

Finding Napa

Rong Feng
New York University
New York, USA
rf1316@nyu.edu

Cody Gilbert
New York University
New York, USA
cjl507@nyu.edu

Gary Ng
New York University
New York, USA
yjn214@nyu.edu

Abstract—

The time-tested traditional methods of wine production have begun to be impacted with the advent of rapid climate change. This analysis seeks to quantify and predict those changes to allow grapevine and wine stakeholders to find new regions for growing grapes in a climate-altered future that best replicate the current conditions of the Napa Valley region. This analysis aggregates weather, soil, and sunlight data over a set of US geographic regions and time periods to select areas that best replicate those of Napa Valley given a trend of climate change.

Keywords—analytics, grapes, vineyards, climate change

I. INTRODUCTION

Wine production has taken place in the famous Napa Valley for over 150 years. Today wine from the region is known world-wide, and commands a premium compared to other US regions. With the advent of rapid climate change the time-tested traditional methods of wine production have begun to change. Already farmers up and down the Valley are starting to plant different varieties as compared to the last 150 years.

Outside of this valley, production is also booming in regions farther North, in Oregon and Washington, which now have become just slightly warmer to see better vintages coming out of that region is growing in popularity. Despite wine production's traditionally unscientific nature, the quality of grapes is governed by all that occurs above and below the soil line. As a result, this ambitious analytic seeks to identify other regions in the United States that have high similarity to Napa Valley in soil, weather, and sunlight (solar irradiation), where climate can be expected to change to that of Napa Valley in the early to mid-20th century. Therefore, this analysis will attempt to find locations that are expected to match the growing conditions of Napa Valley over time.

II. MOTIVATION

The impacts of climate change are far-reaching and severe for all aspects of life and all economic sectors. Of particular impact are the effects of altered climate patterns on

agricultural development and the resulting economic effects on farmers, investors, and all downstream consumers of agricultural goods. Quantifying the impact of climate change on goods with such an array of stakeholders is of great importance, and the ability to react flexibly to changes in environmental conditions reaps substantial benefits.

In this analysis, we intend to provide a tool and a guideline for US vineyard stakeholders. This analysis will provide an outlook to farmers and investors on the best locations for future growth and land development. It is therefore our goal to contribute to the mitigation of detrimental climate change effects that impacts a large population of people.

III. RELATED WORK

This project focused on three areas of interest as they relate to climate change and its impact on grapevine growth.

Temperature and Precipitation

Recent works have suggested that while the premium status of renowned wine producing regions such as Napa, Bordeaux, Burgandy have come very slowly over time [5], those luxury statuses do shift slowly over time. As the quality of the grapes change with the climate they are grown in, there is risk to the downside for these elite regions. However, in this risk there is opportunity for speculators to find potential new regions that can overtake old names in the world of premium wine.

This change is already occurring in existing regions around the world. In particular, a German vineyard is experience great success with varietals not traditionally grown there [6] and has reported great excitement to be able to plant previously unsuccessful types.

Grape growing for wine production is a delicate process and small changes in the environmental factors can lead to large differences in results. Another author reports on this narrow niche of the environment and raises alarms for the future of existing vineyards in the face of rapid climate change [7].

Solar Effects

A basic conclusion from elementary earth science is the fact that sunlight affects the growth of plants via several reactions, the most vital of which is photosynthesis. Despite this bedrock foundation of the botanical sciences, studies analyzing the effects of solar variance on grape growth which quantified levels of lighting are few and far between. Older research produced generalized results that various cultivars exhibit significant responses to light, temperature, and daylength to vine growth [2] and berry weight and composition [3].

