ALTERNATIVE DATASETS FOR IDENTIFICATION OF EARTH SCIENCE EVENTS AND DATA

Kaylin Bugbee[1], Robert Griffin[1], Brian Freitag[1], Jeffrey Miller[1], Rahul Ramachandran[2], Jia Zhang[3]

[1]University of Alabama in Huntsville
[2]NASA Marshall Space Flight Center
[3]Carnegie Mellon University

ABSTRACT

Alternative, or non-traditional, data sources can be used to generate datasets which can in turn be analyzed for temporal, spatial and climatological patterns. Events and case studies inferred from the analysis of these patterns can be used by the remote sensing community to more effectively search for Earth observation data. In this paper, we present a new alternative Earth science dataset created from the National Weather Service's Area Forecast Discussion (AFD) documents. We then present an exploratory methodology for identifying interesting climatological patterns within the AFD data and a corresponding motivating example as to how these data and patterns can be used to search for relevant events or case studies.

Index Terms— Alternative data, big data, event identification, geospatial analysis, weather forecasting

1. INTRODUCTION

The volume of Earth observation data has grown significantly since the first satellites were launched in the 1960s and is expected to grow exponentially in the future with the launch of more powerful and accurate sensors [1]. For example, the upcoming NASA/Indian Space Research Organization (ISRO) Synthetic Aperture Radar (NISAR) mission is expected to generate as much as 85 TB of data per day or 140 PB of data over a three-year period [2]. In light of this data deluge, new scientific and technological approaches will be needed in order to better manage, understand and analyze Earth observation data.

Alternative data sources offer a solution to more accurately pinpointing interesting scientific events within these large Earth observation datasets. Alternative data sources are broadly defined as data which are extracted or generated from non-traditional sources and can include social media data, point of sale transactions and product reviews. While the idea of alternative data originates in the investment world, there are other non-traditional, domain specific data sources that can be similarly leveraged within the Earth sciences. These data sources include, but are not limited to, the information found in numerous unstructured text documents such as flight reports for airborne field campaigns, agricultural reports and weather forecast discussions. Information extracted from these documents can be used to generate datasets that can be analyzed for spatial, temporal and climatological patterns in order to more effectively identify interesting events or trends. Events and trends extracted from these alternative data sources assist the remote sensing community in more efficiently identifying interesting events or use cases and can also help decision makers better understand reporting of anticipated hazards and disasters. These data can in turn be leveraged to build an event database that will help the remote sensing community more effectively discover and use Earth observation data for research and for labelling data for deep learning applications.

In this paper, we describe the creation of an alternative Earth science dataset from the extraction of key vocabulary terms in the National Weather Service's Area Forecast Discussion (AFD) documents. A brief description of the original unstructured data and the methodology used to extract the terminology will be included. In section three, we describe the methods and results of an exploratory use case in which we test the viability of our newly created dataset as an alternative data source for identifying events. Finally, we conclude the paper by describing lessons learned from this project and future directions of this work.

2. ALTERNATIVE DATASET CREATION

The National Weather Service's Area Forecast Discussion (AFD) documents were identified as a viable candidate for an alternative Earth science data source. The National Weather Service (NWS) operates 122 weather forecast offices (WFO) across the United States. Each office has a geographic area of responsibility for which they issue forecasts and severe weather warnings. The AFD is one forecast document that the WFOs are required to write. The AFD is "intended to provide our customers and partners with the scientific reasoning forecasters used to develop forecasts and warnings, an understanding of the forecaster's

confidence in the forecast, and a summary of watches, warnings and/or advisories in effect" [3]. Each forecast focuses on the most significant weather issues facing the WFO's coverage area and includes a description of the forecast and a summary of all outlooks, watches, warnings or advisories issued. On average, an AFD is written every six hours but this frequency can vary depending on the weather patterns occurring in the WFO's coverage area. The NWS makes the last 50 versions of each office's AFD documents available at any given time but the NWS does not have a publicly accessible database of all previous AFDs. Fortunately, the Iowa State University, Iowa Environmental Mesonet website [4] archives many NWS text documents including the AFDs back to the year 2001. The Iowa Environmental Mesonet website was therefore used to obtain the historical AFD documents. Since the Iowa Environmental Mesonet website provides the AFDs as web pages and not text documents, it was necessary to scrape each page for the text needed.

