Journal of Quantitative and Qualitative Research

2025 (Vol 3)

FRANKLIN EDUCATION
FOUNDATION

Contents

MESSAGES FROM MENTORS 2

COMPARISON OF DIFFERENT ROCKET FUEL AND OXIDIZER MIXTURES AND HOW THEY AFFECT ROCKET ENGINE THRUST (LUJIA CHEN) 4

EXAMINING THE RELATIONSHIP BETWEEN RETURN ON EQUITY AND RETURN ON ASSETS AS INDICATORS OF CORPORATE LONGEVITY (XIONG YIN) 12

A UNIFIED DESIGN AND EVALUATION FRAMEWORK FOR LIQUID ROCKET ENGINES INCORPORATING ENVIRONMENTAL AND ECONOMIC TRADEOFFS (CAYDEN SHI) 21

QUANTIFYING BALLOT REDISTRIBUTION TURBULENCE IN RANKED-CHOICE VOTING THROUGH CLASSICAL AND TROPICAL WASSERSTEIN DISTANCES (KEVIN LIU AND AUSTIN WEI) 35

PREDICTING RESIDENTIAL PROPERTY VALUES IN SAN MARINO USING A MULTIVARIABLE REGRESSION MODEL (ZHIYING (MICHELLE) YIN) 51

Research has become all the more important for tackling multifaceted new challenges and exploring exciting opportunities in our increasingly fast changing world. It has been a vital part of the mission of higher education in the United States. In institutions of higher education, especially top universities, research is becoming an integral part of the educational experience of undergraduate students. At the University of California, for example, many students participate in a wide range of research programs and activities.

The Franklin Research Internship Program (FRIP) offers high-school students a precious educational opportunity to gain valuable research experience through their exciting research projects. It is a well-designed program that covers the process of research (including methodology, ethics, data collection, critical thinking, etc.) plus technical skills (in areas like software programming, writing, and data analysis). The program was not only focused on STEM fields but also made a fruitful effort to expose students to research in the Humanities and Social Sciences. This is significant because many of the challenges we face in our life and society cannot be solved with science and technology alone but require interdisciplinary approaches and solutions.

Having attended numerous activities of the program from its beginning to its conclusion, I think that it has been a big success. The success of the program was also due to the dedication from teachers and the professional expertise of mentors. Finally, FRIP's success reflected the diligent and fruitful efforts of students. In their research projects, students asked and explored important topics in different fields, ranging from self-driving cars to educational programs in schools. The student research papers that I have read are remarkably well-written and thoughtful, and they would earn a good grade in college classes. I am glad that the Franklin Education Foundation will publish student papers. The collection of these papers will illustrate the fruit of the extraordinary work of students and serve as a source of inspiration for other students in the future.

Professor Y. Chen
Franklin Research Internship Program Mentor

The high-school experience has a great influence on a person's life in terms of career choices, personal passions, and even the direction of one's lifelong journey. Take my own high-school experience, for example, of coming across a small book describing medicine that can treat diseases. The book illustrated the chemical composition as well as the structure of the molecules. It ushered me into a new world of structures and functions. As a result, I developed a lifelong passion in associating chemical structures with how they treat diseases and improve patients' lives.

Similarly, the Franklin Research Internship Program (FRIP) provides this kind of opportunity to high-school students. Motivated students can seek and find their own passions in humanities, sciences, technology, and mathematics. Students have gained hands-on research experience on trendy topics in a myriad of fields, including artificial neural network, aerodynamic engineering, molecular biology, human psychology, physics, and more.

I feel happy to meet some of these wonderful students. Each of them exhibited a unique set of talents. Some were able to grasp concepts better than those in a college class; others were able to author articles with excellent writing skills; and still others were able to design artificial neural networks and train them for solving real-life problems.

Assembled in this journal are the resulting papers of the projects completed by some of these talented students. I sincerely hope that this publication provides a sound platform for the career development of these students and foreshadows a bright future for their scientific research pursuits. I wish the best of success for all students who have already participated in FRIP and those who will join this program in the future.

Dr. Peng
Franklin Research Internship Program Mentor

Comparison of Different Rocket Fuel and Oxidizer Mixtures and How They Affect Rocket Engine Thrust

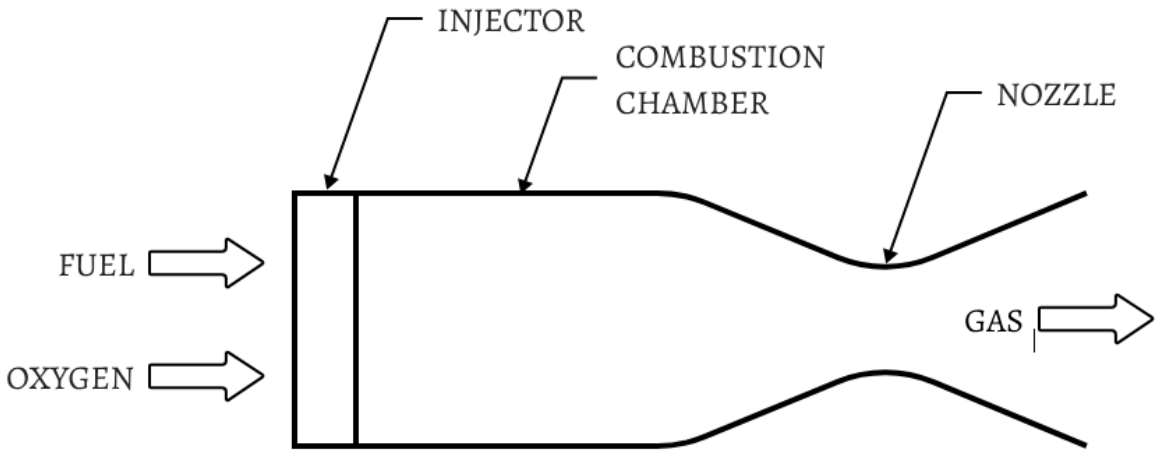
Lujia Chen

ABSTRACT

A rocket engine is designed with a combustion chamber with fuels and oxygen feeding into it and a nozzle at one end where the combusted gases are expelled to create thrust. There is a challenge to properly design rocket engines based on proper fuel and oxidizer combination. However, predicting the thrust of different fuels and oxidizers is difficult because it is impossible to know exactly how the combusted fuels and oxidizers will behave. In this work, a computational fluid dynamic model was developed to study the impact of different fuel and oxidizer combinations on the thrust of the engine. Different fuel and oxidizer combinations such as liquid oxygen and gasoline, gaseous oxygen and methyl alcohol, and liquid oxygen and hydrogen have been compared based on the thrust, pressure, and temperature. Liquid oxygen and hydrogen should be used because it has the highest thrust while having the lowest temperatures.

Introduction

A rocket is a vehicle that propels itself using exhaust that flows out at high speeds. There are two different types of rockets, solid rockets and liquid rockets. A solid rocket engine uses a solid mixture of fuel that when given a heat source will ignite and burn until all the fuel is used up. [2] A liquid rocket engine burns pressurized liquid propellants in a combustion chamber, forming hot gases that are then accelerated and ejected at high speeds out of a nozzle to impart work onto the engine. [1] However no matter what type of rocket is used, all rocket engines use Newton's third law of motion that every force has an equal and opposite reaction. This is because by the hot gases ejecting out of the nozzle, the gases will push on each other and the air around it which then pushes the rocket. [2] This research paper will focus on the latter of the rocket engine types because a model with a combustion chamber and liquid propellant parameters were used.
Below is a simple diagram of a liquid rocket engine. [1]



The fuel and oxygen are injected into the combustion chamber combusted which causes an expansion by producing gas that is then ejected from the nozzle. While all parts of the rocket engine can affect how the rocket performs, this paper will be focused on the propellants. More specifically it will be a comparison of the different propellant types and how it will impact thrust. While the different propellant types can cause different temperature combustions and generally, lower temperature is better, this paper will not cover the effects of temperature from different propellants on the rocket engine's thrust.

Methodology

To collect data and see which propellants create the most thrust, Autodesk CFD 2024 was used to simulate a rocket engine and the way a gas would travel through it. Autodesk Fusion 360 was used to model the simulated rocket engine in a 3D space. From the thrust and the total propellant weight flow rate according to [1], specific impulse or the thrust developed per weight unit (newton) of total propellant burned per second can be calculated using this formula: $I_{sp} = \frac{thrust}{total\ propellant\ weight\ flow\ rate}$. To calculate thrust, mass flow rate first needs to be converted from $\frac{g}{s}$ to $\frac{kg}{s}$ which was found by dividing the mass flow rate by 1000 and then can be multiplied to the I_{sp} value given in the table [1].

Propellant Combination Oxidizer/Fuel	Combustion Pressure, MPa	Mixture Ratio	Flame Temp (°C)	I_{sp}, sec
Liquid Oxygen & gasoline	2.068	2.5	3020	242
Gaseous Oxygen & gasoline	2.068	2.5	3170	261
Gaseous Oxygen & gasoline	3.450	2.5	3240	279
Liquid Oxygen & JP-4 (jet fuel)	3.450	2.2	3250	255
Liquid Oxygen & methyl alcohol	2.068	1.25	2860	238
Gaseous Oxygen & methyl alcohol	2.068	1.2	2880	248
Liquid Oxygen & hydrogen	3.450	3.5	2480	363
Red fuming nitric acid & JP-4	3.450	4.1	2840	238

Of each of their fuel sources, they each had different temperatures.

Results and Discussion

In the contour plot, the redder a color is, the more positive velocity in the X direction it has while the bluer it is, the more velocity in the negative x direction it has.

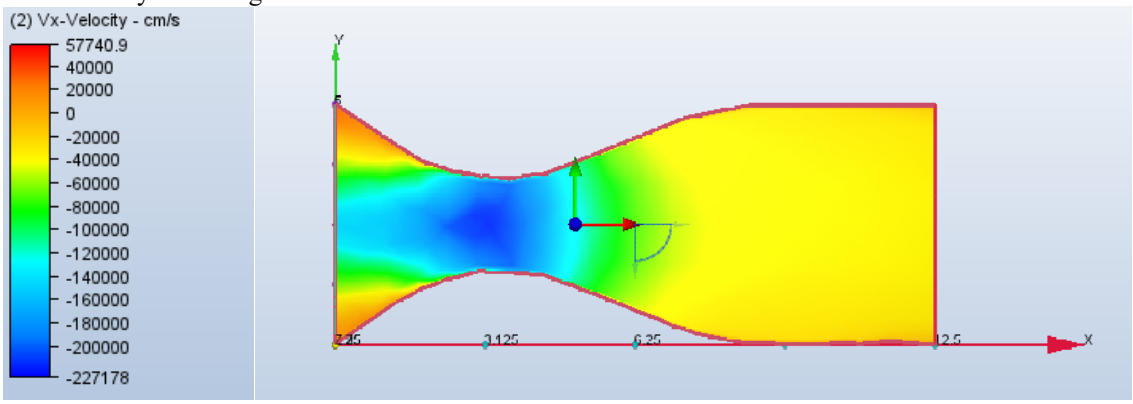


Image 1

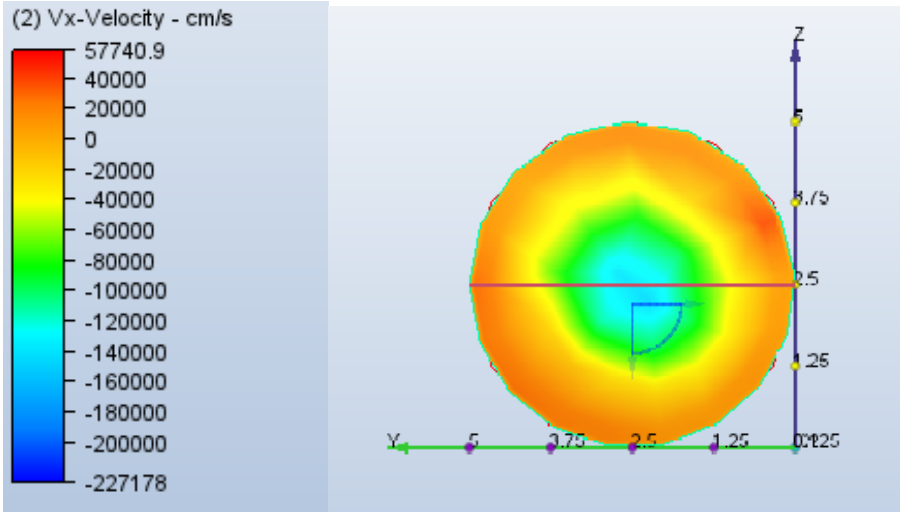


Image 2

Images 1 and 2 are for the rocket fuel types that produced 3.450 MPa or mega Pascals of pressure. Image 1 is a top-down view of the simulation showing how when the opening constricts, an area with high negative velocity forms. Image 2 is a nozzle-end view and shows how at the nozzle, the inside of the nozzle has a high negative velocity while the outside of the nozzle has a high positive velocity.

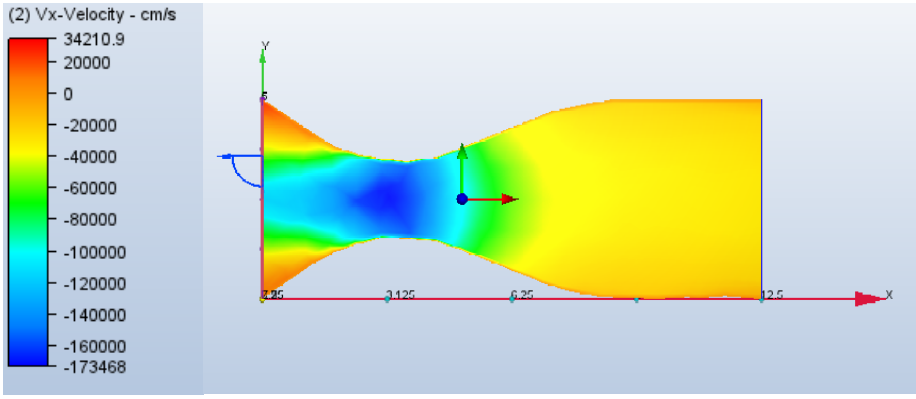


Image 3

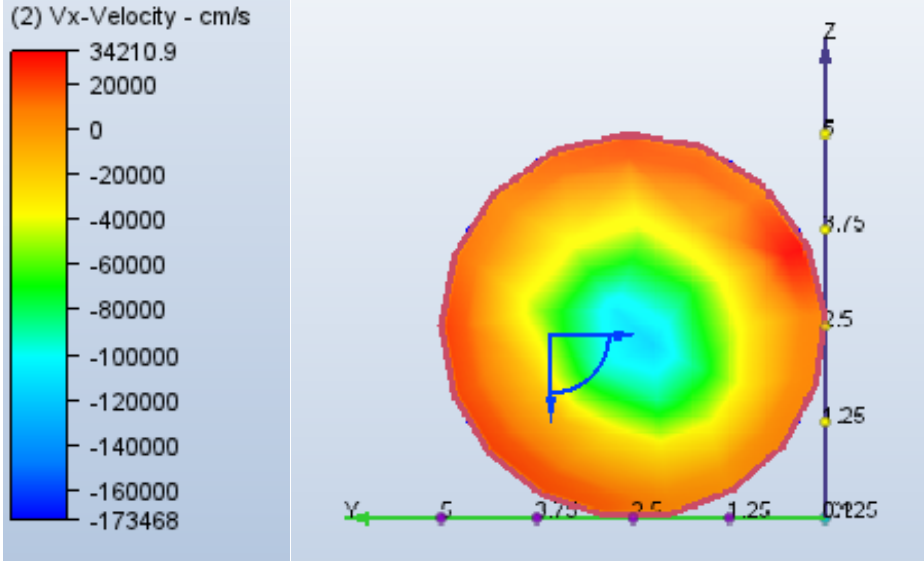


Image 4

Images 3 and 4 are for the rocket fuel types the produced 2.068 MPa of pressure. Both were run with turbulence, which is a closer model to what happens in reality. Like images 1 and 2, Image 3 is a top-down view while image 4 is a nozzle-end view, showing a similar pattern. Seeing how the simulation that was run with less pressure showed lower velocities as observed in the chart showing what color corresponds to the velocity value, the fuels that generate lower pressures will generate less thrust. This is corroborated by the data taken from the simulation and put into the chart below.

MFR: Mass Flow Rate | VFR: Volumetric Flow Rate | Vx: Velocity in the X direction |

Setup	MFR ($\frac{g}{s}$)	VFR ($\frac{cm^3}{s}$)	Vx ($\frac{cm}{s}$)	Vy ($\frac{cm}{s}$)	Vz ($\frac{cm}{s}$)	D ($\frac{g}{cm^3}$)	P ($\frac{dyne}{cm^2}$)	PF (dyne)	T (℃)	Vi (poise)	I_{sp} ($\frac{N \times s}{kg}$)	Th (N)
Red fuming nitric acid & JP-4 (jet fuel)	930.615	-508608	-26506.4	2406.91	-5725.3	0.001830	-1.013*10^6	-1.944*10^7	2840	0.000149	238	221.48637
Liquid Oxygen & hydrogen	930.615	-508608	-26506.4	2406.91	-5725.3	0.001830	-1.013*10^6	-1.944*10^7	2480	0.000149	363	337.813245
Gaseous Oxygen & methyl alcohol	794.923	434448	-22656	114.997	-2090.18	0.001830	-1580.92	-18816.8	2880	0.000149	248	197.140904
Liquid Oxygen & methyl alcohol	794.923	434448	-22656	114.997	-2090.18	0.001830	-1580.92	-18816.8	2860	0.000149	238	189.191674
Liquid Oxygen & JP-4	930.615	-508608	-26506.4	2406.91	-5725.3	0.001830	-1.013*10^6	-1.944*10^7	3250	0.000149	255	237.306825
Gaseous Oxygen & gasoline	930.615	-508608	-26506.4	2406.91	-5725.3	0.001830	-1.013*10^6	-1.944*10^7	3240	0.000149	279	259.641585
Gaseous Oxygen & gasoline	794.923	434448	-22656	114.997	-2090.18	0.001830	-1580.92	-18816.8	3170	0.000149	261	207.474903
Liquid Oxygen & gasoline	794.923	434448	-22656	114.997	-2090.18	0.001830	-1580.92	-18816.8	3020	0.000149	242	192.371366

Legend:
Vy: Velocity in the Y direction | Vy: Velocity in the Y direction | D: Density | P: Pressure |
PF: Pressure Force | T: Temperature | Vi: Viscosity | I_{sp} : Specific Impulse | Th: Thrust
The trials with more pressure resulted in higher mass and higher volume flow rates. This caused a higher thrust as I_{sp} is in the numerator and has a direct relationship with thrust. However, the since there were only two pressures given, there were only two distinct results from the simulation as the due to not running a proper temperature simulation, temperatures didn't affect anything.

Discussion

Of the many fuels and oxidizer combinations tested, it seems of the groups of them that combusted with the high and low pressure, liquid oxygen and hydrogen, and gaseous oxygen and gasoline resulted in the most thrust respectively. Because liquid oxygen and hydrogen had a greater mass flow rate and a greater I_{sp} than Gaseous oxygen and gasoline due to differing fuel and oxidizer combinations, it had a greater thrust.

While the gaseous oxygen and gasoline resulted in a lower thrust, it also did combust with lower pressure. This could result in that combination not needing as thick of walls for the combustion chamber and could cause the entire engine to be lighter. So, depending on how much more thrust the higher-pressure fuel of liquid oxygen and hydrogen generated, it could still be beneficial to use the gaseous oxygen and gasoline combination just because it has lower pressure.

Temperature can also affect the decision on which fuel type to use as a mixture with very high temperatures runs a risk of melting the walls of the combustion chamber. To avoid the risk of melting, potentially more expensive or heavier materials that have higher melting points would need to be used. This is where the liquid oxygen and hydrogen combination is better because it runs almost 700 degrees Celsius cooler than the gaseous oxygen and gasoline.

Conclusion

From the chart and the images taken from the simulation, the fuels that generate higher pressures are the most ideal for generating the most amount of thrust as all the fuels that had the higher pressure value generated higher amounts of thrust than the fuels with the lower pressure value. This is because the simulation that used a higher pressure value of 3.450 MPa resulted in a higher mass flow rate of $930\frac{g}{s}$ while the lower pressure 2.068 MPa simulation resulted in a lower mass flow rate of $795\frac{g}{s}$.

