. It has been estimated that only two-thirds of fires of any type, are reported to the system of record for U.S. fire departments, NFIRS. Subsequently, this dataset is likely to be an incomplete picture of wildfires in the US.

Data Contents

The database contains both geospatial and time series data as well as key data points on individual wildfires including the size.

Data Relevance

Hypotheses

The frequency and geographical range of wildfires have increased over time.
Due to higher temperature climates, southern states have a higher risk of wildfires.

Historical wildfire data can be used to determine trends over time and also help identify high risk states.

Data Profile

The data set contains 2,166,753 rows and 37 columns (listed below). The 14 columns kept for data analysis are highlighted.

FOD_ID = Unique numeric record identifier.

FPA_ID = Unique identifier that contains information necessary to track back to the original record in the source dataset.

SOURCE_SYSTEM_TYPE = Type of source database or system that the record was drawn from (federal, nonfederal, or interagency).

SOURCE_SYSTEM = Name of or other identifier for source database or system that the record was drawn from. See Table 1 in Short (2014), or \Supplements\FPA_FOD_source_list.pdf, for a list of sources and their identifier.

NWCG_REPORTING_AGENCY = Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report (BIA = Bureau of Indian Affairs, BLM = Bureau of Land Management, BOR = Bureau of Reclamation, DOD = Department of Defense, DOE = Department of Energy, FS = Forest Service, FWS = Fish and Wildlife Service, IA = Interagency Organization, NPS = National Park Service, ST/C&L = State, County, or Local Organization, and TRIBE = Tribal Organization).

NWCG_REPORTING_UNIT_ID = Active NWCG Unit Identifier for the unit preparing the fire report.

NWCG_REPORTING_UNIT_NAME = Active NWCG Unit Name for the unit preparing the fire report.

SOURCE_REPORTING_UNIT = Code for the agency unit preparing the fire report, based on code/name in the source dataset.

SOURCE_REPORTING_UNIT_NAME = Name of reporting agency unit preparing the fire report, based on code/name in the source dataset.

LOCAL_FIRE_REPORT_ID = Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year.

LOCAL_INCIDENT_ID = Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year.

FIRE_CODE = Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (https://www.firecode.gov/).

FIRE_NAME = Name of the incident, from the fire report (primary) or ICS-209 report (secondary).

ICS_209_PLUS_INCIDENT_JOIN_ID = Primary identifier needed to join into operational situation reporting data for the incident in the ICS-209-PLUS dataset.

ICS_209_PLUS_COMPLEX_JOIN_ID = If part of a complex, secondary identifier potentially needed to join to operational situation reporting data for the incident in the ICS-209-PLUS dataset (2014 and later only).

MTBS_ID = Incident identifier, from the MTBS perimeter dataset.

MTBS_FIRE_NAME = Name of the incident, from the MTBS perimeter dataset.

COMPLEX_NAME = Name of the complex under which the fire was ultimately managed, when discernible.

FIRE_YEAR = Calendar year in which the fire was discovered or confirmed to exist.

DISCOVERY_DATE = Date on which the fire was discovered or confirmed to exist.

DISCOVERY_DOY = Day of year on which the fire was discovered or confirmed to exist.

DISCOVERY_TIME = Time of day that the fire was discovered or confirmed to exist.

NWCG_CAUSE_CLASSIFICATION = Broad classification of the reason the fire occurred (Human, Natural, Missing data/not specified/undetermined).

NWCG_GENERAL_CAUSE = Event or circumstance that started a fire or set the stage for its occurrence (Arson/incendiarism, Debris and open burning, Equipment and vehicle use, Firearms and explosives use, Fireworks, Misuse of fire by a minor, Natural, Power generation/transmission/distribution, Railroad operations and maintenance, Recreation and ceremony, Smoking, Other causes, Missing data/not specified/undetermined).

NWCG_CAUSE_AGE_CATEGORY = If cause attributed to children (ages 0-12) or adolescents (13-17), the value for this data element is set to Minor; otherwise null.

CONT_DATE = Date on which the fire was declared contained or otherwise controlled (mm/dd/yyyy where mm=month, dd=day, and yyyy=year).

