STATS401-Final Project

Sissi Wang, Wanqi Hu, & Yijia Xue

May 2022

**S****patial and temporal distribution of**

**cherry blossom-viewing date in the US based on Twitter data**

**1 Introduction**

Phenology has become the "diagnostic fingerprint" of global climate change (Cleland et al. 2007). As an essential phenological indicator, the blossom-viewing date is a sensitive and intuitive reflection of the environmental changes. However, most blossom phenological site observation data show a pattern of incompleteness, which can hardly support the overall spatial and temporal analysis (Liu et al., 2019). As a result, the big data from Twitter, a widely used social media platform, has become a complementary source to solve this problem.

Different flowers’ blossom-viewing phases vary significantly. To more accurately show the influence of climate change on the viewing dates, this project tends to focus on one specific kind of blossom with a short blossom-viewing phase. In Figure.1, the average starting date and ending date of the ten most popular ornamental flowers’ (Frankie et al. 2005) blossom phases are shown on the x-axis, and the length of the phase is shown on the y-axis in the unit of days. From Figure.1, the cherry blossom shows the shortest blossom-viewing phase, from late March to middle April.

图表, 折线图

描述已自动生成

Figure.1 Comparison of the blossom-viewing phases of ten different flowers

Thus, this project takes the cherry blossom as the research object, Twitter as the data crawling source, and the United States as the geographical research scope to analyze cherry blossom viewing dates' spatial and temporal distribution patterns.

**2 Data**

The main data source for this project is Twitter. Tweets are scrawled using snscrape (Blair et al. 2021), a scraper for social networking services based on Python. With the keyword “cherry blossom,” 44915 pieces of tweets are found from 2011 to 2021 with the location in the US. The information for each tweet includes the time (ordinal), location (categorical), and content (categorical).

The data of different flowers’ blossom-viewing phases (ordinal) from Wikipedia (Wikipedia 2004) and data of elevation distribution (numerical) in the US from Google earth (Lisle 2006) are used as complementary data in this project.

**3 Methods**

*3.1 Data cleaning*

After scrawling, the dataset was first cleaned by deleting the data without information of state in location and further filtered to get rid of the keywords like “Starbuck,” “shampoo,” and “cake,” with which tweets are not describing the true blossom. The final version was manually checked.

*3.2 Analysis of spatial and temporal distribution*

Spatial distribution is described in two parts: latitude and altitude. Linear regression was done respectively with each state’s average viewing date among 11 years. Temporal distribution is based on year. The viewing date is compared over the years to find whether it was significantly advanced or backward.

**4 Results**

*4.1 Dataset*

13132 pieces of tweets are left after cleaning. Figure.2 is a word cloud of the contents of tweets. From the figure, “Washington-DC” and “festival” show high frequency, which is in correspondence to a spring celebration event called National Cherry Blossom Festival in Washington, D.C (McClellan 2005). Other keywords, like "spring,” “park,” and “photo,” illustrate that, after cleaning, most tweets describe the blossom-viewing, thus verifying the rationality of using Twitter data to determine the cherry blossom viewing phase.

The interactions for this figure are when hovering over the words

1. There will pop up a tooltip that shows the frequency of the selected word.

2. The chosen word will be blown up, rotated, and changed the color.

图片包含 游戏机, 风筝, 飞行, 花

描述已自动生成

Figure.2 Word cloud of tweet contents after data cleaning

*4.2 Spatial distribution*

Blossom-viewing dates vary in different states of the US. Figure.3 is a map demonstrating the starting cherry blossom viewing date in each state. From the map, it can be concluded that the viewing date slips from south to north in general, and the date in the west is slightly higher than that in the east due to the elevational difference. Taking 2021 as an example, the earliest date is March 6th in CO, and the latest date is April 7th in ID. The viewing date varies by more than 32 days.

The interactions for this figure are:

1. The user can select the state and see the exact viewing date by clicking.
2. The user can choose the year by dragging the button to see different years’ distribution.

地图

描述已自动生成

Figure.3 Map with the starting cherry blossom viewing date over 11 years

(take 2021 as an example)

Figure.4 is a regression plot between the latitude of each state and the average viewing date in that state in 2021. The MAE of the regression is 4.71, which suggests a clear relationship between the latitude and viewing date. Based on the regression, the viewing date for each degree increase in latitude will be delayed by 0.70 day.

图表, 散点图

描述已自动生成

Figure.4 Regression between latitude and average viewing date in 2021(Not included in the five main visualizations)

Figure.5 is another regression plot between the altitude of each state and the average viewing date in that state in 2021. The MAE of the regression is 5.35. Based on the regression, the viewing date for each meter increase in altitude will be delayed by 0.0017 day.

图表, 散点图

描述已自动生成

Figure.5 Regression between altitude and average viewing date in 2021(Not included in the five main visualizations)

*4.3 Temporal distribution*

Blossom-viewing dates show different patterns in different years. Figure.6 is a lollipop plot with the starting blossom viewing date and the ending viewing date among all states each year. From the figure, the viewing date gradually became earlier from 2011 to 2021. The difference between the starting date between 2011 and 2021 is about seven days, and the difference between the ending date is 14 days.