Although there are limitations to this research stemming from the fact that an actual combustion was not used but only a pressure simulation was used. This could have caused the data to only have two distinct sets rather than having the 8 that was supposed to appear. The model that was used could have also had some issues with accuracy and not being completely shaped correctly. Along with the fact that the model used did not do thermal calculations with the walls of the rocket, the results would not be exact to what could happen in real life.

There was also a limitation in that I did not do a proper pressure model in that I never calculated the effects of higher pressure on the walls of the combustion chamber. If I had done that, the results may have turned out differently as I may have had a more nuanced decision on which fuel and oxidizer mixture was truly the most optimal rather than simply the fuel and oxidizer combination that resulted in the highest pressure.

From this research, I have learned how to write a technical paper on a scientific research topic. I have also learned how to do a CFD simulation along with learning how to model a system for that simulation.

Future Work

In the future, if a more accurate conclusion were to be taken from this topic, if an actual combustion model was used rather than simply a pressure simulation, more accurate data could be generated from each fuel and oxidizer combination. This is because each fuel and oxidizer mixture can combust differently based on their molecular structure and amount that they were mixed into the chamber. With an actual combustion simulation, a more varied data set could have been found, and each individual fuel and oxidizer combination could have yielded their own data which would have been more effectively compared.

There could also be different fuel and oxidizer combinations that I did not test within my own research that could have yielded different thrust and pressure values. This could lead to different conclusions in how pressure affects thrust as they could get to a point where increasing pressure beyond that could be unrealistic. This is because the benefits that it would get would be unnoticeable but the drawbacks that could arise from increased pressure like decreased part lifespan and more catastrophic failures could outweigh the benefits.

References

[1] Miika Sikala, “How to design, build and test small liquid-fuel rocket engines – Leroy J. Krzycki,” *Github.io*, 2025. https://spacha.github.io/How-to-Rocket/#propellant-choice-and-properties (accessed Feb. 11, 2025).

[2] NASA, “Rocket Propulsion,” *Nasa.gov*, 2021. https://www.grc.nasa.gov/www/k-12/airplane/rocket.html (accessed Apr. 17, 2025).

Examining the Relationship Between Return on Equity and Return on Assets as Indicators of Corporate Longevity

Xiong Yin

ABSTRACT

This study investigates the correlation between Return on Equity (ROE) and Return on Assets (ROA) and their efficacy in distinguishing between active and inactive publicly traded companies. Analyzing financial data extracted from SEC filings, we find that while extreme outliers exist, the majority of firms cluster within a stable range of -15% to 15% for ROE and -5% to 15% for ROA. Companies that ultimately became inactive exhibited greater volatility and dispersion in these metrics over time. A longitudinal case study of Tesla demonstrates a transition from negative values into this stable cluster, correlating with its improved market position. The results indicate that sustained stability in ROE and ROA, rather than extreme short-term performance, is a stronger indicator of long-term corporate viability.

Introduction

In today's competitive environment, analysis of a firm's financial performance is essential to understanding its operational efficiency and long term viability. Two key financial metrics that are widely used for this purpose are Return on Equity (ROE) and Return on Assets (ROA). ROE, which is a firm's net income divided by shareholder equity, indicates how efficiently a company is using their investor capital to generate profits. ROA, on the other hand, defines how efficiently a firm is employing its assets to generate earnings. These ratios are particularly helpful to utilize when comparing companies between industries and when analyzing trends in financial well being.

This project both looks at the correlation between ROE and ROA, also how each of these values drift over time for a wide range of companies. Based on publicly available financial reports, we plot the correlation between the two and categorize companies by whether they remain active or become inactive. The research also considers how extreme outliers in either ROE or ROA can skew trends, and how eliminating those outliers allows more stable patterns to emerge. By considering both average values and trends over time, we hope to gain a clearer sense of what these financial metrics imply about company performance and stability.

Methods

To compare companies' performance based on ROE and ROA, the data are directly from the United States Securities and Extraction Commission (SEC) website. The website offers structured financial reports filed by public companies via its EDGAR database. These reports, which contain numerous detailed financial "facts" like Net Income (NetIncomeLoss), Stockholders' Equity (StockholdersEquity), and Total Assets (Assets), were extracted via machine and converted into format amenable to analysis.

In order to fully analyze the data, I downloaded a sample of around 1,000 companies' JSON files and put them into a folder named "/succeedcompanies". I used python in Google Colab to go through every JSON file and pull out the financial values that mattered.

For each company, I computed:

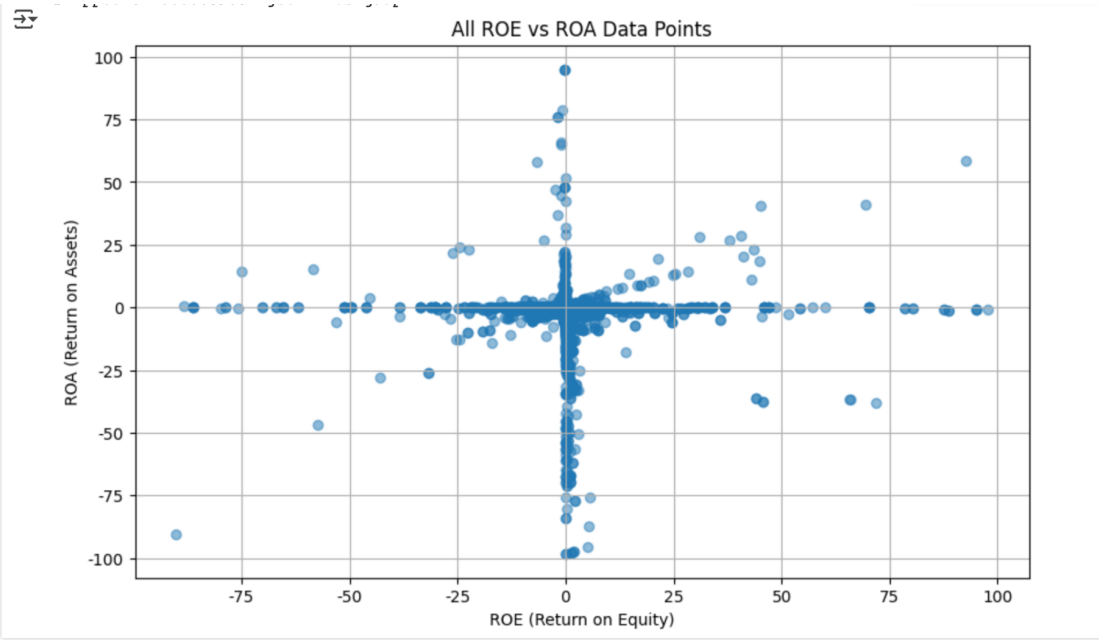
ROE (Return on Equity) = Net Income/Stockholders' Equity
ROA (Return on Assets) = Net Income/Total Assets

These values were computed for each reporting present in each file. The script aligned values based on their fiscal year end date, which enabled me to calculate both ROE and ROA for the same time point.

To illustrate the process, the following is a condensed version of the primary loop that pulled and processed the data:

```
for filename in os.listdir(folder_path):
    if filename.endswith(".json"):
        with open(os.path.join(folder_path, filename), 'r') as file:
            data = json.load(file)
            net_income_list = data['facts']['us-gaap'].get('NetIncomeLoss', {}).get('units', {}).get('USD', [])
            equity_list = data['facts']['us-gaap'].get('StockholdersEquity', {}).get('units', {}).get('USD', [])
            assets_list = data['facts']['us-gaap'].get('Assets', {}).get('units', {}).get('USD', [])
```

Having gathered the complete dataset, I plotted the results in a scatter plot of ROE vs ROA values for companies and years:



As illustrated, the majority of companies are closely grouped in the low values of ROA and in the broad interval of ROE. Yet, there are also some extreme outliers, from those data points with extremely high or extremely low values.

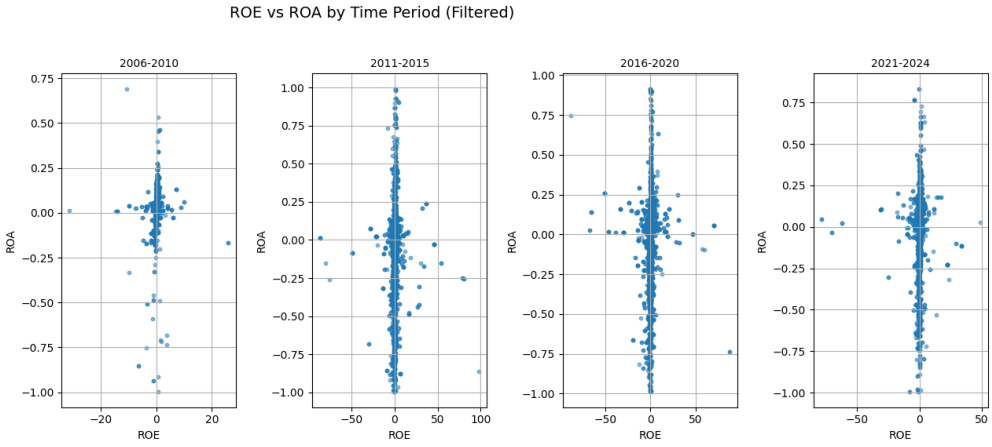
To concentrate on more realistic performance figures. I excluded those extreme values where either ROE or ROA was outside of the range, from -100% to 100%. This was an eyeball decision and prevents distortion due to accounting quirk. These outliers were not removed but put aside for individual examination later.

Then, I had to separate the companies that are still active and those that were probably inactive, for example bankrupt, acquired or no longer publicly traded. To accomplish this, I examined the latest fiscal filing date for each company: Latest ROE and ROA dates. By setting up a cutoff date, any companies that had not filed reports after December 31, 2023 will be marked as inactive.

These categories were appended to the data and enabled me to graphically compare trends in ROE and ROA between companies that are still operating and those that are no longer reporting. In subsequent phases, I also considered alternative methods of segmenting the data, including comparing changes over time in ROE and ROA, looking for clusters or patterns that could be used to forecast long term success or failure. In general, all of the choices made during the methods phase, like what data points to pull, how to qualify outliers, and how to classify company status, were all aimed to create a cleaner, more readable dataset that still captures the variability of real world financial performance.

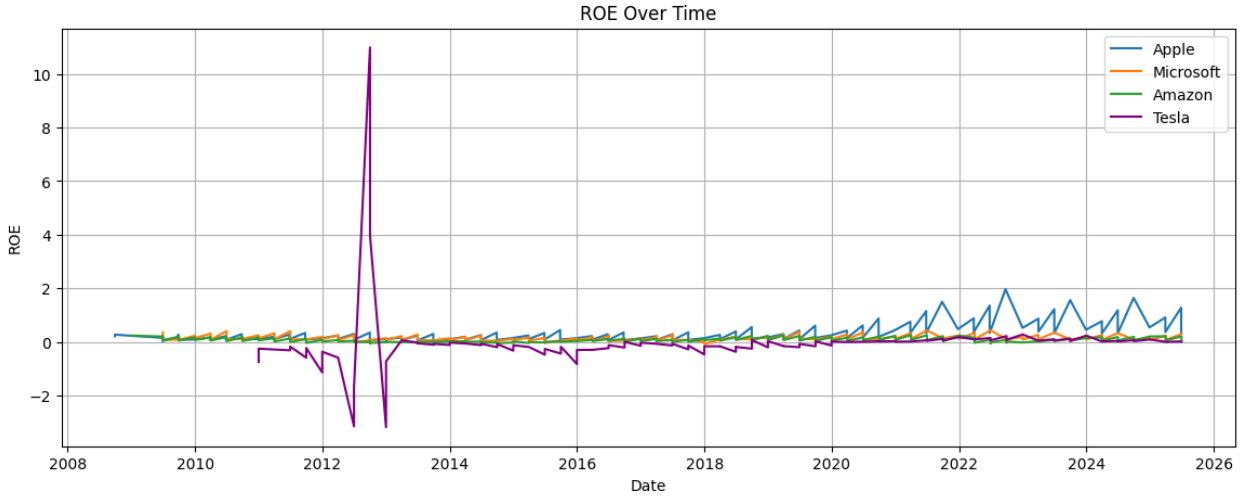
Results

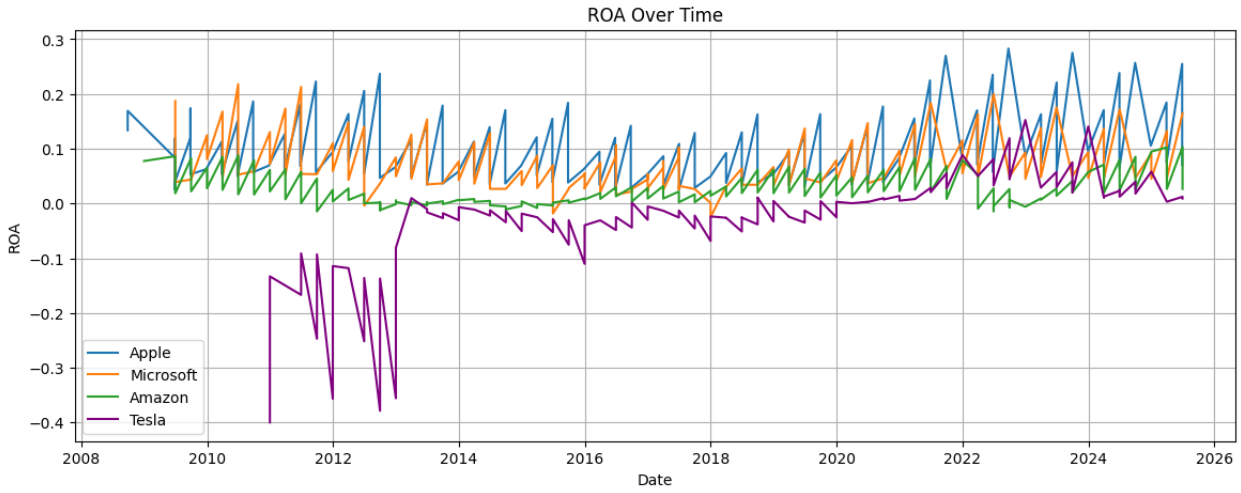
1. Different periods of companies' ROE vs ROA



   - These graphs are help to examine the relationship between Return on Equity (ROE) vs Return on Assets (ROA) across different time periods:
     i. 2006-2010 (Pre/Post 2008 Financial Period)
     ii. 2011-2015 (Recovery Period)
     iii. 2016-2020 (Economic Expansion)
     iv. 2021-2024 (Covid-19 and Aftermath)
   - X-axis: ROE (%)
   - Y-axis: ROA (%)
   - Based on the graphs, we can see few key findings:
     i. Core Cluster (approximately 78% of the firms):
        - Consistently centered near (0,0) with modest profitability
        - Represents stable, Low volatility businesses
     ii. Expanded Dispersion (2010-2015):
        - Wider during the end of the financial crisis and the recovery period
     iii. Temporal Trends:
        - Pre-2010: Higher volatility, with clusters stretching toward extreme values (ROE up to 50% in 2005–2009 due to leveraged buyouts)
        - Post-2010: Tighter cluster, which suggests stricter restrictions and improved risk management practices
     iv. Outliers:
        - Frequency: around 5-10 outliers per period
        - Causes: Accounting anomalies, Near zero equity

2. A closer look at successful vs failing companies





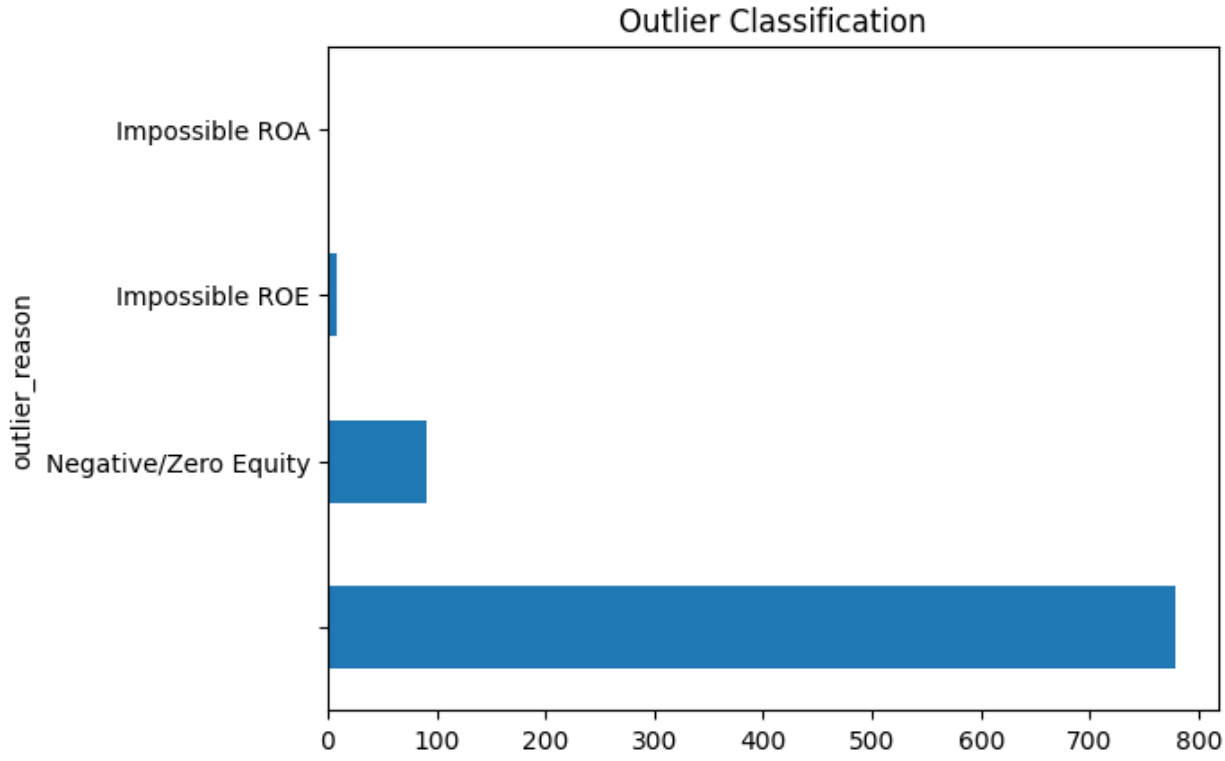
1. Introduce of Trends:

- **Apple:**
  Apple displayed a steady trend of stability and gentle enhancement in both ROE and ROA. ROE is nearly 0% in the early years of the dataset but steadily increases to 2% during 2022(post Covid-19 period). This suggests the effective utilization of shareholder capital, which is more likely to maintain its market dominance and ongoing product innovation. Apple's ROA commenced at a high 0.15%, which signified robust asset utilization from the onset, and reached a high of about 0.29% in 2023. The coupling of stable improvement in ROE with continuously high ROA illustrates Apple's long term operating efficiency and resilience, placing it squarely in the "stable cluster" of ongoing companies
- **Microsoft:**
  Microsoft has a credible consistent financial performance, with ROE stable between 0.2% and 0.5% since 2008, spiking briefly at the beginning of 2022. This low volatility profile is indicative of disciplined capital management and persistent market dominance. Its ROA started at 0.19%, fell slightly to about 0.06% between 2014-2020, then recovered to 0.2% during the Covid and Post recovery period. The recovery indicates a return boost in asset efficiency, maybe because of the heightened demand for cloud and software services globally during lockdowns. Microsoft's stability

in both metrics over so long a horizon puts it firmly among the most financially successful companies in the dataset.

- **Amazon:**
  Amazon has a more volatile profile, its ROE started near 0%, rose between 0.1% and 0.3% in the early years, then jumped around 0.5% in 2022 before stabilizing around 0.4%. These variations are associated with phrases of stable growth and scaling. Amazon's ROA follows the same narrative: initially at a healthy 0.08%, it fell sharply to near 0-0.02% from 2012 to 2016 as it reinvested heavily in logistics, cloud infrastructure, and market growth. In recent years, ROA rebounded to around 0.10%, indicating better asset utilization as the cost of expansion normalized. This path is indicative of Amazon's approach of trading off short-term efficiency for long-term market leadership, eventually returning to healthier fundamentals.
- **Tesla:**
  Tesla offers the best example for a turnaround from a troubled company to a more stable one. During its early years(2010-2012), Tesla's ROE and ROA is strongly negative, its ROE is around -1% to -3%, indicating intense operational and financial distress. ROA has had even more negative trouble, beginning with -0.4% in 2013 following the Model S launch, as the company wrestled with high production expenses and liquidity issues. Although there were temporary surges in performance, instability continued until 2010. From that time, however, Tesla's metric starts to consistently improve. Its ROE starts to move into positive territory, and ROA reached about 0.15% in the Post-Covid era, positioning Tesla in the stable range of active companies such as Apple and Microsoft. This shift shows how strengthening fundamentals can push a company from the "at-risk" profile toward longer-term viability.