A recent report from the University of Iran [1] investigates the effect of light on the growth of Crimson Seedless grapes under various light conditions. The researchers in [1] exposed the axillary buds of the Crimson Seedless grape variety to several light frequencies and intensities. Light frequency was shown to increase the growth rate, with red and visible light displaying the fastest growth rates. Light intensity dramatically impacted photosynthesis. Low-intensity light ranges that might hinder the growth of buds through a reduction in photosynthesis were not included in the study, however high-intensity light was shown to reduce photosynthesis through dynamic and chronic photoinhibition. In high-intensity light, heat-distribution effects reduce photosynthesis in the short term (dynamic photoinhibition) and can lead to permanent damage to the photosynthesis system over the long term (chronic photoinhibition). Based on these conclusions, grapes grown in areas with a higher light intensity than suited for the grape variety can be determinantal on grape growth.

Soil Characteristics

Soil characteristics are also widely acknowledged to have significant influence over wine quality. In particular, Burns (2012) highlighted particular physical characteristics are thought to matter the most: 1) soil depth, which affects the rooting depth and risk of waterlogging, 2) soil pH, which affects the availability of nutrients and root growth, 3) soil texture, which affects the soil's ability to hold water, drain, and cultivate, 4) soil color, which affects its temperature, and in turn the speed at which grapes mature, and 5) availability of specific macronutrients, such as phosphorus, sulfur, and calcium.[8]

IV. DESIGN AND IMPLEMENTATION

A. Design Details

This analysis was built using the data flow structure as shown in the diagram in Figure 1.

RBDA-SU19: Finding Napa Project

Design Diagram

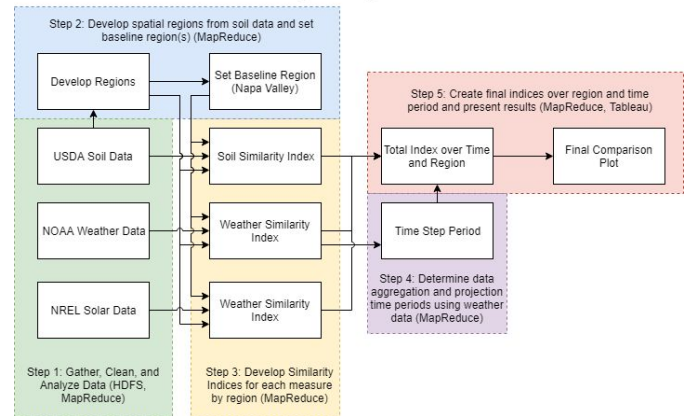


Figure 1: Analysis Design Diagram

The analysis was divided into several steps:

1. Soil, weather, and solar irradiance data sets were collected from their respective sources and stored, cleaned, and profiled on the NYU Dumbo Hadoop cluster using HDFS and Hadoop MapReduce.
2. The soil data, being the most regionally sparse, was used to define US geographical regions in which all three measures of similarity would be produced. The baseline Napa Valley region was defined from this set.
3. Soil, weather, and solar irradiance data were aggregated by the regions developed in step 2, a subset of features were selected, and a similarity measure to the baseline region were calculated.
4. The weather data, being the only data set that could be defined over a temporal dimension, was used to define time intervals over which similarity measures could be calculated.
5. Similarity measures over time and spatial region were weighted and aggregated over all the data sets, creating a single similarity index over time and space, and then plotted on a US map for each time interval to determine the regions that hold the most similarity to Napa Valley.

Quantification of Similarity

Cosine similarity was used as the index of similarity of a vector of numerical features to some baseline vectors. This well-known measure was chosen because of its ease of implementation in MapReduce programs, and its ubiquitous use in data analytics. Cosine similarity also produces indices of common ranges and magnitudes which permit comparison and weighting of data taken from a variety of sources. An index of cosine similarity from each of the weather, solar, and soil data sets can be easily weighted to produce the final similarity measure used for the conclusions of this analysis.