The American Meteorological Society (AMS) Glossary of Meteorology terms were extracted from all available AFDs for the years 2007 to 2017. The AMS maintains the Glossary of Meteorology, a living document that curates over 12,000 important meteorological terms. The Glossary of Meteorology defines "every important meteorological term likely to be found in the literature today" and to provide "definitions that are understandable to the generalist and yet palatable to the specialist" [5]. The highly curated and domain specific nature of the AMS Glossary of Meteorology makes it a reliable source to leverage for relevant term extractions. Therefore, the Glossary of Meteorology terms were extracted from the AFDs using heuristic, rule-based extraction techniques [6]. The rule-based extraction technique is similar to the one described in [6]. Other data is also captured with each term extracted from the AFDs including the WFO name, the date and time of the AFD and the number of times each term is used within the AFD.

3. EXPLORATORY USE CASE

A use case was identified from over ten relevant Glossary of Meteorology terms in order to test the viability of the newly created AFD extraction dataset as a possible alternative data source for identifying events. The term 'hard freeze' was selected for the first exploratory use case. The Glossary of Meteorology defines the term 'hard freeze' as "a freeze in which seasonal vegetation is destroyed, the ground surface is frozen solid underfoot, and heavy ice is formed on small water surfaces such as puddles and water containers" [7]. The term 'hard freeze' is often used to alert a WFO's customers that subfreezing temperature conditions may occur which could in turn cause crop or plant damage. Understanding where, when and how often hard freezes occur is important since freezing temperatures in combination with plant growth timing determine frost damage. Understanding hard freeze events in the past and quickly responding to ongoing hard freeze events is especially important to agricultural mitigation and adaptation planning. The AFD extraction data aids in identifying these hard freeze events and in efficiently finding relevant Earth observation data.

3.1. Geospatial Hot Spot Analysis

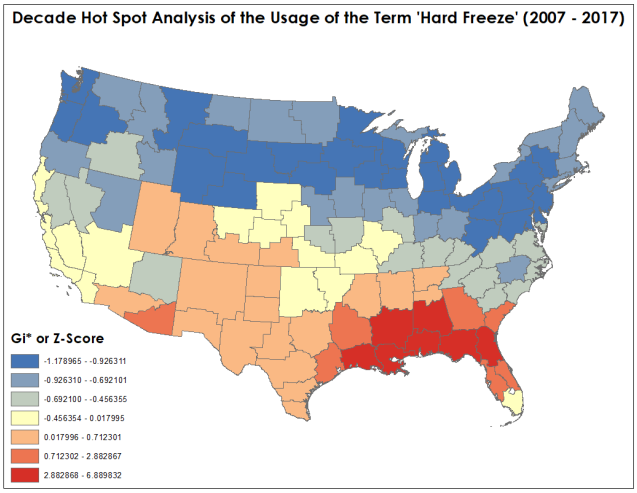

Fig 1: Hot spot analysis results for decadal 'hard freeze' term usage. Each polygon represents the WFO's coverage area. Red and orange colors indicate high z-score values while blues indicate low z-score values.

The distribution of many atmospheric phenomena are often related to or dependent on spatial location. Since AFDs describe the issues occurring in a WFO's coverage area, it can be assumed that these geospatial relationships would also be present within the AFD extraction data. Building on Tobler's first law of geography, which states that "everything is related to everything else but near things are more related than distant things" [8], spatial statistical methods can be used to identify geospatial hot spots within the AFD extraction data. Hot spot analysis looks at each feature in a dataset within the context of neighboring features and identifies statistically significant hot spots where features with high values are surrounded by other features with high values [9]. For this analysis, features are each WFO's coverage area while values are the number of times a term is used.

To conduct the hot spot analysis, the AFD extraction data was subsetted to the winter season where a winter season is defined as occurring from the first of October to the 31st of May. Data was analyzed for each winter season occurring between 2007 and 2017. Additionally, a decade long climatology of the winter seasons was also created. This hot spot analysis was limited

to the continental United States due to incomplete AFD data for Alaska, Hawaii, Guam and Puerto Rico. For each WFO, the number of times the term 'hard freeze' was used was summed up for each winter season and for the entire decade. The summed valued was then used as a weight when calculating spatial statistics and for measuring spatial clustering within the data. An assessment for patterns both across the entire study area for global statistics and amongst the various features for local statistics was conducted for each winter season and the entire decade. The results of the 2007 - 2017 decadal hot spot analysis are shown in Figure 1. The results show statistically significant hot spots occurring in the southeastern United States with positive Z-scores occurring across the southern portions of the United States.