3. Here are the main trend

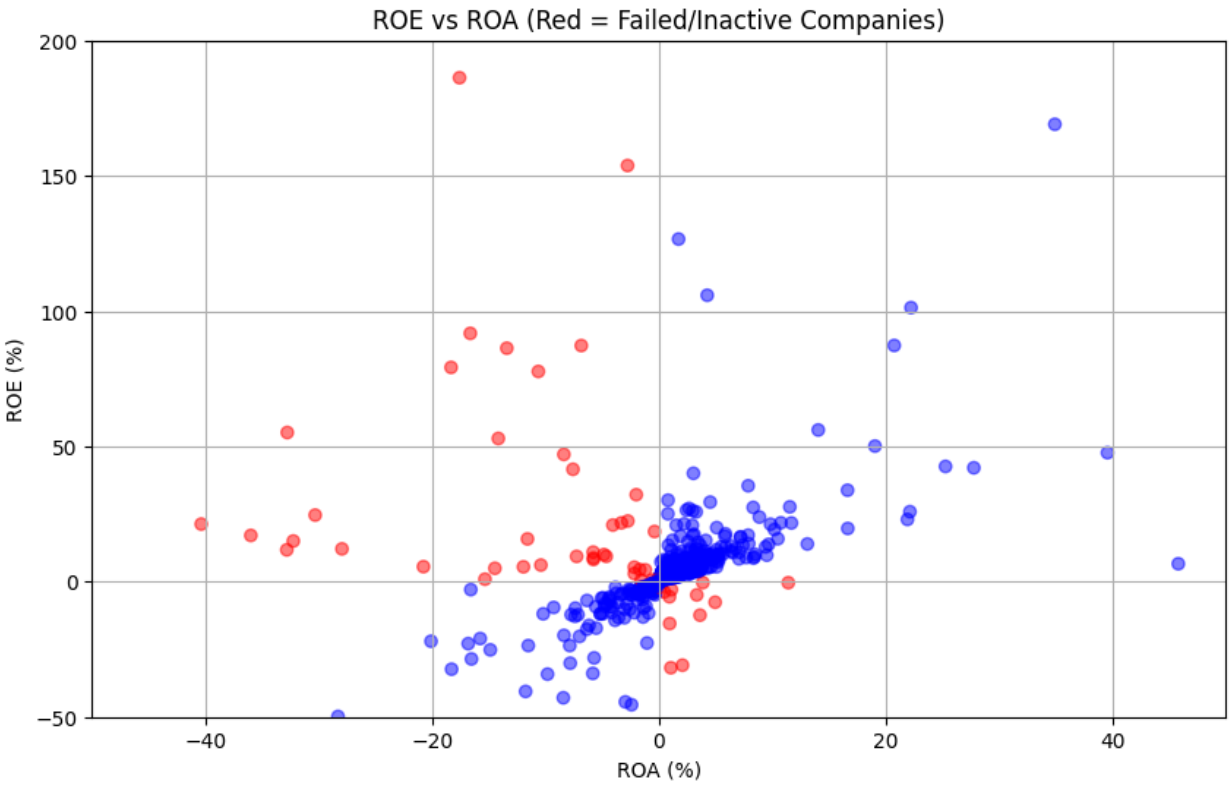


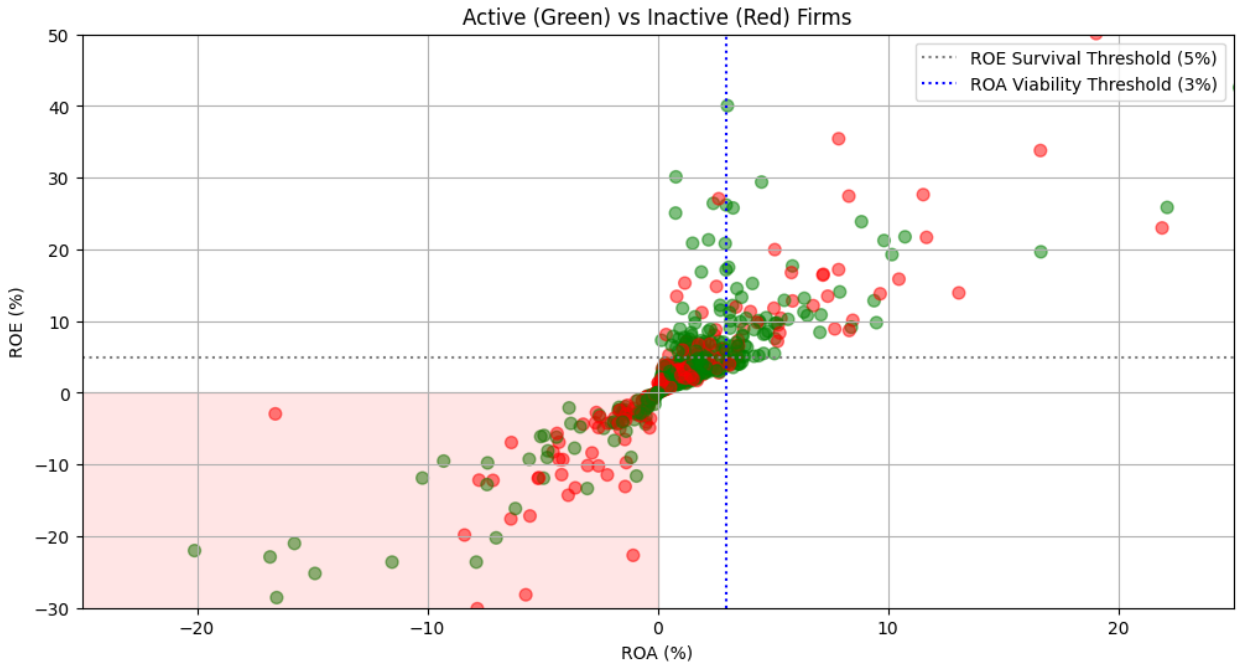
- Outliers disproportionately skewed the initial distribution. Their removal revealed the true central tendency, with most firms clustered near ROE 5–15% and ROA 1–5%
- The cleaned dataset excludes outliers, ensuring robust analysis. Remaining values align with industry norms: Cleaned Data

- As preserving outlier records for transparency. Common issues included data errors and leverage distortions. Here is the file for all companies: Organized outliers
- Here are some extreme outliers cases:
  1. **SOOUM CORP (2013 Q3):**
     - ROE: 4.3% (normal range)
     - ROA: -25,162.6% (extreme outlier)
     - This is obviously mathematical impossibly for a company that is operating, due to the volatility in the commodities market, making it hard to use its assets

- Another example:
  1. Hill-Rom Holdings INC. (2019 Q3)
     - ROE: 31,061,224% (Absurdly outlier)
     - ROA: 3.1% (normal)
     - This company has like near zero equity, the likely issues would be having too much net income while the equity was a very small part

Visualization Active vs Inactive companies:





- Active (green) and inactive (red) firms show substantial overlap in the 0-5% ROA/ROE range. Dashed lines indicate theoretical survival thresholds that failed to separate groups
- As we can see in graph 1 that companies that are active have similar ROE and ROA when encountered, making it more stable across a long period of time. While the inactive companies are more likely to be spread out on the graphs that are often unstable on the correlation between ROE and ROA

Conclusion

By examining ROE and ROA, together they offer a clearer view of a company's health and development. Most successful companies are clustered in a more moderate and stable range: -15%-15% for ROE and -5-15% for ROA. While inactive companies show more volatility and prolonged periods of poor performance. Tesla's journey from deep negative values and instability to a profile resembling the most stable firms demonstrates that improving these metrics can be a leading indicator of long-term success. Conversely, companies such as SOOUM CORP and Hill-Rom Holdings INC. are showing that extreme imbalances, even when only one metric is abnormal, often precede failure.

References

[1] U.S. Securities and Exchange Commission. (n.d). EDGAR — Search and Accesshttps://www.sec.gov/edgar

A Unified Design and Evaluation Framework for Liquid Rocket Engines Incorporating Environmental and Economic Tradeoffs

Cayden Shi

Abstract

As current rates of global warming accelerate, it becomes increasingly important to evaluate environmental impacts from all sources, including those historically considered to be marginal contributors. Rocket propulsion, while currently a small contributor to global carbon emissions, emit a disproportionately high amount relative to the frequency and scale of launches. Amidst increasing interests and investment in space, it also becomes increasingly important to evaluate the environmental impacts of popular rocket fuels. While current rocket design practices focus on solely performance, cost, and dimensions, environmental trade-offs remain unexplored due to lack of accessible tools that integrate all axes of consideration. This research presents a modular dual-function Python-based software toolkit, based on classic rocket equations and novel environmental and economic scoring algorithms, that allows users to both design a liquid bipropellant rocket engine, while also analyze and visualize the performance–environment–cost tradeoffs of various rocket propellants through a mission-level trade study, focusing on CO_2 Emission and cost per unit thrust metrics. The study emphasizes allowing engineers to make informed, conscious rocket design choices ethical with current climate protection considerations in mind.

Introduction

Liquid bipropellant rocket engines remain foundational for current spaceflight. These engines rely on the combustion of fuel and oxidizer in a high-pressure chamber to generate hot gases that expand through a nozzle to produce thrust.

Traditionally, rocket engine optimization has focused on metrics like specific impulse (Isp), which measures the efficiency of a propellant in generating thrust per unit mass, allowing for lower launch masses and higher payload capacities, as well as cost per pound of fuel and oxidizer. However, as climate concerns grow and the space industry expands, the environmental

cost of propellants, like greenhouse gas emissions, have become an increasingly relevant design consideration.

Despite these shifts, design tools rarely provide comparisons between the environmental impacts of various propellants. Many commonly used fuels like RP-1, UDMH, and ethanol emit significant carbon dioxide or toxic byproducts, while others like liquid hydrogen or methane offer cleaner combustion profiles. This begs the need for accessible, integrated tools to evaluate these tradeoffs, so engineers fully understand the consequences that come with the type of fuel they choose to use.

This research addresses that gap by presenting a modular Python-based software toolkit that supports both the design of a custom liquid rocket engine and a quantitative trade study of propellant candidates. The toolkit guides users through 9 modules for calculating mass flow, thermodynamics, nozzle sizing, chamber dimensions, structural thickness, cooling needs, injector sizing, and cost estimation. In addition to these standard engineering calculations, the toolkit includes a novel analysis framework module that evaluates propellants based on CO_2 emissions per unit thrust (lb/N·s), cost per unit thrust ($/N·s), and toxicity weighting.

To compare environmental impacts between fuels, the tool implements a multi-criteria scoring algorithm, using normalized I_{sp}, CO_2 emissions, and cost, allowing for a weighted environmental score (Ecoscore) that weighs the tradeoffs between environmental impact and fuel performance, and an overall score (TotalScore), that adds in cost evaluations to Ecoscore. The results are then presented through tables to help visually explore fuel tradeoffs.

This accessible data-driven framework that allows for performance environment cost tradeoffs allows for more responsible mission planning propulsion choices. In an era of exponentially increasing climate risks, a toolkit that ethically keeps the planet in mind is crucial.

Method

This section outlines the computational framework and equations used in each module of the rocket design and evaluation toolkit. Some modules have variable dependences calculated in previous modules.

1. Propellant and Performance Module

This module determines the total propellant mass flow rate needed for a specific thrust and impulse, then divides it into fuel and oxidizer flows based on the mixture ratio. This is done through the equations:

$$W_t = \frac{F}{I_{sp}} \quad \text{Eq. 1}$$

$$W_f = \frac{W_t}{1+R} \qquad W_o = W_t - W_f = \frac{W_t \cdot R}{1+R} \quad \text{Eq. 2}$$

Where:

- Wt: Total propellant mass flow [lb/s]
- F: Thrust [lbf]
- Isp: Specific impulse [s]
- Wf: Fuel mass flow [lb/s]
- Wo: Oxidizer mass flow [lb/s]
- R: Mixture ratio (oxidizer/fuel)

2. Thermodynamics Module

The Thermodynamics Module calculates key flow properties at various stages of the engine, specifically, the temperatures and pressures at the chamber, throat, and nozzle exit. This is done through the equations:

$$T_c = T_{c,^\circ F} + 460 \qquad T_t = \frac{T_c}{1+\frac{\gamma-1}{2}} \quad \text{Eqs. 4-5}$$

$$P_t = P_c \cdot \left(1 + \frac{\gamma-1}{2}\right)^{-\frac{\gamma}{\gamma-1}} \quad \text{Eq. 6}$$

$$M_e = \sqrt{\frac{2}{\gamma-1}\left[\left(\frac{P_c}{P_{atm}}\right)^{\frac{\gamma-1}{\gamma}} - 1\right]} \quad \text{Eq. 7}$$

$$\frac{T_e}{T_c} = \frac{1}{1+\frac{\gamma-1}{2}M_e^2} \qquad T_e = T_c \cdot \frac{1}{1+\frac{\gamma-1}{2}M_e^2} \quad \text{Eqs. 8-9}$$

Where:

- Tc: Chamber temperature [°R]
- Tc,°F: Chamber temperature input [°F]
- Tt: Throat temperature [°R] (isentropic flow)
- Pc: Chamber pressure [psi]
- Pt: Throat pressure [psi] (isentropic relation)
- Patm: Ambient pressure [psi]
- Me: Exit Mach number [dimensionless] (from pressure ratio)
- Te: Exit temperature [°R]

- γ: Ratio of specific heats [dimensionless]
- R: Gas constant [ft·lbf/(lb·°R)]

3. Nozzle Design Module

This module determines the geometry of the converging-diverging nozzle, including the throat and exit areas and their corresponding diameters. These calculations ensure that the nozzle is properly sized to handle the mass flow and chamber conditions required. This is accomplished through the equations:

$$A_t = \frac{W_t}{P_t \cdot 144} \cdot \sqrt{\frac{R \cdot T_t}{\gamma \cdot g_c}} \qquad D_t = 2 \cdot \sqrt{\frac{A_t}{\pi}} \cdot 12 \qquad \text{Eqs. 10-11}$$

$$\frac{A_e}{A_t} = \frac{1}{M_e}\left(\frac{1+\frac{\gamma-1}{2}M_e^2}{\frac{\gamma+1}{2}}\right)^{\frac{\gamma+1}{2(\gamma-1)}} \qquad A_e = A_t \cdot \left(\frac{A_e}{A_t}\right) \qquad \text{Eqs. 11-12}$$

$$D_e = 2 \cdot \sqrt{\frac{A_e}{\pi}} \cdot 12 \qquad \text{Eq. 13}$$

Where:

At: Throat area [ft²]

Wt: Total mass flow rate [lb/s]

Pt: Throat pressure [psi]

Tt: Throat temperature [°R]

R: Gas constant [ft·lbf/(lb·°R)]

γ: Ratio of specific heats [dimensionless]

gc: Gravitational constant [32.174 ft/s²]

Dt: Throat diameter [in]

Ae: Exit area [ft²]

De: Exit diameter [in]

Me: Exit Mach number [dimensionless]

Ae/At: Nozzle expansion ratio [dimensionless]do

4. Combustion Chamber Sizing Module

This module estimates the internal geometry of the chamber based on the required propellant volume and flow characteristics. This is done through the equations:

$$V_c = L^* \cdot A_t \cdot 144 \quad \text{Eq. 14}$$

$$A_c = \text{contraction_ratio} \cdot A_t \cdot 144 \quad \text{Eq. 15}$$

$$D_c = 2 \cdot \sqrt{\frac{A_c}{\pi}} \quad \text{Eq. 16}$$

$$L_c = \frac{V_c}{A_c} \quad \text{Eq. 17}$$

Where:

Vc: Chamber volume [in³]

L*: Characteristic length [in], an empirical value based on engine type

At: Nozzle throat area [ft²]

Ac: Chamber cross-sectional area [in²]

contraction_ratio: Ratio of Ac to At, typically between 3–5 [dimensionless]

Dc: Chamber diameter [in]

Lc: Chamber length [in]

5. Structural Sizing Module

This module calculates the required wall thickness and outer diameter of the combustion chamber to safely contain the internal pressure during operation. This is accomplished through the following equations:

$$t_w = \frac{P_c \cdot D_c}{2 \cdot \sigma_{allow}} \quad \text{Eq. 18}$$

$$D_{outer} = D_c + 2 \cdot t_w \quad \text{Eq. 19}$$

Where:

tw: Required wall thickness [in]

Pc: Chamber pressure [psi]

Dc: Inner chamber diameter [in]

σallow: Allowable material stress [psi]

Douter: Outer chamber diameter including wall thickness [in]

6. Cooling and Heat Transfer Module

This module estimates the cooling requirements for the combustion chamber by calculating the heat load, coolant flow rate, and the outer jacket diameter needed to accommodate the cooling flow. It assumes jacket-style cooling using water or another coolant. This is accomplished through the following equations:

$$A_{surface} = \pi \cdot D_c \cdot L_c \qquad Q = q'' \cdot A_{surface} \qquad \text{Eqs. 20-21}$$

$$\dot{W_w} = \frac{Q}{\Delta T} \qquad A_{annular} = \frac{\dot{W_w}}{\rho \cdot v} \qquad \text{Eqs. 22-23}$$

$$D_2 = \sqrt{\frac{4 \cdot A_{annular}}{\pi} + \left(\frac{D_1}{12}\right)^2} \cdot 12 \qquad \text{Eq. 24}$$

Where:

Asurface: Internal surface area of the chamber wall [in²]

Dc: Inner chamber diameter [in]

Lc: Chamber length [in]

q": Heat flux [Btu/in²/s]

Q: Total heat load [Btu/s]

Ww: Coolant mass flow rate [lb/s]

ΔT: Allowable temperature rise of coolant [°F]

ρ: Coolant density [lb/ft³]

v: Coolant velocity [ft/s]

Aannular: Required annular flow area [ft²]

D1: Outer diameter of the inner wall [in]

D2: Outer diameter of the cooling jacket [in]

7. Injector Sizing Module

This module calculates the flow area and corresponding orifice diameter required for delivering fuel or oxidizer through the injector plate. This is accomplished through the equations:

$$A = \frac{\dot{W}}{C_d \cdot \sqrt{2 \cdot g_c \cdot \rho \cdot \Delta P}} \quad \text{Eq. 25}$$

$$D = 2 \cdot \sqrt{\frac{A}{\pi}} \cdot 12 \quad \text{Eq. 26}$$

Where:

- A : Injector flow area [ft^2]
- D : Injector orifice diameter [in]
- Ẇ: Mass flow rate (fuel or oxidizer) [lb/s]
- Cd: Discharge coefficient [dimensionless]
- gc: Gravitational constant [32.174 ft/s^2]
- rho : Fluid density [lb/ft^3]
- ΔP: Pressure drop across the injector [psi]
- π: Pi, mathematical constant ≈ 3.1416

8. Efficiency and Cost Module

This module estimates the total propellant cost and cost per unit of delivered thrust over the engine burn duration. It calculates how much fuel and oxidizer is consumed during operation, multiplies these masses by their respective costs, and computes a normalized cost per lbf-second of thrust. This is accomplished through the following equations:

$$m_f = \dot{W}_f \cdot t \qquad m_o = \dot{W}_o \cdot t \quad \text{Eqs. 27-28}$$

$$\text{Cost} = m_f \cdot c_f + m_o \cdot c_o \qquad \text{Cost/lbf-s} = \frac{\text{Cost}}{F \cdot t} \quad \text{Eqs. 29-30}$$

Where:

- mf: Total fuel mass used [lb]
- mo: Total oxidizer mass used [lb]
- Wfdot: Fuel mass flow rate [lb/s]
- Wodot: Oxidizer mass flow rate [lb/s]
- t: Burn time [s]
- cf: Fuel cost per pound [$ / lb]
- co: Oxidizer cost per pound [$ / lb]
- Cost: Total propellant cost [$]
- F: Thrust [lbf]
- Cost/lbf-s : Cost per pound-force-second [$ / (lbf·s)]

9. Unified Mission Environmental & Cost Trade Study Module

This module compares the value of different fuels through a variety of metrics including performance, climate damage, and cost. To establish a shared baseline, this module assumes all fuels are being used to accomplish the same mission, and simulates the cost, CO2 emitted, and performance required to complete that mission. The user chooses how many propellants they want to compare and inputs preliminary data for each propellant configuration. Results appear in the form of a table with scores for direct and streamlined comparison. The calculations for this module are divided into several parts. The first three sections (propellant mass calculation, engine-level flow parameters, and fuel/oxidizer breakdown) reuse equations already defined in earlier modules (Modules 1, 2, 3) to determine thrust, specific impulse relationships, mass flow rate, mixture ratios, oxidizer/fuel flow breakdowns, exhaust velocity, structural and initial mass, and total propellant mass in both kilograms and pounds. Building on the reused equations, we introduce three new sections.