The interactions for this figure are:

1. The user can filter the visualization by the elevation height of each state.
2. The user can click on the pink ball to get the starting and ending viewing date information for each year.

图表, 条形图

描述已自动生成

Figure.6 Starting and ending blossom viewing date in each year

The change over the years varies in different states. Figure.7 is a circular bar chart with the change of the earliest viewing date from 2011 to 2021 in different states. The larger the difference is, the farther the top of the bar is from the baseline, and the more saturated the bar is. If the date becomes earlier, the bar is pink, while if it becomes later, the bar is blue. From the figure, the viewing date in 43(82%) states becomes earlier from 2011 to 2021, and the averagely becomes 1.16 days earlier per year. Date in 3(6%) states becomes later and becomes 0.21 days later per year. The largest difference is in 2020, which is 1.83 days earlier.

图表, 雷达图

描述已自动生成

Figure.7 Change of the starting viewing date from 2011 to 2021 in different states

The interactions for this figure are:

1. The user can select a year by dragging the year selector bar.
2. The user can hover the mouse over each bar to get the exact variation number of each year.

**Discussion**

***What***

The data used in this project is crawled tweets with the keyword ‘cherry blossom’ and different flowers’ blossom-viewing phases. The output of the project is 5 visualizations with design idioms: 1. Line chart 2. Word cloud 3. Choropleth Maps 4. Lollipop chart 5. Circular bar chart.

***Why***

Visualizations in this project intend to show why and how we use the tweets data to explore the temporal and spatial distribution of the cherry blossom viewing date.

***How***

The audience can grasp the information in the following user tasks

The user tasks for this visualization include: 1) Retrieve value (blooming phase, word frequency, starting and ending blossom viewing date, number of variation days); 2) Determine range for the blooming phase; 3) Filter the data by elevation; 4) Find extreme (word with the highest frequency, state with the largest variation viewing date, the earliest viewing date); 5) Characterize distribution (temporal distribution, spatial distribution).

The main strength of the project is that the visualizations are innovative. For example, we draw ten types of flowers and use them in Vis.1 to make it more vivid. We use three types of visualization tools (Plotly, D3.js, and Observable) to fully explore the efficiency of different visualization tools. What’s more, the interactions are served for visualization and research goals. For example, in Vis.3 and Vis.5, the interaction allows the user to see the temporal changes of the cherry blossom and provide intuitive comparisons between years. However, since the number of the data is limited, the estimation of the actual blooming date using the date of the tweets which contain the keyword ‘cherry blossom’ is coarse. The possible resolutions for this limitation are 1. Enlarge the data by crawling data with other keywords and different social platforms. 2. Find official phenological observation data for the cherry blossom.

**Conclusion**

To conclude, extracting cherry blossom-viewing date from Twitter is effective and reliable. The cherry blossom viewing date in the US has a clear spatial and temporal distribution pattern. Spatially speaking, the viewing date correlates to latitude and slightly correlates to altitude. Temporally speaking, the viewing date becomes earlier from 2011 to 2021, which provides evidence for global warming.

Word Count: 1506

**Roles and Responsibilities:**

Sissi Wang: Data crawling, data visualization of Vis.1, report writing and poster making.

Wanqi Hu: Data crawling, data cleaning and data visualization of Vis.3, Vis.4.

Yijia Xue: Data cleaning, data visualization of Vis.2,Vis.5 and report revising.

References

Blair, Johnna, Chi-Yang Hsu, Ling Qiu, Shih-Hong Huang, Ting-Hao Kenneth Huang, and Saeed Abdullah. 2021. ‘Using Tweets to Assess Mental Well-Being of Essential Workers during the Covid-19 Pandemic’. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–6.

Cleland, Elsa E, Isabelle Chuine, Annette Menzel, Harold A Mooney, and Mark D Schwartz. 2007. ‘Shifting Plant Phenology in Response to Global Change’. *Trends in Ecology & Evolution* 22 (7): 357–65.

Frankie, Gordon W, Robbin W Thorp, Mary Schindler, Jennifer Hernandez, Barbara Ertter, and Mark Rizzardi. 2005. ‘Ecological Patterns of Bees and Their Host Ornamental Flowers in Two Northern California Cities’. *Journal of the Kansas Entomological Society* 78 (3): 227–46.

Lisle, Richard J. 2006. ‘Google Earth: A New Geological Resource’. *Geology Today* 22 (1): 29–32.

Liu Jun, Wang Shenghong, Jin Mengmeng, and Li Ningxin. 2019. ‘Reconstruction of Peach Blossom-viewing Date of China Using Weibo Big Data.’ Geographical science 39 (9): 1446–54.

McClellan, Ann. 2005. *The Cherry Blossom Festival: Sakura Celebration*. Bunker Hill Publishing, Inc.

Wikipedia. 2004. *Wikipedia*. PediaPress.