Hot spots or positive z-score values of the term 'hard freeze' are inversely related to winter temperature trends spatially. While the northern United States faces more days of hard freezing temperatures, the term 'hard freeze' is used minimally or not at all in those areas. Southern areas of the United States use the term more frequently but do not face as many days of hard freezing temperatures. These statistics imply that it is not the number of days of hard freezing temperatures that matter to forecasters but the significance of the occurrence of those hard freeze temperature events to the office's customers. While forecasters are tasked to report scientific and factual information, each forecast is, in the end, curated to focus on weather issues that impact customers. This curation effect is illustrated in the 'hard freeze' hot spot analysis.

Understanding the decadal hot spot patterns for the term 'hard freeze' enables the identification of interesting outlier events in the yearly winter season data. A hot spot analysis was conducted for each winter season from 2007 to 2017. This hot spot analysis shows hot spots of the usage of the term 'hard freeze' occurring in the southeastern United States for seven out of ten winter seasons. The three remaining anomalous seasons indicate hard freeze events happening late in the winter season. For example, the 2011 - 2012 winter season shows a hot spot in the Buffalo, New York area. This hot spot is due to a late freeze event occurring around April 28. Mid to late April is the recommended time to plant spring vegetable crops in the Buffalo area and would therefore be of concern to customers in the Buffalo WFO's coverage. Similarly, the 2012 - 2013 winter season indicates a hot spot in the Albuquerque, NM and central Texas area. This is also due to a late freeze event occurring around April 18. Many spring vegetable crops have already been planted in the area at this time in this area. This preliminary year to year analysis suggests that these types of analyses could be used to look for anomalous events in the forecast discussions.

3.2. 'Hard Freeze' Event Identification Use Case

Identifying hard freeze events is extremely important to the agricultural community including "large and small growers, agricultural agencies, extension agents, retailers and commercial enterprises" [10]. These groups leverage remote sensing data to better understand how crop health is affected

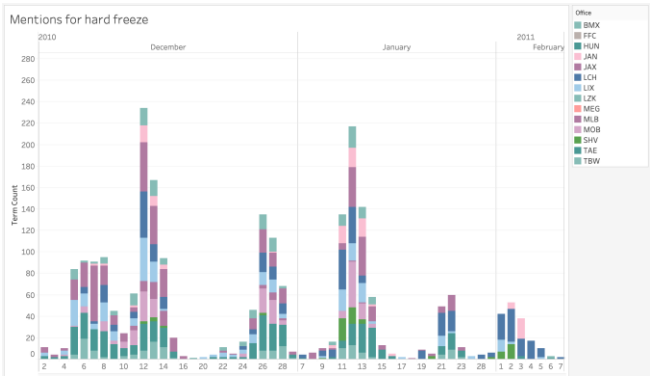

Fig 2.: Graph of 'hard freeze' term usage for the 2009/2010 winter season. The offices have been subsetted to the southeastern United States based on a hot spot identified through geospatial analysis. The offices are color coded and listed at right. The horizontal axis represents the day of the month. Peaks in term usage can clearly be identified and used to identify hard freeze events.

by hard freeze events and how these events will subsequently affect crop yields. Identifying significant hard freeze events in the past and immediately identifying hard freeze events as they occur will help "to identify patterns of impacts on agriculture with a view to improving impact assessments, including impact forecasting, mitigation, adaptation and emergency operations" [11]. For example, hard freeze events are potentially significant events for citrus growers in the southeastern United States. Citrus crops are a 9-billion-dollar industry [12] with around 75% of the United States' citrus production occurring in the southeast alone [10]. Therefore, identifying hard freeze events is essential to the southeastern citrus industry and to the broader agricultural community.

Using the annual or decadal hot spot analysis method described previously, regions of interest can be identified for specific events. In the case of the 'hard freeze' term, the southeastern United States is a consistent hot spot throughout the 10-year study period (Fig. 1). A monthly analysis of the data subsetted to this region reveals several potential cases over the study period for the region of interest. For example, several peaks in term usage can be seen in the winter 2009/2010 season (Fig. 2) when sustained cold temperatures caused significant damage to citrus crops throughout the state of Florida [12]. Once these dates are identified, relevant remote sensing data, including Landsat imagery, MODIS imagery and NDVI derived products [10]

can then be more quickly discovered and used for agricultural applications.

4. LESSONS LEARNED & FUTURE WORK

4.1. Lessons Learned

Scraping the large number of AFD web pages on the Iowa Environmental Mesonet website was challenging because it was easy to miss pages. Similarly, data of this volume is difficult to check for quality. While broad exploratory checks may not always find data gaps, more specific dives into use cases often revealed data that was missed.