The first new section involves a normalized metric for cross-comparing efficiency and sustainability between propulsion systems. The three equations below are used to quantify the environmental impact of the fuel by calculating the total amount of carbon dioxide emitted during the burn, the average rate of CO_2 emission per second, and the normalized CO_2 emission per unit of delivered impulse (in Newton-seconds).

$$\mathrm{CO}_2 = m_f \cdot \mathrm{CO}_{2,per_lb} \quad \text{Eq. 31}$$

$$\mathrm{CO}_2/s = \frac{\mathrm{CO}_2}{t_{burn}} \quad \text{Eq. 32}$$

$$\mathrm{CO}_2/(N \cdot s) = \frac{\mathrm{CO}_2}{F \cdot 4.44822 \cdot t_{burn}} \quad \text{Eq. 33}$$

Where:

CO2 = Total carbon dioxide emitted [kg]

mf = Fuel mass [lb]

CO2_per_lb = CO_2 emission per pound of fuel [kg/lb]

t_burn = Total burn time [s]

F = Thrust [lbf]

CO2/s = CO_2 emission per second [kg/s]

CO2/(N·s) = CO_2 emission per Newton-second of impulse [kg/(N·s)]

4.44822 = Conversion factor from lbf to N [N/lbf]

The second new section involves introducing cost evaluation into the analysis, enabling economic comparison between different propulsion configurations. The total cost is calculated based on the cost of fuel and oxidizer, and this total is then normalized per unit of delivered impulse (N·s).

$$\mathrm{Cost} = m_f \cdot c_f + m_o \cdot c_o \quad \text{Eq. 34}$$

$$\mathrm{Cost}/(N \cdot s) = \frac{\mathrm{Cost}}{F \cdot 4.44822 \cdot t_{burn}} \quad \text{Eq. 35}$$

$$\frac{\mathrm{CO}_2}{N \cdot s} = \frac{\mathrm{CO}_2}{F \cdot 4.44822 \cdot t_{\mathrm{burn}}} \quad \text{Eq. 36}$$

Where:

Cost = Total cost of propellant [$]

mf = Fuel mass [lb]

cf = Cost per pound of fuel [$ / lb]

mo = Oxidizer mass [lb]

co = Cost per pound of oxidizer [$ / lb]

t_burn = Total burn time [s]

F = Thrust [lbf]

4.44822 = Conversion factor from lbf to N [N/lbf]

Cost/(N·s) = Cost per unit impulse [$ / (N·s)]

CO2/(N·s) = CO_2 emitted per unit impulse [kg / (N·s)]

This final section introduces the core novelty of the module: a custom scoring framework that translates raw propulsion, environmental, and economic data a universal score for direct comparison across propellant configurations. First, three preliminary scores are calculated: PerfScore (Performance Score), EnvScore (Environmental Score), and CostScore (Cost Score). These three values normalize each propellant type into a score between 0-1 by dividing each propellant type by the maximum value in the comparison group. EnvScore and CostScore are reversed by having that value subtracted from 1, rewarding lower scores and penalizing higher scores with higher costs and CO2 emissions.

These three normalized metrics are then combined in two composite scores. EcoScore weighs performance and emissions equally (40% each), subtracts a fixed penalty based on fuel toxicity (20%), and scales the result to a 100-point system, directly incorporating both quantitative and qualitative sustainability factors. In parallel, TotalScore provides a broader benchmark by assigning 40% each to performance and environmental efficiency, and 20% to cost-effectiveness. This dual-score system allows designers to target either sustainability (EcoScore) or holistic value (TotalScore), depending on mission priorities.

$$\text{PerfScore} = \frac{I_{sp}}{I_{sp,\max}} \quad \text{Eq. 37}$$

$$\text{EnvScore} = 1 - \frac{\text{CO}_2/(N \cdot s)}{\left(\text{CO}_2/(N \cdot s)\right)_{\max}} \quad \text{Eq. 38}$$

$$\text{CostScore} = 1 - \frac{\text{Cost}/(N \cdot s)}{\left(\text{Cost}/(N \cdot s)\right)_{\max}} \quad \text{Eq. 39}$$

$$\text{EcoScore} = 100 \cdot (0.4 \cdot \text{PerfScore} + 0.4 \cdot \text{EnvScore} - 0.2 \cdot \text{Toxicity}) \quad \text{Eq. 40}$$

$$\text{TotalScore} = 100 \cdot (0.4 \cdot \text{PerfScore} + 0.4 \cdot \text{EnvScore} + 0.2 \cdot \text{CostScore}) \quad \text{Eq. 41}$$

Where:

PerfScore = Normalized specific impulse [unitless]

Isp = Specific impulse of engine [s]

Isp_max = Maximum Isp among all configurations [s]

EnvScore = Environmental performance score [unitless]

CO2/(N·s) = CO_2 emitted per unit impulse [kg / (N·s)]

(CO2/(N·s))_max = Maximum CO_2 per unit impulse among all configs [kg / (N·s)]

CostScore = Cost efficiency score [unitless]

Cost/(N·s) = Cost per unit impulse [$ / (N·s)]

(Cost/(N·s))_max = Maximum cost per unit impulse among all configs [$ / (N·s)]

Toxicity = Binary value (1 = toxic propellant, 0 = non-toxic) [unitless]

EcoScore = Environmental sustainability score [0–100]

TotalScore = Overall propulsion configuration score [0–100]

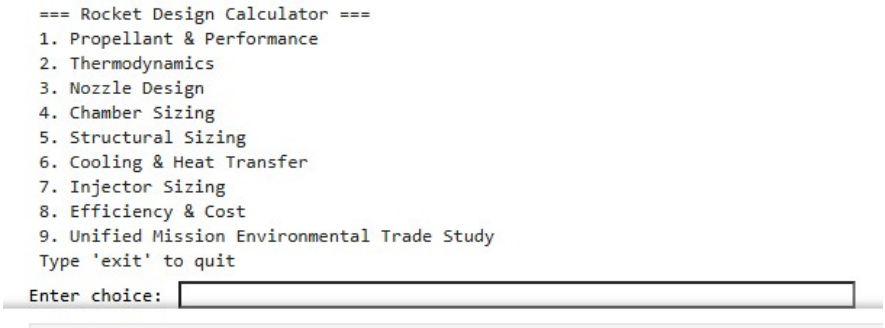


Figure 1 Program Design Interface

The final user interface, as seen in Figure 1 of the Rocket Design Calculator presents a clear, menu-driven layout where users can select from nine core modules.

Result & Discussion

To explore the trade-offs between performance, environmental impact, and cost in liquid rocket propellant selection, a comparative mission-level trade study was performed using module 9 of the toolkit. Six commonly used propellant combinations, LH_2/LOX, Methane/LOX, RP-1/LOX, Ethanol/LOX, UDMH/N_2O_4, and Hydrazine/N_2O_4, were evaluated across key metrics: specific impulse (I_{sp}), CO_2 emissions per unit thrust (CO_2/N·s), and cost per unit thrust ($/N·s). The values entered for each fuel type was found through various sources including the NASA Glenn Research Center and various other articles on propellants. Using EcoScore, TotalScore, and other calculated metrics, the propellants were ranked.

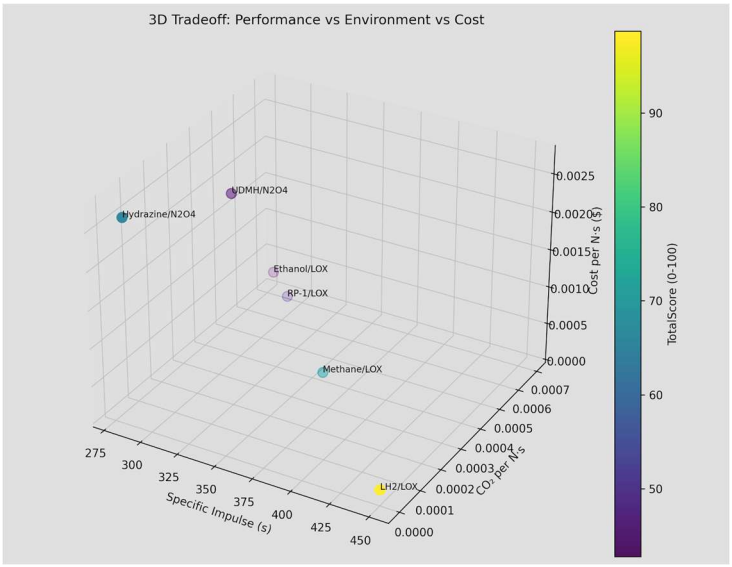


Figure 2 Performance vs Environment vs Cost Tradeoffs

Figure 2 shows a 3D scatter plot comparing propellants by performance (Isp), environmental impact (CO_2 per thrust), and cost per thrust. LH_2/LOX stands out as the clear leader, with the highest Isp (450 s), zero CO_2 emissions, and low cost—earning the top TotalScore of 98.72. All other propellants fall short in one or more areas: Methane/LOX performs moderately well with lower emissions and cost, while RP-1/LOX, Ethanol/LOX, UDMH/N_2O_4, and Hydrazine/N_2O_4 are penalized by high emissions, cost, or toxicity.

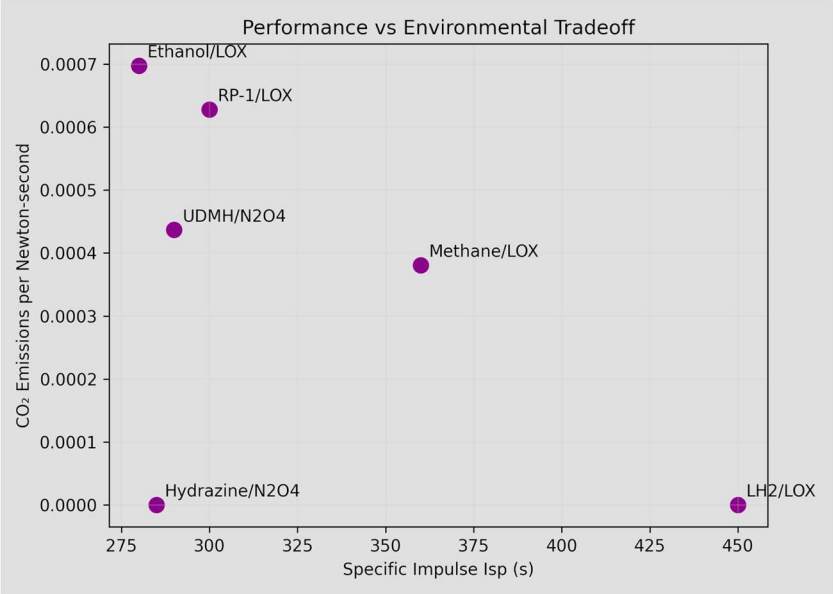


Figure 3 Performance vs Environmental Tradeoffs

Figure 2 simplifies the comparison by plotting specific impulse against CO_2 emissions per Newton-second, revealing a general trend: higher efficiency tends to correlate with lower emissions. LH_2/LOX again emerges as the top performer, offering unmatched efficiency and zero carbon output. Methane/LOX stands out among carbon-based options for its strong balance of performance, emissions, and future viability through in-situ resource utilization. In contrast, RP-1/LOX and UDMH/N_2O_4 deliver high thrust but at the cost of increased emissions and toxicity, making them better suited for specialized or expendable missions. Ethanol/LOX performs the worst across all metrics.

Conclusions & Future Work

This research presents a unified, modular Python-based toolkit for liquid bipropellant rocket engine design and environmental trade analysis. Unlike traditional design approaches that prioritize thrust and cost alone, this toolkit incorporates performance, environmental impact, and cost-efficiency into an integrated decision-making framework.

Through eight modular engineering calculators and an expanded mission-level analysis module, the toolkit enables users to not only size and analyze custom engines, but also quantitatively compare multiple propellants using consistent metrics such as CO_2 emissions per Newton-second and cost per unit thrust. A custom scoring algorithm that balances normalized performance, environmental, and cost scores provides a clear, interpretable ranking of candidate fuels across different mission priorities.

The results of a study conducted through the toolkit demonstrate that LH_2/LOX offers the most favorable overall profile, combining the highest specific impulse with zero CO_2 emissions and low cost. Methane/LOX emerged as a strong alternative for reusable or long-duration missions, offering lower emissions and cost than RP-1/LOX. In contrast, fuels like UDMH/N_2O_4 and Ethanol/LOX scored poorly across most axes, reflecting their limited viability in modern, sustainability-conscious aerospace applications.

Ultimately, this work underscores the importance of incorporating environmental and economic considerations into early-stage rocket design as climate change accelerates, as well as providing an easy to use and efficient rocket design calculator to help novel rocket builders make planet-conscious, price-conscious, and performance-maximizing decisions while building rockets.

Future improvements could include more accurate modeling of gas properties and nozzle flow using real-gas or chemical equilibrium data. The current toxicity input is a simplified, qualitative score assigned based on human judgment and available literature, which may introduce bias or inconsistency.

References

[1] G. Sutton and O. Biblarz, *Rocket Propulsion Elements*, 9th ed. Hoboken, NJ, USA: Wiley, 2016.

[2] NASA Glenn Research Center, “Chemical Propellants,” *NASA CEA Documentation*, 2023. [Online]. Available: https://cearun.grc.nasa.gov/

[3] E. W. Price, “Historical Review of Solid Propellant Rocket Motors,” *NASA Technical Memorandum*, TM-1982-84978, 1982.

[4] J. D. Hager, “Performance and Cost Comparison of Liquid Propellants,” *AIAA*, 1989.

[5] S. Pacha, “How to Rocket,” *spacha.github.io*, [Online]. Available: https://spacha.github.io/How-to-Rocket/#propellant-choice-and-properties, 2023.

Quantifying Ballot Redistribution Turbulence in Ranked-Choice Voting through Classical and Tropical Wasserstein Distances

Kevin Liu[1] and Austin Wei[2]

[1]*Mission San Jose High School*
[2]*Franklin Academy*

ABSTRACT

Ranked-choice voting (RCV) reassigns ballots across elimination rounds, making outcomes sensitive to how preferences change as candidates are removed. In this paper, we use a transportation distance to quantify how the ballot distribution varied over different rounds. The voter choices are encoded as fix-length rank vectors and we compare distributions before and after removing a single candidate using the classical 2-Wasserstein distance with squared Euclidean ground cost and a tropical p-Wasserstein variant whose cost is the range of coordinate differences. We develop two simulations based on an actual data from the San Francisco 2024 Mayoral Election as well as self generated sets to model skewed and more uniform distributions; the first simulation utilizes a sample and sort model to create rankings while the second uses an annealed model with an exhaustion temperature that increases with the ranks to model ballot truncation. Our simulation uses 150 people, which generates a 150x13 matrix of vectors that represent voter preferences from which we remove a candidate to obtain a 150x12 matrix which is then compared to the original matrix to solve an optimal transport problem under uniform marginals despite being different dimensions. Traditional Euclidean Wasserstein distances between rounds reflect the general trend of shifted preferences, but by employing tropical distances, we are able to track the largest per-ballot change and the fraction of ballots that contain the removed candidate. Furthermore, to create a more accurate comparison of Wasserstein distances across rounds, we weight coordinates in both the tropicalized and Euclidean matrices to emphasize the importance of earlier ranks. Extending the novel field of tropical algebra, we are able to measure new different things beyond what traditional calculations could show, and can be used for real precinct data or alternative ballot encodings.

1. INTRODUCTION

Ranked-choice voting has been adopted by a steadily increasing number of U.S. jurisdictions with the aim of producing majority winners, moderating campaign incentives, and potentially increasing voter participation. Ballot exhaustion, or the ballots in later rounds that no longer be counted because all chosen candidates have eliminated, has been documented to be raising concerns about voter disenfranchisement as well as the overall representativeness. On the other hand, some find limited systematic effects of turnout or socioeconomic disparities in voter participation, but theoretical work reveals the possibility of monotonicity failure in RCV. These observations suggest that understanding the voter distribution round to round matters for interpreting RCV outcomes.

Tropical geometry and algebra is a relatively new field that currently has promising applications inside of statistics and machine learning. Foundational papers have appeared in the past two decades, and applications of Tropical Optimization Theory on the tropical projective torus are even more recent with computational algorithms for tropical $\mathbf{W}_1$ and $\mathbf{W}_2$ appearing only in the past few years. Talbut et al. (2024) extends this theory to comparisons across different dimensions, proving the equivalence of projecting and embedding distances and showing that the tropical matrices are non-expansive with respect to $\mathbf{d}_{tr}$. While their study demonstrated the its application in comparing phylogenetic trees, the comparisons across different dimensions served as an impetus for comparing voter distributions in RCV across different rounds.

San Francisco's November 5, 2024 mayoral elections served as a concrete and reliable place to begin and start simulating. While this distribution matched a lot of other ones across the world, we decided to use simulated distributions to demonstrate more skewed and uniform distributions. Analyzing how the distribution changed across rounds when each candidate is hypothetically removed one at a time allowed us to quantify the sensitivity of the outcomes as well as the patterns.

In this paper, our goal is to compare the empirical distribution of simulated ballots of dimension **d** with the following round where a single candidate is removed with dimension **d** − 1. Traditional optimal transport problems and the resulting Wasserstein distances are geometric and are widely used to compare probability measures in statistics, but are traditionally confined to comparisons within the same space. To compare $\mathbf{R}^d$ with $\mathbf{R}^{d-1}$, we build upon the framework of Talbut et al who prove that the two distances from projecting the higher dimensional measure down or embedding the lower dimensional measure up are actually equal. In regards to the actual calculation, we build off of the framework of Cai and Lim who provide a calculation utilizing Lasserre sum-of-squares as a means to calculate the Wasserstein distance between our discrete probability measures.

Following the simulation and generation of a $150 \times d$ and a $150 \times (d-1)$ matrix by removing one candidate at a time, we calculate the Euclidean and tropical p-Wasserstein distances between the measures. While this information did give us insight on the change in distribution as well as extreme movements between rounds, we also used a weighted Euclidean and Tropical variant of this calculation to reflect the importance of the early ranks and the less-important nature of later preferences which likely provided too much noise for their respective importance during our original calculations.

2. BACKGROUND

In this section we introduce the tropical setting of our project, recall the definition of Wasserstein distances, and outline the approach for comparing probability distributions supported on spaces of different dimensions in both Euclidean and tropical geometries. Throughout, we write $[n] := \{1, 2, \ldots, n\}$.

2.1. *Essentials of Tropical Algebra and Geometry*

We will begin by introducing tropical algebra and geometry. See Brugallé et al. (2014) for a comprehensive introduction.

2.1.1. *Tropical Algebra*

Definition 2.1 (Tropical algebra)**.** The tropical semiring is defined on $\mathbb{R} \cup \{-\infty\}$ with the operations:

$$a \oplus b = \max(a, b),$$
$$a \odot b = a + b.$$

The neutral elements are $-\infty$ for $\oplus$ and 0 for $\odot$. This algebra is idempotent under addition ($a \oplus a = a$), which leads to combinatorial and geometric interpretations not found in classical algebra.

Using tropical algebra we can define tropical analogues of many classical objects. Of central importance is the tropical projective torus and its metric.