B. Description of Datasets

Weather Data

NOAA publishes hourly data on all global weather stations for as far back as 1901. However, the data before the 90s were spotty at best and error prone in some of the sampled data that we saw. Since other datasets further down only include US, this dataset also only used US information and excluded those elsewhere. The dataset included numerical fields on temperature, humidity, pressure, cloud coverage, wind speed and precipitation. The data was downloaded directly using FTP and amounted to 270 GB on disk. The files are fixed-width raw text files with each row representing a particular weather on a particular hour on a given day, followed by numerical data fields as described above. The data was loaded into Hive and aggregated to the month-station level by averaging. We felt that by averaging over the month, it gives us a smoother feature that represented the broad weather patterns of a region without the daily random fluctuations of weather being overly factored into the feature.

Finally, an alternate dataset that represents potential weather in the future was constructed by adding a flat 5 degrees Celsius to the raw data to simulate what might happen in the next 80 years. The 5 degrees was a more extreme scenario than the “do-nothing” scenario as projected by consensus studies as seen in Figure 2. The reason that a more extreme scenario was picked was to magnify the cosine similarity changes attributable to the changing temperature so that they are easier to identify.

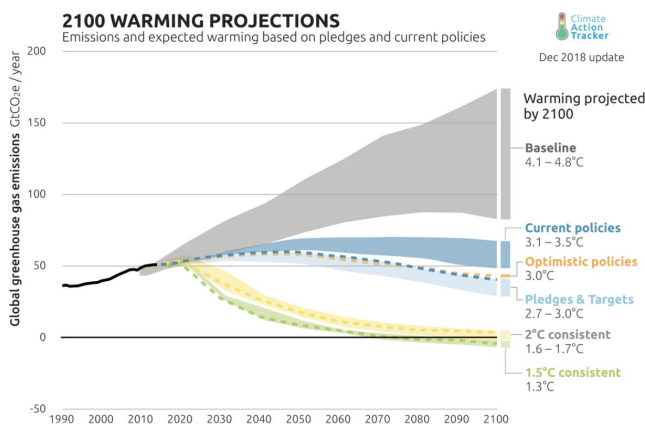


Figure 2: Projected Global Temperature Paths
[climateactiontracker.org]

Solar Irradiance Data

Data that quantifies the sunlight conditions in which a grapevine is expected to grow were taken from the United

State National Renewable Energy Laboratory (NREL) National Solar Radiation Database (NSRDB). The NSRDB comprises 30-minute solar and meteorological data for approximately 2 million 0.038-degree latitude by 0.038-degree longitude surface pixels within the US region of interest from 1998-2017. These data are modeled with the Physical Solar Model Version 3 (PSM3). Additional information regarding the NSRDB models, methods, and validation can be found in [4].

The primary NSRDB feature used in this analysis is the modeled Global Horizontal Irradiance (GHI) that estimates the total amount of direct and diffuse solar radiation received on a horizontal surface. This feature was determined to best quantify the amount of direct and indirect sunlight to which grapevine located in the region would be exposed. To smooth variations in solar irradiance over time, GHI was aggregated by month. As the studies in mentioned in Section III show that both the total exposure and overexposure have strong effects on plant growth, the mean GHI and maximum GHI were selected as features for each region and monthly interval.

NSRDB data was collected using API calls from a local Python script. Each call produced a yearly collection of solar data in 60-minute intervals for a given latitude-longitude point. The NSRDB API allows only 300 API calls per day from a single access key, which for the over 3000 soil regions limited the amount of data that could be collected for each region. Data was only collected for the 2017 year in all regions because of this restriction. The temporal sparsity of the data was considered acceptable, as year solar irradiance variation is assumed to be small in comparison to the more erratic weather data.

Data was collected and stored on the NYU Dumbo Hadoop cluster, profiled, and the GHI mean and maximum were aggregated by month and region using Hadoop MapReduce. The data was then normalized and a cosine similarity measure was applied to each region with respect to the baseline region.