CONT_DOY = Day of year on which the fire was declared contained or otherwise controlled.

CONT_TIME = Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, mm=minutes).

FIRE_SIZE = The estimate of acres within the final perimeter of the fire.

FIRE_SIZE_CLASS = Code for fire size based on the number of acres within the final fire perimeter (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).

LATITUDE = Latitude (NAD83) for point location of the fire (decimal degrees).

LONGITUDE = Longitude (NAD83) for point location of the fire (decimal degrees).

OWNER_DESCR = Name of primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.

STATE = Two-letter alphabetic code for the state in which the fire burned (or originated), based on the nominal designation in the fire report.

COUNTY = County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.

FIPS_CODE = Five-digit code from the Federal Information Process Standards (FIPS) publication 6-4 for representation of counties and equivalent entities, based on the nominal designation in the fire report.

FIPS_NAME = County name from the FIPS publication 6-4 for representation of counties and equivalent entities, based on the nominal designation in the fire report.

Data Wrangling

The data set was downloaded as an SQLite database and uploaded in Python as a data frame. Out of the 27 columns in the table, only 14 were imported into Python (as highlighted above).

After checking the data frame column data types, the "DISCOVERY_DATE" column was changed from an "object" to "datetime64[ns]" type.

Consistency Checks

Upon investigation, the 'DISCOVERY_TIME' column was found to have a statistically significant number of missing values and was dropped.

The "COUNTY" code and "FIPS_NAME" county columns also had missing values but they were not dropped as it could be worth exploring to replace the missing values later on using the "LATITUDE" and "LONGITUDE" coordinates.

Although on the face of it, the "OWNER_DESCR" did not have any NaN values, there are ov er one million rows with "MISSING/NOT SPECIFIED" owner description. Although statistically significant, I am not dropping the column at this stage without further analysis as it could still be useful.

The were no columns with mixed data types and there were no duplicate records found.

Data Profile

Variable	Time-variant / invariant	Structured / Unstructured	Quantitative / Qualitative	Nominal / Ordinal	Discrete / Continuous
FOD_ID	Invariant	Structured	Qualitative	Ordinal	
FIRE_YEAR					
DISCOVERY_DATE					
DISCOVERY_DOY					
DISCOVERY_TIME					
NWCG_GENERAL_CAUSE				Nominal	
FIRE_SIZE			Quantitative	Ordinal	Continous
FIRE_SIZE_CLASS			Qualitative	Nominal	
LATITUDE					
LONGITUDE					
OWNER_DESCR					
STATE					
COUNTY		Unstructured			

Basic Statistical Descriptive Analysis

	FOD_ID	FIRE_YEAR	DISCOVERY_DOY	FIRE_SIZE	LATITUDE	LONGITUDE	DISC_MONTH
count	2166753.00	2166753.00	2166753.00	2166753.00	2166753.00	2166753.00	2166753.00
mean	100699748.95	2005.32	164.99	75.99	36.89	-96.19	5.94
std	150380118.92	7.54	89.99	2536.04	6.02	16.65	2.95
min	1.00	1992.00	1.00	0.00	17.94	-178.80	1.00
25%	582842.00	1999.00	89.00	0.10	32.96	-110.85	3.00
50%	1320811.00	2006.00	165.00	0.97	35.64	-93.11	6.00
75%	201662150.00	2011.00	230.00	3.00	40.81	-82.46	8.00
max	400482086.00	2018.00	366.00	662700.00	70.33	-65.26	12.00

Limitations and Ethics

As mentioned above, although the data set may be the most complete and official of wildfires in the US, due to the fact that in some states reporting wildfires is voluntary, this is likely not the complete picture. The implication is that incorrect fire management planning and budgeting decisions could be made if used for that purpose.

There is no personally identifiable information (PII) in the data set, meaning there are no ethical concerns from that perspective.

Questions to Explore

- What are the main causes of wildfires?
- What percentage of wildfires are caused by humans?
- When do wildfires happen? i.e., in which month and day of the week
- Have incidents been increasing or decreasing over time?
- Where do fires occur? i.e., in which states and counties?
- On which type of land do wildfires occur e.g., government owned etc.