Embedding $\mathbb{R}^4/\mathbb{R}\mathbf{1} \to \mathbb{R}^3$*.*—We use

$$\varphi : \mathbb{R}^4/\mathbb{R}\mathbf{1} \longrightarrow \mathbb{R}^3, \qquad (x_0, x_1, x_2, x_3) \longmapsto (u, v, w) = (x_1 - x_0,\ x_2 - x_0,\ x_3 - x_0).$$

The tropical norm is $\|x\|_{\mathrm{tr}} = \max\{0, u, v, w\} - \min\{0, u, v, w\}$, hence the tropical ball $\{x : \|x\|_{\mathrm{tr}} \le r\}$ projects to the polyhedron

$$\mathcal{B}_r = \Big\{(u, v, w) \in \mathbb{R}^3 : |u| \le r,\ |v| \le r,\ |w| \le r,\ |u - v| \le r,\ |u - w| \le r,\ |v - w| \le r\Big\}.$$

Definition 2.2 (Tropical projective torus, tropical metric)**.** The n–dimensional *tropical projective torus* is the quotient

$$\mathbb{R}^{n+1}/\mathbb{R}\mathbf{1} \ = \ \big\{x \in \mathbb{R}^{n+1}\big\}/\big\{x \sim x + c\mathbf{1} \ : \ c \in \mathbb{R}\big\},$$

equipped with the tropical Hilbert metric

$$d_{\mathrm{tr}}(x, y) \ = \ \max_i (x_i - y_i) \ - \ \min_i (x_i - y_i),$$

induced by the tropical norm $\|x\|_{\mathrm{tr}} = \max_i x_i - \min_i x_i$.

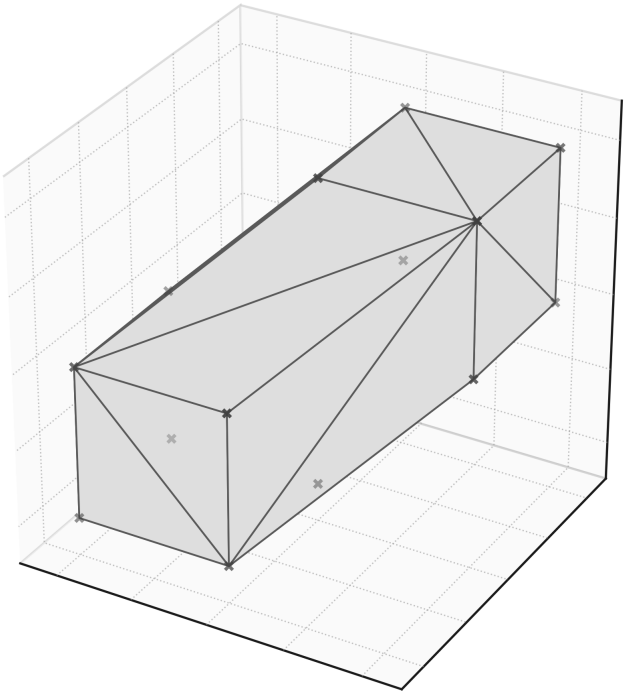


Figure 1. Tropical ball of radius $r = 1$ under $\varphi : \mathbb{R}^4/\mathbb{R}\mathbf{1} \to \mathbb{R}^3$. The polyhedron is cut out by the six families of inequalities $|u| \le 1$, $|v| \le 1$, $|w| \le 1$, $|u - v| \le 1$, $|u - w| \le 1$, $|v - w| \le 1$, and has 14 vertices listed in the text.

For visualization we use the embedding

$$\varphi : \mathbb{R}^{n+1}/\mathbb{R}\mathbf{1} \longrightarrow \mathbb{R}^n \qquad (x_0, \dots, x_n) \longmapsto (x_1 - x_0, \dots, x_n - x_0),$$

to identify $\mathbb{R}^{n+1}/\mathbb{R}$ with a subset of $\mathbb{R}^n$ because it removes the redundancy of global translation and centers the data around one coordinate which allows for a unique representation of all tropical vectors in the Euclidean space. It's particularly convenient when plotting tropical polytopes or computing pairwise distances since it provides an efficient representation of otherwise equivalence-class-defined data. i.e., we normalize the first coordinate to 0. Figure 1 depicts a tropical ball under φ in the case $n = 3$.

2.1.2. *Tropical Matrices*

To respect tropical structure we use tropical matrix maps and their convexity. The basic objects of tropical convexity are defined as follows.

Definition 2.3 (Tropical line segment; tropical convex hull)**.** For $a, b \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$, the tropical line segment between a and b is

$$\gamma_{ab} \;=\; \big\{\alpha \odot a \oplus \beta \odot b \;\big|\; \alpha, \beta \in \mathbb{R}\big\}.$$

For a finite set $X = \{x_1, \dots, x_r\} \subset \mathbb{R}^n/\mathbb{R}\mathbf{1}$, the tropical convex hull is

$$\operatorname{tconv}(X) \;=\; \Big\{\alpha_1 \odot x_1 \oplus \cdots \oplus \alpha_r \odot x_r \;\Big|\; \alpha_1, \dots, \alpha_r \in \mathbb{R}\Big\}.$$

We study tropical matrices $M \in \mathbb{R}^{m \times n}$ with at least one finite entry in every row (a nondegeneracy condition). The image of M in $\mathbb{R}^m/\mathbb{R}\mathbf{1}$ is the tropical convex hull of its columns.

Example 2.4 (A 4×5 matrix)**.** Let

$$M = \begin{pmatrix} 0 & 2 & -1 & 1 & 0 \\ 3 & -2 & 0 & 2 & -1 \\ 1 & 1 & 3 & -1 & 0 \\ 2 & 0 & 4 & -2 & 1 \end{pmatrix}.$$

For columns c_j of M we obtain

$$\varphi(c_1) = (3, 1, 2), \quad \varphi(c_2) = (-4, -1, -2), \quad \varphi(c_3) = (1, 4, 5), \quad \varphi(c_4) = (1, -2, -3), \quad \varphi(c_5) = (-1, 0, 1).$$

The image of M in $\mathbb{R}^4/\mathbb{R}\mathbf{1}$ is the tropical convex hull of its columns. For this configuration it coincides with the Euclidean convex hull of the five embedded points.

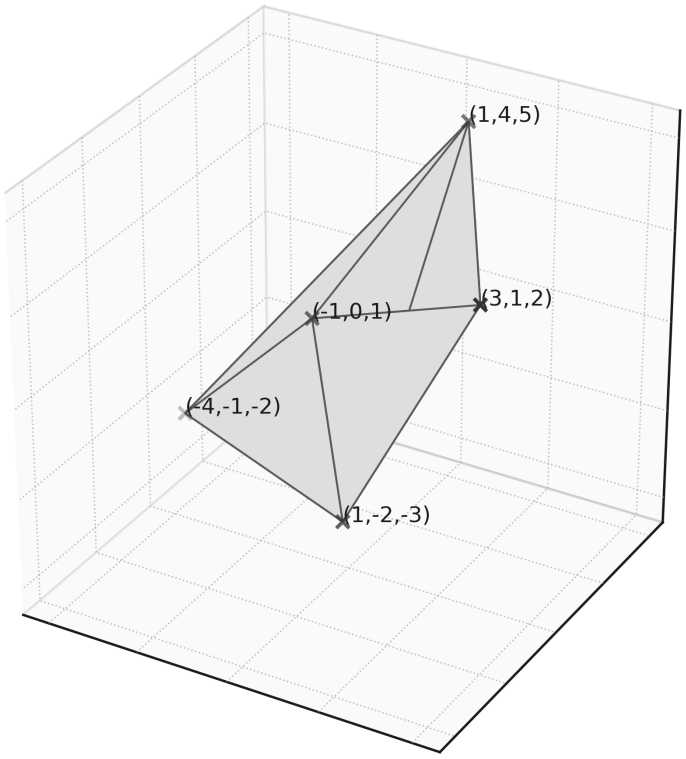


Figure 2. Image of the 4×5 matrix under $\varphi : \mathbb{R}^4/\mathbb{R}\mathbf{1} \to \mathbb{R}^3$. The polygonal surface is a triangulation of the convex hull of the embedded columns $\varphi(c_j)$.

Definition 2.5 (Type of a point)**.** Let $M \in \mathbb{R}^{m \times n}$. The *type* of $x \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$ with respect to M is the n–tuple $S = (S_1, \ldots, S_n)$ with

$$S_j \;=\; \big\{ i \in [m] : (Mx)_i = M_{ij} + x_j \big\}.$$

We write $\text{type}(x) = S$. The *type cell* of S is

$$X_S \;=\; \big\{ x \in \mathbb{R}^n/\mathbb{R}\mathbf{1} : S \subseteq \text{type}(x) \big\}.$$

The type cells $\{X_S\}$ form a polyhedral cell decomposition of $\mathbb{R}^n/\mathbb{R}\mathbf{1}$. On each cell the map $x \mapsto Mx$ is classically linear up to the projective equivalence.

2.2. *Wasserstein Distances Between Probability Distributions*

Wasserstein distances are generated from an optimal transport (OT) plan. Given a ground metric on a state space, the cost of moving mass between probability measures gives a distance that respects the geometry of the space and metrizes weak convergence under moment conditions.

Definition 2.6 (Wasserstein distance)**.** In Euclidean settings we take the ground metric to be the usual norm $\|x - y\|$ on $\Omega \subset \mathbb{R}^d$. Let $\rho_1 = \sum_{i=1}^{k} p_i \, \delta_{x_i}$ and $\rho_2 = \sum_{j=1}^{\ell} q_j \, \delta_{y_j}$ be discrete probability measures with $p_i, q_j \geq 0$ and

$\sum_i p_i = \sum_j q_j = 1$. The *Wasserstein distance* (i.e. 1-Wasserstein distance) is

$$W(\rho_1, \rho_2) = \inf_{\pi \in \Gamma(\rho_1, \rho_2)} \sum_{i=1}^{k} \sum_{j=1}^{\ell} \pi_{ij} \, \|x_i - y_j\|, \tag{1}$$

where $\Gamma(\rho_1, \rho_2) = \{\pi \in \mathbb{R}_+^{k \times \ell} : \sum_j \pi_{ij} = p_i, \ \sum_i \pi_{ij} = q_j\}$.

More generally, on a Polish metric space $(\Omega, \|\cdot\|)$ and for $\mu, \nu \in \mathcal{P}_1(\Omega)$,

$$W(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} \|x - y\| \, d\pi(x, y), \tag{2}$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν.

Remark 2.7 (Kantorovich–Rubinstein duality). The distance in (2) admits the dual representation

$$W(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \int_{\Omega} f(x) \, d(\mu - \nu)(x), \qquad \|f\|_{\mathrm{Lip}} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|}.$$

Definition 2.8 (p-Wasserstein distance)**.** Let $(\Omega, \|\cdot\|)$ be a Polish metric space and $p \in [1, \infty)$. For $\mu, \nu \in \mathcal{P}_p(\Omega)$ the *p-Wasserstein distance* is

$$W_p(\mu, \nu)^p = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} \|x - y\|^p \, d\pi(x, y). \tag{3}$$

In the discrete case of ρ_1, ρ_2 as above,

$$W_p(\rho_1, \rho_2) = \left(\inf_{\pi \in \Gamma(\rho_1, \rho_2)} \sum_{i=1}^{k} \sum_{j=1}^{\ell} \pi_{ij} \, \|x_i - y_j\|^p \right)^{1/p}. \tag{4}$$

Remark 2.9. W_1 in (1)–(2) is the special case $p = 1$ of (3)–(4). Larger p penalize long transports more heavily; the limit $W_\infty(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \operatorname{ess\,sup}_{(x,y) \sim \pi} \|x - y\|$ measures the maximal move under an optimal plan.

Remark 2.10 (Monge maps and Brenier's theorem). When $p = 2$, $\Omega = \mathbb{R}^d$ and μ is absolutely continuous, the optimal plan is induced by a transport map $T = \nabla \phi$ (gradient of a convex potential), yielding $W_2(\mu, \nu)^2 = \int \|x - T(x)\|^2 \, d\mu(x)$. See (Santambrogio 2015, §2.2).

Remark 2.11 (Equal-mass discrete supports). If $k = \ell$ and $p_i = q_j = 1/k$, the problem reduces to a minimum-cost bipartite matching (assignment) with cost matrix $c_{ij} = \|x_i - y_j\|^p$; network simplex / auction methods are standard solvers. See (Peyré and Cuturi 2019, §3.4).

2.3. *Wasserstein Distances Between Euclidean Spaces of Different Dimensions*

To compare measures on Euclidean spaces of different dimensions, Cai and Lim (2022) consider affine maps

$$\mathcal{M} \ = \ \big\{ \phi : \mathbb{R}^n \to \mathbb{R}^m \ : \ \phi(x) = Ax + b, \ A \in O(m, n), \ b \in \mathbb{R}^m \big\}.$$

They define the sets of projection and embedding measures

$$\Phi^-(\nu, m) := \big\{ \beta = \phi_{\#}\nu : \phi \in \mathcal{M} \big\}, \qquad \Phi^+(\mu, n) := \big\{ \alpha : \mu = \phi_{\#}\alpha \text{ for some } \phi \in \mathcal{M} \big\},$$

and the associated distances

$$W_p^-(\mu, \nu) := \inf_{\beta \in \Phi^-(\nu, m)} W_p(\mu, \beta), \tag{5}$$

$$W_p^+(\mu, \nu) := \inf_{\alpha \in \Phi^+(\mu, n)} W_p(\alpha, \nu). \tag{6}$$

We are able to map from lower to higher dimensional spaces or vice versa from Cai and Lim who show that for $m \le n$, $p \in [1, \infty]$, $\mu \in \mathcal{P}_p(\mathbb{R}^m)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^n)$,

$$W_p^-(\mu, \nu) \ = \ W_p^+(\mu, \nu).$$

3. RCV MEASURES AND INTER-ROUND MAPPINGS

Definition 3.1 (Round measures)**.** Let a round with d active candidates be represented on the tropical projective torus $\mathbb{R}^d/\mathbb{R}\mathbf{1}$ with ground metric

$$d_{\mathrm{tr}}(x,y) = \max_i(x_i - y_i) - \min_i(x_i - y_i).$$

A ballot distribution in that round is a probability measure $\mu \in \mathcal{P}_p(\mathbb{R}^d/\mathbb{R}\mathbf{1})$. The following round where one candidate is removed lives in $\mathbb{R}^{d-1}/\mathbb{R}\mathbf{1}$ with measure $\nu \in \mathcal{P}_p(\mathbb{R}^{d-1}/\mathbb{R}\mathbf{1})$. The tropical projective torus is Polish, hence $\mathcal{P}_p(\mathbb{R}^k/\mathbb{R}\mathbf{1})$ is well defined for all k and $p \geq 1$. (Monod et al. 2018, L1–L3)[3]

Remark 3.2 (Why this model is natural). The tropical Hilbert metric d_{tr} measures the width of the smallest tropical box that contains two points and in voting terms, it quantifies the largest change in log-scores a candidate receives when moving from x to y. Modeling a round as a probability measure on the tropical torus treats each ballot as an evaluation of these score differences where multiplying every score vector by the same tropical scalar translates all candidates equally, leaving the election outcome unchanged as required by the quotient $\mathbb{R}^d/\mathbb{R}\mathbf{1}$.

Proposition 3.3 (Non-expansivity of tropical matrices)**.** *For every tropical matrix M and all $x, y \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$,*

$$d_{\mathrm{tr}}(Mx, My) \leq d_{\mathrm{tr}}(x,y).$$

Source: *(Talbut et al. 2024, Prop. 9).*

Lemma 3.4 (Surjectivity criterion)**.** *A tropical matrix map $M : \mathbb{R}^n/\mathbb{R}\mathbf{1} \to \mathbb{R}^m/\mathbb{R}\mathbf{1}$ is surjective iff for every row i there exists a column $c(i)$ with entries*

$$M_{r,c(i)} = \begin{cases} -\infty, & r \neq i, \\ \textit{real}, & r = i. \end{cases}$$

Source: *(Talbut et al. 2024, Lemma 10).*

Example 3.5 (Verification in Small Dimension)**.** Take $M = \left(\begin{smallmatrix} -\infty & 0 & 0 \\ 0 & -\infty & 1 \end{smallmatrix}\right)$. Row 1 has its unique finite entry in column 2 and row 2 in column 1, so Lemma 3.4 gives the surjectivity $M : \mathbb{R}^3/\mathbb{R}\mathbf{1} \twoheadrightarrow \mathbb{R}^2/\mathbb{R}\mathbf{1}$.

For a target class $y = [(a,b)]$ choose

$$x \;=\; (x_1, x_2, x_3) \;:=\; \big(\max\{b,\, 2b-1-a\},\; a,\; b-1\big).$$

A quick check shows $Mx = (a,b) + (\max\{a, b-1\} - b)\mathbf{1}$, so Mx represents the same class as y. Adding any scalar to all three coordinates of x yields the whole fiber $M^{-1}(y)$.

Definition 3.6 (Simple projection)**.** A tropical matrix $M \in \mathbb{R}^{m \times n}$ $(n > m)$ is a simple projection if each column has at most one real entry. Writing $J_i := \{j : M_{ij} \in \mathbb{R}\}$, these sets are pairwise disjoint, nonempty, and

$$(Mx)_i = \max_{j \in J_i}\{M_{ij} + x_j\}, \qquad i = 1, \ldots, m.$$

Simple projections are surjective and every maximal fiber cell has dimension $n - m$. *Sources:* (Talbut et al. 2024, Def. 15; L49–L53)

Lemma 3.7 (Product decomposition)**.** *Let M be a simple projection $\mathbb{R}^n/\mathbb{R}\mathbf{1} \to \mathbb{R}^m/\mathbb{R}\mathbf{1}$ and let F_0 be the fiber at 0. Define $f_M(x) = (Mx,\, x - z_x)$ with*

$$(z_x)_j = \begin{cases} (Mx)_i - \max_k (Mx)_k, & j \in J_i, \\ 0, & j \notin \bigcup_i J_i. \end{cases}$$

Then f_M is well defined with image in $\mathbb{R}^m/\mathbb{R}\mathbf{1} \times F_0$. Source: *(Talbut et al. 2024, Lem. 17).*

[3] Cited via the statement that the tropical projective torus is Polish and p-Wasserstein is well-defined.

Proposition 3.8 (Homeomorphism and metric equivalence)**.** *The map f_M is a homeomorphism*

$$f_M : \mathbb{R}^n/\mathbb{R}\mathbf{1} \longrightarrow \mathbb{R}^m/\mathbb{R}\mathbf{1} \times F_0$$

and satisfies the two-sided bounds

$$\frac{1}{3}\, d_{\mathrm{tr}}(Mx_1, Mx_2) + \frac{1}{3}\, d_{\mathrm{tr}}(x_1 - z_{x_1}, x_2 - z_{x_2}) \;\le\; d_{\mathrm{tr}}(x_1, x_2) \;\le\; d_{\mathrm{tr}}(Mx_1, Mx_2) + d_{\mathrm{tr}}(x_1 - z_{x_1}, x_2 - z_{x_2}).$$

Source: *(Talbut et al. 2024, Prop. 18).*

Proposition 3.9 (Push-forward is non-expansive on measures)**.** *For all $p \in [1, \infty)$, any $\varphi \in \mathcal{M}_{\mathrm{tr}}$ (simple projection) and $\alpha, \nu \in \mathcal{P}_p(\mathbb{R}^n/\mathbb{R}\mathbf{1})$,*

$$W_{\mathrm{tr},p}(\varphi_{\#}\alpha, \varphi_{\#}\nu) \;\le\; W_{\mathrm{tr},p}(\alpha, \nu).$$

Source: *(Talbut et al. 2024, Prop. 19).*

Corollary 3.10 (Finite pth moments after projection)**.** *If $\nu \in \mathcal{P}_p(\mathbb{R}^n/\mathbb{R}\mathbf{1})$ and $\varphi \in \mathcal{M}_{\mathrm{tr}}$, then $\varphi_{\#}\nu \in \mathcal{P}_p(\mathbb{R}^m/\mathbb{R}\mathbf{1})$.*

We will occasionally write $\varphi_{\#} \colon \mathcal{P}_p(\mathbb{R}^n/\mathbb{R}\mathbf{1}) \to \mathcal{P}_p(\mathbb{R}^m/\mathbb{R}\mathbf{1})$ for the push-forward operator to emphasize the fact that it preserves the finiteness of p-moments.