Soil Characteristics Data

The Natural Resources Conservation Service (NRCS) of the U.S. Department of Agriculture (USDA) provides public access to a database containing geolocation data of over 3000 soil areas and very granular data on their physical characteristics, such as consistency, fragmentation of soil, size and shape of the fragments, distribution of soil that pass through various sieves, sand content, clay content, moisture level, concentration of various chemicals, pH level of the water found, etc.

One challenge of the dataset is the sheer level of technical detail provided, with over several hundreds of soil characteristics surveyed. After reviewing the literature and documentation, we selected ten characteristics as our soil features, namely: ph1to1h2o (pH), cec7 (amount of readily

exchangeable cations), awc (water content), om (decomposed residue), slope, elev (elevation), tfact (erosion feasibility), claytotal (clay composition), silttotal (silt composition), and sandtotal (sand composition).

Another challenge is NRCS defines four levels of soil aggregation:

1. an *area* is composed of several map units
2. a *map unit* is composed of several components
3. a *component* is composed of several horizons/layers
4. a *horizon/layer* is the most granular unit

For instance, “Napa County, California” is a soil area, but it contains X map units, Y components, and Z horizons. Because most of the soil characteristics are provided at the horizon- and component- level, they need to be aggregated up the soil hierarchy to form our dataset. For our purpose, horizons are aggregated to components based on their depth, components are aggregated to map units based on a handy attribute called “component percent of map unit” provided by NCRS, and map units are aggregated to areas based on their number of acres. Hive queries are used to perform these aggregations.

V. RESULTS

[This section will be added in the final version of this paper]

VI. FUTURE WORK

[This section will be added in the final version of this paper]

VII. CONCLUSION

[This section will be added in the final version of this paper]

ACKNOWLEDGMENT

[This section will be added in the final version of this paper]

REFERENCES

1. F. Fallah, D. Kahrizi. Effect of light spectrum and intensity on growth of grape (*Vitis vinifera*) under in vitro conditions. Journal of Applied Biotechnology Reports, Tehran, Iran, February 2017.
2. M.S. Buttrose. Fruitfulness in grape-vines: The response of different cultivars to light, temperature and daylength. In *Vitus: Journal of Grapevine Research* Vol 9 No 2. Siebeldingen, Germany. 1970.
3. N.K. Dokoozlian, W.M. Kliewer. Influence of Light on Grape Berry Growth and Composition Varies during Fruit Development. In *Journal of the American Society for Horticultural Science*, Volume 121, Issue 5. September 1996.
4. M. Sengupta, et. al. The National Solar Radiation Data Base (NSRDB). In *Renewable and Sustainable Energy Reviews*, Volume 89, Pages 51-60. June 2018.
5. Krysta Suzanne Gingue, "Wine About It: How Climate Change is Affecting International Wine Markets", University of New Hampshire Scholars' Repository <https://scholars.unh.edu/cgi/viewcontent.cgi?article=1419&context=honors>
6. Cheslow, Daniella. "Climate Change Ripens Prospects For German Winemakers." NPR. November 17, 2017. Accessed May 21, 2018. <https://www.npr.org/sections/thesalt/2017/11/17/564099490/climate-change-ripensprospects-for-german-winemakers>
7. Jones, Gregory V. "Climate Change: Observations, Projections, and General Implications for Viticulture and Wine Production." *INFOWINE.COM Internet Journal of Viticulture and Enology*6 (April 2007): 1-12. 2007. <https://www.infowine.com/intranet/libretti/libretto4594-01-1.pdf>
8. Burns, Scott. “The Importance of Soil and Geology in Tasting Terroir with a Case History from the Willamette Valley, Oregon.” In: Dougherty P. (eds) *The Geography of Wine*. Springer, Dordrecht. 2012. https://link.springer.com/chapter/10.1007/978-94-007-0464-0_6
9. USDA Natural Resources Conservation Service Web Soil Survey Database. <https://websoilsurvey.sc.egov.usda.gov>.