Since the AFD documents are written by humans, it is unclear how each author's perception of relevant events, understanding of key terminology definitions and writing styles affect the results of any analysis conducted using this alternative data. Each author and perhaps even each WFO makes an assumption of relevance for each forecasted event [13]. Additionally, each author/WFO has different thresholds of concern for various events in the coverage area [13]. The concepts of relevance and thresholds of concern are clearly evident in the hot spot analysis maps generated for the term 'hard freeze' (Fig. 1). While most places in the U.S. experience temperatures that induce hard freeze conditions, only offices that are concerned with the impacts of a hard freeze explicitly mention the term in the forecast discussions. While the uncertainties and ambiguities due to human communication make drawing rigorous scientific conclusions from the AFD data impractical, the data is still useful as an alternative data source for identifying important atmospheric events and trends. Forecasting leverages heuristics based decision making that is built on experience [14] therefore making the AFDs a reliable source for identifying events that are of interest to the remote sensing community.

4.2. Conclusions and Future Work

In this paper, we presented the creation of an alternative Earth science dataset from the NWS's AFD documents. We also presented one methodology, using the term 'hard freeze,' for identifying patterns and events within the AFD data. A motivating example related to agriculture and hard freeze events was demonstrated. While we have presented one example of leveraging the AFD dataset to discover interesting Earth science events, we can imagine any number of effective approaches to interrogating the AFD data. In the future, we plan to both explore usage patterns of other relevant event terms with the AFD data and to also investigate other data interrogation techniques with a focus on methodologies that can be automated in order to scale up the use of these data.

5. REFERENCES

[1] S. Nativi, P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia and O. Ochiai, "Big Data challenges in building the Global Earth Observation System of Systems," in *Environmental Modelling & Software*, vol. 68, pp. 1 - 26, Jun. 2015.

[2] J. Blumfield, "Getting Ready for NISAR—and for Managing Big Data using the Commercial Cloud,": https://earthdata.nasa.gov/getting-ready-for-nisar, Oct 2018.

[3] National Weather Service, "WFO Public Weather Forecast Product Specifications," http://www.nws.noaa.gov/directives/sym/pd01005003curr.pdf, Aug. 2017.

[4]Iowa State University, Iowa Environmental Mesonet, https://mesonet.agron.iastate.edu/.

[5] American Meteorological Society, "Glossary of Meteorology History," http://glossary.ametsoc.org/wiki/Wiki/History, Jun. 2000.

[6] X. Duan, J. Zhang, R. Ramachandran, P. Gatlin, M. Maskey, J. Miller, K. Bugbee and T. Lee, "A Neural Network-Powered Cognitive Method of Identifying Semantic Entities in Earth Science Papers," in *IEEE International Conference on Cognitive Computing*, 2018.

[7] American Meteorological Society, "Hard Freeze," http://glossary.ametsoc.org/wiki/Hard_freeze, Jan. 2012.

[8] W. Tobler, "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic Geography*, vol. 46, pp. 234 - 240, Jun. 1970.

[9] ESRI ArcMap Documentation " How Hot Spot Analysis (Getis-Ord Gi*) works," http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm

[10] R. Shrivastava and J. Gebelein. "Land cover classification and economic assessment of citrus groves using remote sensing," *ISPRS Journal of Photogrammetry & Remote Sensing*, vol. 61, pp. 341 - 353, Nov. 2006.

[11] R. Guerreiro, "Accessibility of Database Information to Facilitate Early Detection of Extreme Events to Help Mitigate Their Impacts on Agriculture, Forestry and Fisheries," *Natural Disasters and Extreme Events in Agriculture: Impacts and Mitigation*, Springer, 2005.

[12] P. Fletcher, "Florida citrus growers reel under prolonged freeze," https://www.reuters.com/article/us-florida-citrus-freeze/florida-citrus-growers-reel-under-prolonged-freeze-idUSTRE60913020100112, Jan 12, 2010.

[13] B. Fischoff, "What forecasts (seem to) mean," *International Journal of Forecasting*, vol. 10, pp. 387 - 403, 1994.

[14] N. Stuart, D. Schultz and G. Klein, "Maintaining the Role of Humans in the Forecast Process: Analyzing the Psyche of Expert Forecasters," *Bulletin of the American Meteorological Society*, pp. 1893 - 1898, Dec. 2007.