Source: *(Talbut et al. 2024, Cor. 21).*

Definition 3.11 (Families of maps and inter-dimensional distances)**.** Let

$$\mathcal{M}_{\mathrm{tr}} := \{\varphi_M : \mathbb{R}^n/\mathbb{R}\mathbf{1} \to \mathbb{R}^m/\mathbb{R}\mathbf{1} \;:\; \varphi_M(x) = Mx,\ M \text{ simple projection}\}.$$

Define sets of projected and embedded measures

$$\Phi^-_{\mathrm{tr}}(\nu, m) = \{\beta \in \mathcal{P}(\mathbb{R}^m/\mathbb{R}\mathbf{1}) : \ \beta = \varphi_{\#}\nu \text{ for some } \varphi \in \mathcal{M}_{\mathrm{tr}}\},$$

$$\Phi^+_{\mathrm{tr}}(\mu, n) = \{\alpha \in \mathcal{P}(\mathbb{R}^n/\mathbb{R}\mathbf{1}) : \ \mu = \varphi_{\#}\alpha \text{ for some } \varphi \in \mathcal{M}_{\mathrm{tr}}\},$$

and the corresponding distances

$$W^-_{\mathrm{tr},p}(\mu, \nu) = \inf_{\beta \in \Phi^-_{\mathrm{tr}}(\nu, m)} W_{\mathrm{tr},p}(\mu, \beta), \qquad W^+_{\mathrm{tr},p}(\mu, \nu) = \inf_{\alpha \in \Phi^+_{\mathrm{tr}}(\mu, n)} W_{\mathrm{tr},p}(\alpha, \nu).$$

Source: (Talbut et al. 2024, Sec. 4.1).

Theorem 3.12 (Equivalence of projection and embedding distances)**.** *For $p \in [1, \infty)$, $\mu \in \mathcal{P}_p(\mathbb{R}^m/\mathbb{R}\mathbf{1})$ and $\nu \in \mathcal{P}_p(\mathbb{R}^n/\mathbb{R}\mathbf{1})$,*

$$W^-_{\mathrm{tr},p}(\mu, \nu) = W^+_{\mathrm{tr},p}(\mu, \nu) < \infty.$$

Source: *(Talbut et al. 2024, Thm. 22).*

Remark 3.13 (Euclidean analogue). Cai and Lim prove the same equality for Euclidean spaces using semi-orthogonal maps and translations (their Theorem 2.2). Talbut et al.'s construction mirrors theirs in the tropical setting. *Source:* (Cai and Lim 2022, Theorem 8 and eqs. (2)–(3)).

Proposition 3.14 (Lipschitz continuity)**.** *$W^{m,n}_{\mathrm{tr},p}$ is 1-Lipschitz:*

$$\left|W^{m,n}_{\mathrm{tr},p}(\mu_1, \nu_1) - W^{m,n}_{\mathrm{tr},p}(\mu_2, \nu_2)\right| \;\le\; W_{\mathrm{tr},p}(\mu_1, \mu_2) + W_{\mathrm{tr},p}(\nu_1, \nu_2).$$

Source: *(Talbut et al. 2024, Prop. 25).*

Corollary 3.15 (Consistency of the empirical estimator)**.** *Let $\widehat{\mu}_r, \widehat{\nu}_s$ be empirical measures from i.i.d. samples of μ, ν. Then $W^{m,n}_{\mathrm{tr},p}(\widehat{\mu}_r, \widehat{\nu}_s) \to W^{m,n}_{\mathrm{tr},p}(\mu, \nu)$ in probability as $r, s \to \infty$.* Source: *(Talbut et al. 2024, Cor. 26).*

3.1. *Formulation we will compute*

For finite samples $X = \{x^{(r)}\}_{r=1}^{R} \subset \mathbb{R}^n/\mathbb{R}\mathbf{1}$, $Y = \{y^{(s)}\}_{s=1}^{S} \subset \mathbb{R}^m/\mathbb{R}\mathbf{1}$, let $\widehat{\nu} = \frac{1}{R}\sum_r \delta_{x^{(r)}}$ and $\widehat{\mu} = \frac{1}{S}\sum_s \delta_{y^{(s)}}$. Adapting (Talbut et al. 2024, Sec. 4.1), introduce a translated/simple-projection class M', t (with a technical enlargement J_i' ensuring feasibility on a bounded ball) and optimize

$$W_{\mathrm{tr},p}^{m,n}(\widehat{\mu},\widehat{\nu}) = \inf_{t,\ M'} \inf_{\lambda \in \mathbb{R}_+^{R\times S}} \left(\sum_{r,s} \lambda_{rs} \left\| M'\big(x^{(r)} + t\big) - y^{(s)} \right\|_{\mathrm{tr}}^{p} \right)^{1/p}, \quad \sum_s \lambda_{rs} = \tfrac{1}{R},\ \sum_r \lambda_{rs} = \tfrac{1}{S}. \tag{7}$$

Source: (Talbut et al. 2024, Eq. (24)).

4. COMPUTATION

We develop two methods for generating ranked ballots (Algorithm 1 and Algorithm 2). From the simulated ballots we build empirical measures on the appropriate state spaces and compute both Euclidean and tropical 2-Wasserstein distances. When successive rounds have different numbers of candidates we use the projection/embedding framework of Definition 3.11 and Theorem 3.12 and compute the projection form, which equals the embedding form by Theorem 3.12.

4.1. *Simulations*

Fix a set of candidates indexed by $[d]$. Each simulation produces a table whose row is a single ballot and whose columns `Rank_1`,...,`Rank_r` record the candidate labels placed at each rank (with `0` meaning the ballot exhausted before that rank).

Simulation 1 (independent weighted scores).—Let $p = (p_1, \ldots, p_d)$ be nonnegative base shares with $\sum_i p_i = 1$ and let r be the number of recorded ranks (default $r = \#\{i : p_i > 0\}$). For each ballot we draw i.i.d. uniform scores $U_i \sim \mathrm{Unif}(0,1)$ and set

$$\mathrm{score}_i \ = \ p_i\, U_i.$$

We sort candidates by decreasing score and keep the top r distinct labels. Thus ballots never exhaust, and higher p_i raise a candidate's chance of appearing early. See Algorithm 1.

Algorithm 1 Simulation 1: independent weighted scores

1. **Input:** base shares $p \in \mathbb{R}^d$ with $\sum_i p_i = 1$, number of ballots N, ranks r.
2. For each ballot $b = 1, \ldots, N$:
   (a) Draw $U_i \sim \mathrm{Unif}(0,1)$, set $\mathrm{score}_i = p_i U_i$ for $i \in [d]$ with $p_i > 0$.
   (b) Sort candidates by decreasing score and keep the top r labels. Record them as `Rank_1`, ..., `Rank_r`.

Simulation 2 (sequential choice with exhaustion via annealing schedule).—We sequentially build the ballot. At step $j = 1, 2, \ldots, r$, with the remaining candidate set A, we first flip exhaustion with probability

$$\Pr\{\text{exhaust at step } j\} \ = \ p_{\mathrm{exh}}(j; r) \ = \ \min\{1, p_{\mathrm{end}} \Big(\frac{j - j_{\min}}{\max\{1, r - j_{\min} - 1\}} \Big)_+^{\gamma} \}$$

where $j_{\min}$ is the number of ranks guaranteed to be filled (default $j_{\min} = 8$), $p_{\mathrm{end}} \in [0,1]$ is the exhaustion probability at the final rank (default 0.25), and $\gamma > 0$ shapes the curve. This process, motivated by the cooling in simulated annealing, drops the effective temperature as j grows, increasing the chance of exhaustion relative to a polynomial curve. If no exhaustion occurs, we sample a candidate from A with probability proportional to its base share p_i, remove it, and proceed. Any remaining ranks are padded with `0`. See Algorithm 2.

Algorithm 2 Simulation 2: sequential choice with exhaustion (annealing schedule)

1. **Input:** shares p, N, r, parameters $j_{\min}$, p_{end}, γ.
2. For each ballot $b = 1, \dots, N$:
   (a) $A \leftarrow \{\, i : p_i > 0\}$, *ballot* $\leftarrow []$.
   (b) For $j = 1$ to r:
      i. With probability $p_{\text{exh}}(j; r)$, append zeros until length r and **break**.
      ii. Else sample $i \in A$ with probability $p_i / \sum_{k \in A} p_k$, append i, and set $A \leftarrow A \setminus \{i\}$.
   (c) If length $< r$, pad with zeros.

4.2. *From ballots to empirical measures*

Let $X = \{x^{(r)}\}_{r=1}^{R}$ and $Y = \{y^{(s)}\}_{s=1}^{S}$ denote the two samples of ballot vectors we will compare (e.g., two rounds). We consider two representations.

Euclidean representation.—We map each ballot to a vector $v \in \mathbb{R}^d$ using Borda-type weights $w_1 \geq \cdots \geq w_r > 0$:

$$v_i \;=\; \begin{cases} w_j, & \text{if candidate } i \text{ appears at rank } j, \\ 0, & \text{if not ranked by this voter.} \end{cases}$$

Typical choices are $w_j = 1/j$ or $w_j = r - j + 1$. This produces empirical measures $\widehat{\nu} = \frac{1}{R}\sum_{r=1}^{R} \delta_{x^{(r)}}$ and $\widehat{\mu} = \frac{1}{S}\sum_{s=1}^{S} \delta_{y^{(s)}}$ on $\mathbb{R}^d$.

Tropical representation.—We encode a ballot as $u \in \mathbb{R}^d$ by

$$u_i \;=\; \begin{cases} -j, & \text{if candidate } i \text{ appears at rank } j, \\ -(r+1), & \text{if not ranked (exhausted before including } i). \end{cases}$$

Subtracting a constant does not change d_{tr}, so any centering is allowed; we subtract $\max_i u_i$ to keep vectors near the origin. This yields empirical measures on $\mathbb{R}^d/\mathbb{R}\mathbf{1}$.

4.3. *Euclidean 2-Wasserstein*

For two empirical measures supported on $X = \{x^{(r)}\}_{r=1}^{R} \subset \mathbb{R}^d$ and $Y = \{y^{(s)}\}_{s=1}^{S} \subset \mathbb{R}^d$,

$$\widehat{\mu} \;=\; \tfrac{1}{R}\sum_{r=1}^{R} \delta_{x^{(r)}}, \qquad \widehat{\nu} \;=\; \tfrac{1}{S}\sum_{s=1}^{S} \delta_{y^{(s)}},$$

the discrete 2-Wasserstein distance is defined as

$$W_2(\widehat{\mu}, \widehat{\nu})^2 \;=\; \min_{\lambda \in \mathbb{R}_+^{R \times S}} \sum_{r=1}^{R}\sum_{s=1}^{S} \lambda_{rs} \, \|x^{(r)} - y^{(s)}\|_2^2, \quad \sum_{s=1}^{S} \lambda_{rs} = \tfrac{1}{R}, \; \sum_{r=1}^{R} \lambda_{rs} = \tfrac{1}{S}. \tag{8}$$

Derivation.—Recall that the 2-Wasserstein distance on a Polish metric space $(\Omega, \|\cdot\|_2)$ is defined by

$$W_2(\mu, \nu)^2 \;=\; \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} \|x - y\|_2^2 \, d\pi(x, y), \tag{9}$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν (Cai and Lim 2022, Sec. 2.1).

For empirical measures every coupling π is supported on the finite grid $X \times Y$. Writing $\lambda_{rs} := \pi\big(\{x^{(r)}\} \times \{y^{(s)}\}\big)$, the integral in (9) collapses to the finite sum in (8), while the marginal constraints on π become

$$\sum_{s=1}^{S} \lambda_{rs} = \widehat{\mu}\big(\{x^{(r)}\}\big) = \tfrac{1}{R}, \qquad \sum_{r=1}^{R} \lambda_{rs} = \widehat{\nu}\big(\{y^{(s)}\}\big) = \tfrac{1}{S}.$$

These two sets of linear equations force the row and column totals of the transport matrix λ to match the empirical masses, thereby ensuring $\lambda \in \Gamma(\widehat{\mu}, \widehat{\nu})$. In matrix notation the feasible region is the transport polytope

$$\Big\{\lambda \in \mathbb{R}_+^{R\times S} : \ \lambda \mathbf{1}_S = \tfrac{1}{R}\mathbf{1}_R,\ \lambda^\top \mathbf{1}_R = \tfrac{1}{S}\mathbf{1}_S\Big\},$$

so (8) is a linear programme with cost matrix $C_{rs} := \|x^{(r)} - y^{(s)}\|_2^2$.

Special case: balanced samples.—If $R = S$ and all masses are equal, each row and column must sum to $1/R$. An optimal plan can therefore be taken to be a permutation matrix, turning (8) into a minimum-cost bipartite assignment (Cai and Lim 2022, Lem. 2.3). Network-simplex or auction algorithms solve this in $\mathcal{O}(R^3)$ time, which is our default backend for precinct-level matching in the RCV experiments.

General case: unbalanced or weighted samples.—When $R \neq S$ or the empirical masses differ, the problem remains a linear programme over the full $R \times S$ grid of variables. We solve it with an entropically regularized Sinkhorn method. An outer loop updates any quantities that change the cost matrix and an inner loop runs Sinkhorn to update the coupling λ, just like in Algorithm 3 of our appendix.

4.4. *Tropical 2-Wasserstein*

With tropical metric

$$d_{\mathrm{tr}}(u, v) = \max_i (u_i - v_i) - \min_i (u_i - v_i),$$

we compute

$$W_2^{\mathrm{tr}}(\widehat{\mu}, \widehat{\nu})^2 \ = \ \min_{\lambda \in \mathbb{R}_+^{R\times S}} \sum_{r=1}^{R}\sum_{s=1}^{S} \lambda_{rs}\, d_{\mathrm{tr}}\big(x^{(r)}, y^{(s)}\big)^2, \quad \sum_s \lambda_{rs} = \tfrac{1}{R},\ \sum_r \lambda_{rs} = \tfrac{1}{S}.$$

4.5. *Across Dimensions Via a Fixed Embedding*

If one round has d candidates and the next has $d-1$, we still refer to the projection family $\mathcal{M}_{\mathrm{tr}}$ of Definition 3.6 (following the Cai–Li construction) to organize the discussion. For an eliminated candidate k, the simple maps M_k group columns so that the kth coordinate is discarded and the survivors retain one real entry per column; by Proposition 3.3 and Lemma 3.4 such maps are non-expansive and surjective. In our implementation, however, we do not optimize over $\varphi \in \mathcal{M}_{\mathrm{tr}}$. Instead, we compute distances in a common ambient space through index-aware zero insertion. For each $(d-1)$-vector $y^{(s)}$ obtained by removing candidate k from a row, let $\kappa^{(s)} \in [d]$ be the index at which the removal occurred. Define $\mathrm{pad}(y^{(s)})$ as $y^{(s)}$ with a 0 inserted at the coordinate $\kappa^{(s)}$ (and set $\mathrm{pad}(x^{(r)}) = x^{(r)}$ for already d-dimensional rows).

We then minimize

$$\big(W_{\mathrm{tr},2}^{-}(\widehat{\mu}, \widehat{\nu})\big)^2 = \min_{\lambda \in \mathbb{R}_+^{R\times S}} \sum_{r,s} \lambda_{rs}\, d_{\mathrm{tr}}\big(\mathrm{pad}(x^{(r)}),\, \mathrm{pad}(y^{(s)})\big)^2, \quad \sum_s \lambda_{rs} = \tfrac{1}{R},\ \sum_r \lambda_{rs} = \tfrac{1}{S}. \tag{10}$$

For tropical distances we apply the rowwise centering map $T(u) = u - (\max_i u_i)\mathbf{1}$ after constructing the vectors used in the comparison. For the unweighted tropical metric d_{tr} this does not affect values , but fixes a canonical representative. For the weighted tropical metric $d_{\mathrm{tr},w}$ this gauge is required. All $W_{2,w}^{\mathrm{tr}}$ values are computed with $T(u), T(v)$. We do not center for Euclidean distances.

The padding we use is compatible with the Cai–Li template as it fixes a particular embedding and inherits non-expansiveness from Proposition 3.3. Its value serves as an upper-bound for the optimal embedding distance and by equality of minima, the minimum for the projection as well. Note that this is a design choice for computational convenience and not what Cai-Lim/Talbut optimizes.

For Euclidean baselines we analogously use padding rather than optimizing an inter-round map:

$$\big(W_{\mathrm{eu},2}^{-}(\widehat{\mu}, \widehat{\nu})\big)^2 = \min_{\lambda \in \mathbb{R}_+^{R\times S}} \sum_{r,s} \lambda_{rs} \big\|\mathrm{pad}(x^{(r)}) - \mathrm{pad}(y^{(s)})\big\|_2^2, \quad \sum_s \lambda_{rs} = \tfrac{1}{R},\ \sum_r \lambda_{rs} = \tfrac{1}{S}. \tag{11}$$

Proposition 4.1 (Well-posedness and attainment for padded OT)**.** *Let $\widehat{\mu}, \widehat{\nu}$ be the empirical measures in §4.3, and let both samples be embedded in a common ambient space by the trailing-zero padding of §4.5 such that $x^{(r)}, y^{(s)} \in \mathbb{R}^d$ with $d = \max\{d_X, d_Y\}$. We set $p = 2$.*

(i) The padding-based objectives Equations (10) and (11) are well-posed and finite:

$$0 \le \left(W^{-}_{\mathrm{tr},2}(\widehat{\mu},\widehat{\nu})\right)^2 < \infty, \qquad 0 \le \left(W^{-}_{\mathrm{eu},2}(\widehat{\mu},\widehat{\nu})\right)^2 < \infty.$$

Each objective equals the squared 2-Wasserstein cost between the padded empirical measures $\mathrm{pad}_{\#}\widehat{\mu}$ *and* $\mathrm{pad}_{\#}\widehat{\nu}$ *with the corresponding ground metric (*d_{tr} *or* $\|\cdot\|_2$*). Furthermore, since padding fixes a translated simple projection (cf. §4.5), the tropical value in Equation (10) upper-bounds the inter-round projection distance that minimizes over* $\mathcal{M}_{\mathrm{tr}}$*, with equality whenever the fixed padding coincides with an optimal simple projection.*

(ii) In both Equation (10) and Equation (11) there exists a minimizer $\lambda^\star \in \mathbb{R}_+^{R\times S}$ *attaining the minimum. If* $R = S$ *and all masses are equal, an optimal plan can be chosen to be a scaled permutation matrix.*

Proof. **(i)** After padding, both problems are standard discrete 2-Wasserstein linear programs with cost matrix $C_{rs} = d_{\mathrm{tr}}(\mathrm{pad}(x^{(r)}), \mathrm{pad}(y^{(s)}))^2$ for the tropical case and $C_{rs} = \left\|\mathrm{pad}(x^{(r)}) - \mathrm{pad}(y^{(s)})\right\|_2^2$ for the Euclidean case. Since X and Y are finite, all C_{rs} are finite, and the uniform coupling $\lambda_{rs} = 1/(RS)$ is possible which proves fitness. By §4.3, the objective value is the squared 2-Wasserstein distance between the padded empirical measures with the chosen ground metric. In the tropical setting, trailing-zero padding is a translated simple projection (see §4.5), hence its value upper-bounds the minimum over all simple projections in $\mathcal{M}_{\mathrm{tr}}$, with equality when padding matches an optimal member of $\mathcal{M}_{\mathrm{tr}}$.

(ii) For either ground metric the feasible region is the transport polytope

$$\left\{\lambda \in \mathbb{R}_+^{R\times S} : \ \lambda \mathbf{1}_S = \tfrac{1}{R}\mathbf{1}_R,\ \lambda^\top \mathbf{1}_R = \tfrac{1}{S}\mathbf{1}_S\right\},$$

which is nonempty, compact, and convex. The objective is linear and therefore continuous in λ, so a minimizer $\lambda^\star$ exists by the Weierstrass theorem. When $R = S$ with equal masses, the feasible set is R^{-1} times the Birkhoff polytope. □

Implementation.—We equalize dimensionality by zero-padding the smaller sample and then solving a single optimal transport problem per candidate removal as shown through Algorithms 3 to 6.

Algorithm 3 Euclidean W_2 between round d and $d-1$

for *sim_version* $= 1, 2$ **do** ▷ choose ballot generator
  for *cand_out* $= 1$ **to** n **do** ▷ delete each candidate once
    $X \leftarrow$ SIMULATE_BALLOTS$(p, N, r, seed)$ ▷ $X \in \mathbb{R}^{R\times d}$
    $Y \leftarrow$ REMOVE_ONE_PER_ROW$(X, cand_out)$ ▷ $Y \in \mathbb{R}^{S\times(d-1)}$
    Pad the lower-dimensional matrix so $X, Y \in \mathbb{R}^{\bullet\times\hat{d}}$ with $\hat{d} = \max\{d, d-1\}$
    $C_{rs} \leftarrow \|x^{(r)} - y^{(s)}\|_2^2$
    Solve OT: $\min_{\lambda\ge 0}\langle\lambda, C\rangle$ **s.t.** $\lambda\mathbf{1} = \frac{1}{R}\mathbf{1},\ \lambda^\top\mathbf{1} = \frac{1}{S}\mathbf{1}$
    $W_2 \leftarrow \sqrt{\text{opt. value}}$; **print** W_2
  end for
end for

Remark. Algorithm 3 generates two empirical ballot matrices, removes candidate n to obtain the $(d-1)$-dimensional set, builds the squared Euclidean cost matrix, solves the optimal-transport linear program with uniform weights, and returns the classical 2-Wasserstein distance. Padding makes the two matrices comparable when dimensions differ.

Algorithm 4 Tropical W_p^{tr} $(p = 2)$

for *sim_version* $= 1, 2$ **do**
  for *cand_out* $= 1$ **to** n **do**
    Generate X, Y as in Alg. 3; pad the lower-dimensional matrix
    $C_{rs} \leftarrow \max_i(x_i^{(r)} - y_i^{(s)}) - \min_i(x_i^{(r)} - y_i^{(s)})$
    if $p < \infty$ **then**
      Solve OT: $\min_{\lambda \geq 0} \langle \lambda, C^p \rangle$ **s.t.** $\lambda \mathbf{1} = \frac{1}{R}\mathbf{1}$, $\lambda^\top \mathbf{1} = \frac{1}{S}\mathbf{1}$
      return $(\langle \lambda^\star, C^p \rangle)^{1/p}$
    else
      Binary-search the smallest t s. t. there exists λ with support where $C \leq t$
    end if
    print W_p^{tr}
  end for
end for

Remark. Algorithm 4 mirrors Algorithm 3, but swaps the Euclidean norm for the tropical metric d_{tr} inside the same OT framework. We use $p = 2$ to minimize the sum of squared tropical costs.

Definition 4.2 (Weighted tropical distance)**.** For $u, v \in \mathbb{R}^d$ and weights $w \in \mathbb{R}^d_{>0}$,

$$d_{\mathrm{tr},w}(u, v) = \max_{1 \leq k \leq d} \big(w_k(u_k - v_k)\big) - \min_{1 \leq k \leq d} \big(w_k(u_k - v_k)\big).$$

With $W = \mathrm{Diag}(w)$ we have $d_{\mathrm{tr},w}(u, v) = d_{\mathrm{tr}}(Wu, Wv)$, so $d_{\mathrm{tr},w}$ is a metric on $\mathbb{R}^d$. The weighted tropical $W_{2,w}$ distance is obtained from (10) by replacing d_{tr} with $d_{\mathrm{tr},w}$.

Remark 4.3. Unlike the unweighted d_{tr}, the quantity $d_{\mathrm{tr},w}$ is not invariant under adding $c\,\mathbf{1}$ to either argument unless $w_1 = \cdots = w_d$. Hence it does not descend to a metric on the tropical projective torus $\mathbb{R}^d/\mathbb{R}\mathbf{1}$ in general. We work in $\mathbb{R}^d$ with a fixed gauge and normalize each row by $\tilde{u} = u - (\max_i u_i)\mathbf{1}$ and $\tilde{v} = v - (\max_i v_i)\mathbf{1}$, and evaluate $d_{\mathrm{tr},w}(\tilde{u}, \tilde{v})$. (Any positive rescaling of w rescales $d_{\mathrm{tr},w}$ by the same factor; we set $\max_k w_k = 1$.)

Definition 4.4 (Weighted Euclidean metric)**.** For $u, v \in \mathbb{R}^d$ and weights $w \in \mathbb{R}^d_{>0}$, define

$$d_{\mathrm{eu},w}(u, v) \;:=\; \Big(\sum_{k=1}^{d} w_k\,(u_k - v_k)^2\Big)^{1/2} \;=\; \big\|\mathrm{Diag}(\sqrt{w})\,(u - v)\big\|_2.$$

The weighted Euclidean $W_{2,w}$ distance is obtained from (11) by replacing the Euclidean ℓ_2 norm $\|\cdot\|_2^2$ with $d_{\mathrm{eu},w}^2$

Remark 4.5 (Choice of weights). Any $w \in \mathbb{R}^d_{>0}$ yields a valid metric $d_{\mathrm{tr},w}$. In experiments we use exponentially decaying rank-weights

$$w_k \;=\; \exp\big(-\gamma(k-1)\big), \qquad \gamma \geq 0,$$

and normalize by $\max_k w_k = 1$ (any positive rescaling of w scales $d_{\mathrm{tr},w}$ by the same constant). When rounds differ in dimension, we extend w by appending $w_d = \exp(-\gamma(d-1))$ after padding.

Algorithm 5 Weighted Euclidean $W_{2,w}$

for *sim_version* $= 1, 2$ **do**
  for *cand_out* $= 1$ **to** n **do**
    Build X, Y as before; pad the lower-dimensional matrix to width $\hat{d}$
    $w_k \leftarrow \exp(-\gamma(k-1))$ for $k = 1, \ldots, \hat{d}$; normalize $w \leftarrow w / \max_k w_k$
    $C_{rs} \leftarrow \sum_{k=1}^{\hat{d}} w_k\,(x_k^{(r)} - y_k^{(s)})^2$
    Solve OT: $\min_{\lambda \geq 0} \langle \lambda, C \rangle$ **s.t.** $\lambda \mathbf{1} = \frac{1}{R}\mathbf{1}$, $\lambda^\top \mathbf{1} = \frac{1}{S}\mathbf{1}$
    $W_{2,w} \leftarrow \sqrt{\text{opt. value}}$; **print** $W_{2,w}$
  end for
end for

Remark. Algorithm 5 down-weights later ranks exponentially ($\gamma = 0.2$ by default), reflecting the greater importance of early choices in RCV. Everything else mirrors the logic from Algorithm 3.

Algorithm 6 Weighted Tropical $W_{p,w}^{\mathrm{tr}}$ $p = 2$

for *sim_version* $= 1, 2$ **do**
  for *cand_out* $= 1$ **to** n **do**
    Form X, Y; pad lower-dimensional matrix to width $\hat{d}$; compute w_k as above
    $C_{rs} \leftarrow \max_k w_k(x_k^{(r)} - y_k^{(s)}) \; - \; \min_k w_k(x_k^{(r)} - y_k^{(s)})$
    if $p < \infty$ **then**
      Solve OT: $\min_{\lambda \geq 0} \langle \lambda, C^p \rangle$ **s.t.** $\lambda \mathbf{1} = \frac{1}{R}\mathbf{1}$, $\lambda^\top \mathbf{1} = \frac{1}{S}\mathbf{1}$
      return $(\langle \lambda^\star, C^p \rangle)^{1/p}$
    else
      Binary-search threshold t as in Alg. 4
    end if
    print $W_{p,w}^{\mathrm{tr}}$
  end for
end for

Remark. Algorithm 6 combines the exponential rank weights with the tropical metric. Because the weighted cost is still max–minus–min, everything else mirrors the logic from Algorithm 4.

The code is available at this link.

4.6. *Figures*

This section displays the tables from running the code in Python with seed 1. The full data sheet we ran with other seeds 10 and 100 are available in this link.

4.6.1. *Actual Unweighted*

Euclidean and tropical W_2 distances – Actual ballots, un-weighted

Candidate	Round 6→ Round 7				Round 10→Round 11				Round 12→ Round 13			
	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T
1	9.674	9.475	11.140	9.876	6.063	6.202	7.551	7.509	3.260	3.179	3.828	3.752
2	9.656	9.435	11.165	9.839	6.040	6.145	7.465	7.466	3.128	3.113	3.658	3.641
3	9.646	9.435	11.135	9.881	5.991	6.148	7.410	7.478	3.145	3.150	3.459	3.428
4	9.631	9.376	11.086	9.791	6.003	6.192	7.451	7.523	3.208	3.145	2.968	2.874
5	9.300	9.178	10.640	9.554	5.475	5.704	6.000	6.362		...		
6	9.240	9.078	10.566	9.374	5.442	5.606	5.384	5.659		...		
7	9.010	8.707	9.804	8.960		...				...		
8	8.975	8.621	9.603	8.755		...				...		
9	8.949	8.539	9.408	8.488		...				...		
10	8.928	8.487	8.980	8.204		...				...		

Distances are computed with 150 simulated ballots per round (two generators × two metrics).

4.6.2. *Skewed-Left Unweighted*

Euclidean and tropical W_2 distances – Simulated skewed-left data, un-weighted

Candidate	Round 6→ Round 7				Round 10→Round 11				Round 12→ Round 13			
	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T
1	9.646	9.977	10.892	10.651	5.280	5.365	6.116	6.137	3.596	3.637	4.235	4.298
2	9.553	9.803	10.841	10.543	5.151	5.298	5.973	6.124	3.519	3.570	4.189	4.304
3	9.563	9.848	10.851	10.589	5.058	5.252	5.786	6.050	3.423	3.559	3.917	4.116
4	9.523	9.818	10.754	10.544	4.997	5.229	5.550	5.892	3.431	3.414	3.274	3.265
5	9.443	9.681	10.590	10.361	4.971	5.190	5.292	5.585		...		
6	9.426	9.610	10.476	10.291	4.989	5.109	4.844	5.009		...		
7	9.345	9.532	10.337	10.060		...				...		
8	9.289	9.354	10.018	9.757		...				...		
9	9.188	9.200	9.757	9.419		...				...		
10	8.973	8.798	8.901	8.669		...				...		

Distances are based on 150 ballots per round (two generators × two metrics).

4.6.3. *Uniform Unweighted*

Euclidean and tropical W_2 distances – Uniform ballots, un-weighted

Candidate	Round 6→ Round 7				Round 10→Round 11				Round 12→ Round 13			
	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T
1	8.569	7.746	9.597	8.231	4.662	4.362	5.366	4.992	3.018	2.923	3.583	3.479
2	8.685	7.756	9.524	8.276	4.642	4.279	5.396	4.924	2.948	2.850	3.440	3.295
3	8.596	7.833	9.551	8.404	4.620	4.288	5.255	4.821	2.962	2.898	3.154	3.055
4	8.574	7.844	9.505	8.377	4.539	4.290	4.980	4.604	3.033	2.964	2.722	2.605
5	8.576	7.720	9.389	8.140	4.610	4.334	4.833	4.386		...		
6	8.554	7.766	9.165	8.144	4.548	4.360	4.327	4.060		...		
7	8.641	7.819	8.984	7.972		...				...		
8	8.573	7.768	8.720	7.758		...				...		
9	8.595	7.981	8.610	7.787		...				...		
10	8.565	7.718	8.422	7.391		...				...		

Uniform ballot generator; same formatting conventions.

4.6.4. *Actual Weighted*

Euclidean and tropical W_2 distances – Actual ballots, weighted

Candidate	Round 6→ Round 7				Round 10→Round 11				Round 12→ Round 13			
	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T
1	6.687	6.757	5.917	5.625	4.931	5.009	5.162	5.081	2.907	2.814	3.298	3.173
2	6.637	6.657	5.949	5.561	4.898	4.928	5.076	5.082	2.740	2.726	3.012	2.992
3	6.600	6.669	5.878	5.596	4.840	4.936	4.990	5.059	2.751	2.772	2.751	2.730
4	6.574	6.555	5.780	5.480	4.847	4.995	4.934	5.007	2.811	2.771	2.222	2.142
5	5.962	6.174	4.788	4.896	4.190	4.417	3.597	3.889		...		
6	5.865	5.992	4.630	4.668	4.163	4.314	3.158	3.359		...		
7	5.536	5.567	4.017	4.107		...				...		
8	5.489	5.533	3.814	3.976		...				...		
9	5.434	5.344	3.583	3.529		...				...		
10	5.384	5.337	3.280	3.466		...				...		

Distances are computed with 150 simulated ballots per round (two generators × two metrics).

4.6.5. *Skewed-Left Weighted*

Euclidean and tropical W_2 distances – Simulated skewed-left data, weighted

Candidate	Round 6→ Round 7				Round 10→Round 11				Round 12→ Round 13			
	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T
1	6.814	7.409	6.414	6.610	4.433	4.381	4.817	4.540	3.211	3.225	3.650	3.636
2	6.449	6.918	5.516	5.801	4.282	4.340	4.488	4.299	3.088	3.099	3.420	3.420
3	6.424	6.940	5.443	5.862	4.162	4.242	4.350	4.193	2.977	3.096	3.087	3.090
4	6.310	6.848	5.186	5.711	4.083	4.199	4.117	4.069	2.991	3.008	2.842	2.873
5	6.151	6.679	4.907	5.525	4.080	4.195	3.924	3.873		...		
6	6.144	6.611	4.755	5.387	4.084	4.101	3.588	3.765		...		
7	6.312	6.818	4.580	5.204		...				...		
8	6.222	6.716	4.420	5.008		...				...		
9	6.151	6.584	4.282	4.894		...				...		
10	6.017	6.404	4.121	4.712		...				...		

Same simulation settings as Table 1, but with weight-adjusted distances.

4.6.6. *Uniform Weighted*

Euclidean and tropical W_2 distances – Uniform ballots, weighted

Candidate	Round 6→ Round 7				Round 10→Round 11				Round 12→ Round 13			
	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T	Sim1 NonT	Sim2 NonT	Sim1 T	Sim2 T
1	6.468	6.275	5.854	5.858	3.797	3.550	3.879	3.615	2.678	2.592	3.049	2.956
2	6.654	6.182	6.101	5.859	3.788	3.518	3.931	3.587	2.593	2.502	2.871	2.737
3	6.481	6.316	5.954	5.636	3.696	3.481	3.821	3.512	2.605	2.556	2.521	2.465
4	6.446	6.202	5.853	5.604	3.668	3.486	3.746	3.472	2.699	2.639	2.071	1.947
5	6.385	6.039	5.686	5.533	3.711	3.538	3.712	3.527		...		
6	6.387	6.093	5.557	5.394	3.667	3.564	3.583	3.495		...		
7	6.562	6.177	5.780	5.614		...				...		
8	6.464	6.101	5.659	5.500		...				...		
9	6.358	6.003	5.533	5.363		...				...		
10	6.140	5.964	5.393	5.195		...				...		

Uniform ballot generator; weighted W_2 distances (150 ballots per round, two generators × two metrics).

5. DISCUSSION

In this paper, we simulate RCV ballots and recast each elimination step as an OT problem between successive empirical distributions of the rank vectors. We measure four distances, the classical and tropicalized 2-Wasserstein metrics as well as their weighted counterparts to quantify how ballots redistribute once a candidate is removed. The unweighted Euclidean 2-Wasserstein captures the average quadratic work required to move the entire ballot distribution from one round to the next and in our data, we can see that it fell almost monotonically. In every panel, Euclidean W_2 declines with each elimination. For actual ballots (Table 4.6.1, Sim1), W_2 is ≈ 9.67 at $6 \to 7$, ≈ 6.06 at $10 \to 11$, and ≈ 3.26 at $12 \to 13$. Uniform and skewed-left data show the same trend with lower baselines for the more homogeneous uniform case (Table 4.6.3: $\approx 8.57 \to 3.02$). In the weighted version, the data for actual ballots (Table 4.6.4, Sim1, candidate 1) show that the Euclidean distance drops from $\approx$ 6→7 to ≈12→13; uniform ballots (Table 4.6.6) show ≈6→7 → ≈12→13; and skewed-left ballots (Table 4.6.5) show ≈6→7 → ≈ 12→13. Even after applying the exponential kernel $\omega_k = \exp -\gamma(k-1)$, the curved remained monotonic with an exhibition of a steeper early drop. Specifically, distances shrink by roughly 25–35% in early rounds and 10–15% in late rounds (compare Table 4.6.1 vs. 4.6.4; Table 4.6.3 vs. 4.6.6). This reflects our intent of de-emphasizing changes confined to deep ranks while keeping sensitivity to early-rank moves. This weighted version improves on the traditional calculation measurement of outcome variance after each elimination by essentially neutralizing the effect of the candidate account across each round to round comparison.

Weighted tropical behaves analogously but preserves the sharper peaks when a small bloc makes a large rank jump. While the traditional 2-Wasserstein assess the round to round variation, tropicalizing the Euclidean norm with coordinate-wise range focuses on the largest single rank jump per ballot. The general downward trend persisted, with the curve also demonstrating some pronounced spikes when even a small bloc of voters had to shift a candidate several

ranks up after an elimination which did not always coincide with the peaks in the Euclidean metric. Tropical costs are systematically higher early (≈ 11.14 vs. 9.67 at $6 \to 7$ on actual ballots), then converge later (≈ 3.83 vs. 3.26 at $12 \to 13$). We also observe pronounced tropical spikes in the simulated skewed-left setting (Table 4.6.2, $6 \to 7$ is ≈ 10.89), consistent with a small bloc moving a preferred candidate several ranks after their leader is eliminated.When we tropicalize the weighted norm, we saw a similar pattern and realized it was the best diagnostic for high impact minority shifts. It could measure something like when a niche, but passionate constituency loses its leader.

At each elimination, the round–to–round distance quantifies how disruptive the step was with the largest value pinpointing the single elimination that most altered the ballot distribution (see the weighted panels in Tables 4.6.4–4.6.6). Because rank weights down-weight late positions, we treat $W_{2,w}^{\mathrm{eu}}$ as the main stability metric as it summarizes the steady decline across rounds that indicates smooth redistribution, while sharp peaks isolate the exact round that drove the biggest change. The tropical analogue $W_{2,w}^{\mathrm{tr}}$ adds specificity by reacting strongly when a concentrated bloc makes large rank jumps; when $W_{2,w}^{\mathrm{tr}}$ spikes more than $W_{2,w}^{\mathrm{eu}}$ at the same step, the shift is driven by a minority moving far rather than by diffuse, small adjustments. Election administrators and analysts can use these signals by reporting $\{W_{2,w}^{\mathrm{eu}}, W_{2,w}^{\mathrm{tr}}\}$ round, identifying the maximal round for audit or narrative explanation, and, where data permit, breaking down that round with groupwise distances $W_2(\mu_g^{\mathrm{pre}}, \mu_g^{\mathrm{post}})$ to attribute the change to precincts or demographics, which gives insight into how these factors impact RCV.

Simulations inside of our study could be improved, however, by training a model with real world RCV data and simulating a higher count of voters in order to generate both more accurate ballot choices as well and reduce variability. Furthermore, applying the exponential kernel $\omega_k = \exp -\gamma(k-1)$ could be optimized through cross validations on something like historical upset cases, which would make the weighting adaptive. Applying these simulations and calculations across longer round distances could also potentially reveal redistribution patterns not seen before.

As a whole, there are practical implications for both the Euclidean and Tropicalized calculations. One thing that I touched on and could be explored is embedding geographic, demographic, or partisan attributes into the transport plan to determine which bloc of voters are responsible for larger transport costs. This could be done through the conditional distance $W_2^{\mathrm{cond}} = \sum_{g \in G} \pi_g W_2\big(\mu_g^{\mathrm{pre}},\ \mu_g^{\mathrm{post}}\big)$ and then comparing $W_2\big(\mu_g^{\mathrm{pre}},\ \mu_g^{\mathrm{post}}\big)$ across g points at which blocs experience big rank shifts after a candidate is eliminated. This information could help campaigns choose where to invest persuasion or outreach between rounds which could help with not only the campaign, but also obtaining a better grasp on what the broader population desires. Furthermore, Election administrators can treat unusually high distances as audit triggers and focus on areas where the metric could signal an atypical ballot flow. In addition, policy makers and journalists can utilize this as a scale-free way to compare redistribution changes across jurisdictions with quantifiable quantities rather than anecdotes.

REFERENCES

Erwan Brugallé, Ilia Itenberg, Grigory Mikhalkin, and Kristin Shaw. Brief introduction to tropical geometry. In *Proceedings of the 21st Gökova Geometry-Topology Conference*, Gökova, Turkey, May 2014. Lecture notes for a mini-course at the Gökova Geometry & Topology Conference.

Yuhang Cai and Lek-Heng Lim. Distances between probability distributions of different dimensions. *IEEE Transactions on Information Theory*, 68(6):4020–4031, 2022. https://doi.org/10.1109/TIT.2022.3148923.

Anthea Monod, Bo Lin, Ruriko Yoshida, and Qiwen Kang. Tropical geometry of phylogenetic tree space: a statistical perspective, 2018. URL https://arxiv.org/abs/1805.12400. arXiv preprint.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*, volume 11 of *Foundations and Trends in Machine Learning*. Now Publishers, 2019. https://doi.org/10.1561/2200000073.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015. https://doi.org/10.1007/978-3-319-20892-2.

Roan Talbut, Daniele Tramontano, Yueqi Cao, Mathias Drton, and Anthea Monod. Probability Metrics for Tropical Spaces of Different Dimensions, 2024. URL https://arxiv.org/abs/2011.00629v3. arXiv preprint.

Predicting Residential Property Values in San Marino Using a Multivariable Regression Model

Zhiying (Michelle) Yin

Abstract:

This paper describes the development of a robust statistical tool to predict present and future values of real and hypothetical single-family homes in San Marino, CA. The model is readily duplicatable and generalizable for application to other cities and regions. To create a tool of maximal accuracy and reliability, the data used consisted of every single-family home sale in the city of San Marino, CA, from January 1st, 2022, to September 1st, 2025, as archived by the California Realtors Multiple Listing Service (California Regional Multiple Listing Service). The method of analysis was a multiple regression of a property value (the response variable) against the explanatory variables 'living area,' 'lot area,' 'number of bedrooms,' 'number of bathrooms,' 'time of sale,' 'year built,' and 'distance from Langham Resort.' After the model revealed that only the variables 'living area,' 'lot area,' 'number of bathrooms,' 'distance from Langham Resort,' and 'time of sale' were statistically significant, the remaining variables were eliminated one by one, producing the final model.

Most notably, the year the house was built and the number of bedrooms in the house were found to be statistically insignificant by the model. A novel feature of the model was its ability to capture a time trend, which enabled it to predict both future and current home values.

Introduction:

The valuation of residential property is a central question for homeowners, buyers, developers, and investors. Decisions such as whether to sell a home, renovate it, expand it, or purchase a new one depend on accurate expectations of market value. However, determining the present value of a property—and especially predicting its value in the near future—is a complex process.

For example, homeowners often struggle with questions like:
"Should I sell my house now or should I wait and sell it in a year?"
"Should I build an addition to my house and increase its size?"
"If I buy this house now, how much will it be worth 5 years from now?"

Investors and developers may wonder:
"Should I renovate this house or should I demolish and rebuild it?"
"What kind of house would be most profitable to build on a particular vacant lot?"
"What kind of real estate investment project in San Marino will bring me the highest rate of return?"

For a prospective home buyer, this tool can help answer the question, “Should I buy a house in San Marino or should I buy a house in a different city?”

To understand why these decisions are difficult, it is necessary to examine how real estate valuation is currently performed.

The task of determining a house's value – past, present, or future – is surprisingly complex. By definition, the market value of a property is the price a willing buyer and a willing seller would agree to, and, therefore, at which the sale of the property would actually transpire. So, if a particular house has not undergone a sale recently, how do we truly know what a willing buyer would actually pay for it, or what the owner would need to accept in order to part with it?

The formal evaluation of a real estate property's market value is called an appraisal. Official appraisals are conducted by licensed real estate professionals, commonly referred to as real estate appraisers. Those appraisers generally adhere to accepted methods and principles when determining the value of a house. In particular, there are three appraisal approaches:

1. The comparable sales approach
2. The capitalization rate approach
3. The cost-of-replacement approach

Among these, the comparable sales approach is most commonly used to determine the market value of owner-occupied residential real estate. The real estate appraiser chooses three to six “comparable properties” which have been sold relatively recently. The best choices are considered properties that are:

1) close to the property being appraised
2) in the same neighborhood/development as the property
3) of similar size (living area and lot area) to the property
4) equipped with similar features to the property.

Inevitably, no two houses are identical, so even the best comparable sales provide a value that requires adjustment. The real estate appraiser makes those adjustments based on estimated value differences due to variations in living area, lot area, interior finishes, landscaping, whether the house has a pool, solar system, etc.

Other approaches apply in different circumstances. The capitalization approach is typically used to evaluate commercial real estate, as well as income-generating residential real estate, including apartment buildings, rental homes, and others. The real estate appraiser considers the gross income and the net income the property has generated in the past two years and, using a multiplier called capitalization rate, calculates what the market value of the property should be. The cost of replacement approach is used to evaluate real property for which the comparable sales approach and the capitalization approach are impractical or impossible, such as churches, schools, and museums. Since sales of those properties are exceedingly rare, finding recent

sales of comparable properties in close proximity to the one in question is impossible. Furthermore, these are not income-generating properties, so the capitalization approach is unworkable. Therefore, the appraiser attempts to calculate the cost of rebuilding the identical structure from the ground up, based on local labor and material costs, and adds that value to the land's value as part of the property (Khaled, 2019).

At first glance, the comparable sales approach appears to be sensible, as recent actual sale values are a strong predictor of future sale values for similar properties. However, this method has several structural limitations that make the resulting estimate less transparent and less generalizable. First, it is marked with an element of subjectivity–the real estate appraiser chooses the properties, which, in his opinion, are the most similar to the property being appraised. However, there are so many similarity criteria that, especially in an active real estate market, the choice of comparables becomes almost arbitrary. Indeed, when two real estate appraisers evaluate the same home at about the same time, they rarely choose the same comparables. Second, it is ultimately designed to estimate present market value and does not provide a direct way to project future values. Third, the appraiser needs to make multiple adjustments to each comparable's sale price based on the extra value associated with:

1) the additional living square footage of the bigger of the two houses
2) The additional land area of the bigger lot
3) the difference in safety and ambient noise due to the different locations (e.g., a house in a cul-de-sac vs a corner lot at the intersection of two streets)
4) The difference in the presence and quality of features such as swimming pool, kitchen appliances, floor finishings, bathroom fixtures, and so on.

Standardized values are assigned to the appraiser for every house. However, the reliability of those numbers for each individual property is highly questionable (Cupal, 2014).

These limitations do not mean the comparable sales approach is ineffective, but rather that it is designed to answer a narrower question: "What is the market value of this property today?" It is not designed to:

- Incorporate the full set of available sales data,
- Project how the value of the property may change in the future.
- Or show how much each characteristic of a home contributes to its price,

Meanwhile, widely used automated valuation tools, such as Zillow and Redfin, do incorporate statistical modeling and time trends; however, their methodologies are proprietary. The user can view the output, but not the model's structure, underlying coefficients, or the assumptions guiding the estimate. As a result, homeowners and investors receive a value but cannot apply the method themselves.

The unavailability of a transparent, verifiable, and easily applicable model for present- and future-value estimates of homes, both real and hypothetical, inspired the development of our model. The tool we developed can be used to calculate a predicted value of any house in San Marino, CA, at any time in the near future, as long as select basic parameters of the property are known.

Method:

To develop this calculating tool, we first consider the parameters that most significantly affect the value of a single-family home.

These variables—the explanatory variables in our regression model—are:
1) living area of the house (in square feet)
2) size of the property (area of the land, in square feet)
3) number of bedrooms in the house
4) number of bathrooms in the house
5) year the house was built (using 1900 as year ‘zero’)
6) when the last sale of the house happened, the value of which we use as the value of the property at that moment (measured as the number of months since January 2024)
7) distance of the house from the Langham Huntington Resort (measured in miles)

Additionally, we utilize data from the Multiple Listing Service, which includes every house sale in the City of San Marino from January 1, 2022, to September 1, 2025.

The response variable in our regression model is the market value of a particular house at a particular point in time.

For convenience and ease of recognition, we label our variables are follows:

A = Living Area
L = Land/Lot Size
R = # of rooms (bedrooms)
B = # of bathrooms
Y = year house was built
T = time of sale
D = distance from Langham resort
P = market value of the property

An excerpt of the data set is as follows:

P	A	L	R	B	Y	S	D
4,320,000	3,550	12,150	4	4	29	21	1.4

2,400,000	2,000	10,000	3	2	30	21	3.0
1,450,000	1,270	5,600	2	1	25	21	0.9
2,638,000	2,375	13,422	3	3	42	21	1.4
3,110,000	2,760	13,207	4	4	63	20.5	1.1

Result:

Hence, our initial regression model has the form:

P = (Co) + (C1)*A + (C2)*L+ (C3)*R + (C4)*B + (C5)*Y + (C6)*T + (C7)*D,

where Co, C1, … C7 are the coefficients of the variables in the regression, as calculated by Excel's multiple regression software.

In addition to the coefficients, the software provides us with an Analysis of Variance (ANOVA) table, in which it calculates p-values and 95% confidence intervals for each of the coefficients.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.954811
R Square	0.911664
Adjusted R Square	0.91017
Standard Error	638116.4
Observations	422

ANOVA

	df	SS	MS	F	Significance F
Regression	7	1.74E+15	2.49E+14	610.3759	1.068E-213
Residual	414	1.69E+14	4.07E+11		
Total	421	1.91E+15			

	Coefficients	Std Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	41062.31	164311.1	0.249906	0.802784	-281925.72	364050.3	-281925.72	364050.34
Living Area(sqrt)	467.3667	54.62781	8.555472	2.31E-16	359.98426	574.7492	359.98426	574.74921
Lot Area(sqrt)	85.99967	3.343664	25.72019	7.44E-88	79.426992	92.57235	79.426992	92.572345
# of Bedrooms	-54219.7	51765.23	-1.04742	0.295519	-155975.2	47535.71	-155975.2	47535.708
# of Bathrooms	228382	44570.28	5.124086	4.6E-07	140769.696	315994.3	140769.696	315994.26
Year Built(1900 as "zero")	259.3775	1670.47	0.155272	0.876682	-3024.2828	3543.038	-3024.2828	3543.0377
Sale Date(mo. after Jan 1st, 2024)	7400.348	2258.616	3.276497	0.00114	2960.56284	11840.13	2960.56284	11840.134
Distance from Langham (mi)	-90316	36090.49	-2.50249	0.012717	-161259.43	-19372.5	-161259.43	-19372.509

A p-value of less than 0.05 and a confidence interval that does not contain 0 indicate that the variable is statistically significant. A p-value of more than 0.05 and a confidence interval containing 0 mean that we do not have strong enough evidence (95% certainty) that the variable is significant. If the p-value of a coefficient is far above 0.05, there is a good chance that that coefficient's variable is insignificant and may be excluded from the model. Therefore, we rerun the model, removing the insignificant variables one at a time, starting with the variable whose coefficient has the highest p-value.

In the original ANOVA table, we see that the coefficients of two of our variables have p-values greater than 0.05: the "number of bedrooms" variable with p = 0.30 and the "year built" variable with p = 0.88. Hence, first, we remove the "year built" variable from our model and rerun the regression. All the p-values that were less than 0.05 remain so; our p-value for the number of bedrooms is still far greater than 0.05, so we remove this variable as well and run the regression with just the remaining five variables. Our final model is:

P = -56,345 + 445.476*A + 86.6384*L + 216,877.9*B + 7,355.94*T - 91333.8*D,

The ANOVA table is as follows:

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.9546853
R Square	0.9114241
Adjusted R Square	0.9103595
Standard Error	637443.08
Observations	422

ANOVA

	df	SS	MS	F	Significance F
Regression	5	1.73933E+15	3.48E+14	856.1076	2.1821E-216
Residual	416	1.69035E+14	4.06E+11		
Total	421	1.90836E+15			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-56344.953	123188.3315	-0.45739	0.64763	-298494.15	185804.2	-298494.15	185804.2447
Living Area(sqrt)	445.47594	50.34863644	8.847825	2.58E-17	346.5064895	544.4454	346.506489	544.4453973
Lot Area(sqrt)	86.638464	3.279557288	26.41773	5.26E-91	80.19189418	93.08503	80.1918942	93.08503357
# of Bathrooms	216877.92	41351.43513	5.244749	2.5E-07	135594.1059	298161.7	135594.106	298161.7241
Sale Date(# of months after Jan 1st, 2024)	7355.9392	2255.824752	3.260865	0.001202	2921.703014	11790.18	2921.70301	11790.17532
Distance from Langham (mi)	-91333.814	35789.64223	-2.55196	0.011068	-161684.902	-20982.7	-161684.9	-20982.7263

For example, if an investor purchased a 25,000 square-foot lot, 2 miles away from the Langham Huntington Resort in San Marino, CA, and he took three years to build a 5,000 square-foot house with 7 bathrooms, our model predicts that he can expect the value of the property to be

P = -56,345 + 445.476*5,000 + 86.6384*25,000 + 216,877.9*7 + 7,355.94*60 - 91,333.8*2
P = $6,113,829.

Interpretation of the coefficients:

The meaning of our model is as follows:

1) On average, each additional square foot of living space increases the house value by approximately $445.
2) On average, each additional square foot of land increases the property value by approximately $87.
3) On average, each additional bathroom increases the home's value by approximately $216,878.
4) On average, a property appreciates by approximately $7,356 every month.
5) On average, each mile further away from the Langham Resort (the most expensive part of San Marino) decreases a property's value by approximately $91,334.

Discussion:

Advantages of our Model:

The idea of using qualitative and quantitative parameters of a property to estimate its market value is not novel. However, the tool we have developed has several advantages over the traditional and ubiquitous comparable sales approach and tools like zillow, whose specific methodology is proprietary and, therefore, not discoverable.

First, our tool removes the element of subjectivity. Essentially, an appraiser attempts to make an estimate based on the "best" three available data points, but why use three points if it is possible to use every available point? Our multivariable approach uses every completed single family home sales in the city over a multi-year period, which incorporates the full structure of the market rather than relying on a small number of selected comparables.

The second major advantage of our tool is the ability to not only determine the present value but also project an extrapolated future value of a real or hypothetical home. When an appraiser furnishes a report, it only provides a point estimate of the home's value at that moment; it offers no mechanism for extrapolating a value at any future time. In fact, we are not aware of any public model that includes a "time" variable the way our model does. Appraisal reports specifically state that the appraisal is only to be considered reliable for thirty days from its issuance, as home values continuously change over time. Using "month of sale" as one of our independent variables, together with the home sales over a four-year period allows our model to include a time trend. Therefore, the tool can be used to determine the value of a home at any time in the near future as easily as it is used to determine the value of a home today.

Another advantage of our tool is that it offers inherent flexibility. It is easily updatable: as new sales provide new data points, the model can immediately detect changes and trends in the local real estate market. Furthermore, the model can be adapted to different regions and produce distinct conclusions in various marketplaces simply by substituting the corresponding local sales data. This allows any researcher or homeowner with access to the same type of data to generate a personalized, evidence-based valuation. Variables that were statistically insignificant in San Marino may be significant elsewhere, depending on local demographics, consumer preferences, and other market-specific factors. For example, Nguyen's model found the number of bedrooms in a house in Vietnam to be statistically significant and the number of bathrooms insignificant, while our model found the reverse (Nguyen, 2020). A possible explanation for those findings is that in a marketplace with large average family sizes, like Vietnam, the number of bedrooms would be expected to have an outsized importance; on the other hand, in a marketplace dominated by small families and empty nesters, like San Marino, CA, the importance of the number of bedrooms wanes.

The ready availability of our specific source data, along with our methodology, makes our model easily verifiable, duplicable, and customizable: another researcher can retrace our steps and confirm the validity of our conclusions. Additionally, they could utilize our methodology to develop a similar model for any local real estate market nationwide.

Limitations of our model:

Although our model represents a significant improvement over traditional real estate appraisal methods, it is not without its limitations. In fact, it is based on a handful of assumptions, which are reasonable but not entirely accurate.

The first assumption is the linearity of the variables considered. In other words, the model assumes that, regardless of the value of a particular variable, a unit increase in that variable will result in the same effect on the market value. This assumption is probably most reasonable for the variable "distance from Langham" due to its limited domain (all of San Marino is within 4 miles of the Resort) and natural linearity. On the other hand, we consider the living area: if a 1,000-square-foot home is expanded by 500 square feet, the living area increases by 50%; however, if a 5,000-square-foot house is enlarged by 500 square feet, the living area increases by only 10%. It appears reasonable to suspect that the increase in value is proportional to the percent increase in living area rather than the amount of increase in living area. If that is the case, then a logarithmic relationship may be more appropriate than a linear one. In addition, according to economic models, utility functions feature diminishing returns and are, therefore, concave down (the more of a good you have, the less additional satisfaction an extra unit of that good brings you). An analogous analysis follows for the variables "lot area" and "number of bathrooms." The time variable is also susceptible to the flaw of the linearity assumption, but for a different reason: home values (like most values in economics) tend to appreciate by certain

percentages over fixed time periods, rather than by fixed amounts. If that is the case, then an exponential relationship may be more appropriate for the time variable. Notwithstanding, even if the variables are not linear, the model provides a linear approximation of their effect, which may be very accurate if the curves' concavity is not very large (the absolute values of the second derivatives are small). Still, this assumption makes our model better for houses with parameter values close to the average and less reliable for houses at the extremes. For example, the estimate would generally be better for an average-size house on an average-size lot than for an abnormally small or an abnormally large house.

Another assumption inherent in our multiple regression model is independence of the explanatory variables–the variables are assumed to have little to no correlation. This assumption is most accurate for the variables "time" and "distance from Langham," since there is no obvious way they can affect (or be affected) by the values of the other variables. However, at least some correlation between "living area," "lot area," and "number of bathrooms" is rather apparent: houses with many bathrooms tend to have larger living areas and be built on larger lots. However, the extremely low p-values of all of these variables lend credence to the assertion that their independent effect on the property value is still statistically significant. Indeed, a 5,000-square-foot house with only 2 bathrooms would never be worth as much as a 5,000-square-foot house with 6 bathrooms. Conversely, a 5,000-square-foot house with 6 bathrooms would be worth much more than a 1,000-square-foot house with 6 bathrooms. Hence, ignoring the correlation between these variables presents a much smaller challenge than eliminating one of them would.

A third potential pitfall to which a multiple regression is vulnerable is overfitting–having too many explanatory variables for the size of the data set. We believe that we have successfully mitigated this problem by 1) having a large data set, and 2) eliminating from our model the variables that were not statistically significant.

Another potential problem with a multiple regression model is the omission of important explanatory variables. Limiting our model to the small, homogenous city of San Marino, CA, and adding the "distance from Langham Resort" variable serves to address the most important variable in real estate: location. In addition, we hold that there is no variable outside our model that can significantly alter the value of a single-family home–any other variable is unlikely to affect a home's value by more than 1-3%.

Lastly, regression models are, generally very sensitive to outliers. However, none of the properties in our analysis contained parameters with highly anomalous values relative to the rest of the data set. In addition, the large size of the data set diminishes the potential effect of any outliers (Osborne, 2002).

Conclusion:

Our model would be best used to estimate the value of single-family San Marino homes of average size in the relatively near future. Extrapolating values in the distant future or those of houses that are abnormally large or small is feasible but larger potential errors should be anticipated.

Other researchers can explore a variety of ways to enhance the model. First, the linearity assumption can be relaxed and the model can be adjusted to include some non-linear relationships for some of the variables, if they test better. In particular, some of the first adjustments to explore include: 1) an exponential relationship to the time variable; 2) a logarithmic relationship to the variables “living area” and “lot area”; 3) various non-linear relationships to the other variables.

A further enhancement would be to develop a method to test the model: the data can be split into two sets: a training set (with about 80% of the data points) and a testing set (with the remaining 20% of the data). The training set can be used for the regression (train the model), then the testing set would be used to determine the model’s predictive performance–for those points, the value predicted by the model would be compared to the value observed, and the root-mean-square error calculated. Furthermore, the R^2 values for the training data set and the testing data set should be compared. A substantially smaller value for the test set would imply possible overfitting. Alternatively, a k-cross validation can be performed to confirm that the model is not only descriptive but also predictive (Shmueli, 2011).

Further research should also be performed to determine the validity of the independence assumption. The explanatory variables should be tested for intercorrelation: if a high correlation is found between any of the variables, the model should be adjusted to incorporate that finding.

Lastly, our model considered only what we believed are the primary variables determining the value of the home; we omitted variables that we consider unlikely to affect the value of the home by more than 1%. An enhanced model that incorporates variables with smaller but consistent effects would be more accurate and complete. Such variables include the home’s proximity to parks, daycare/preschool, K-12 schools, restaurants, a grocery store, entertainment, and other commercial establishments; CC&Rs of the development, zoning, easements, and other restrictions; architectural style, privacy, and accessibility, and safety (fire, flood, mudslide hazards). For a model of a larger geographic area, proximity to a freeway and an airport and median household income of the area may be important as well (Lowrance, 2015).

Sources

[1] *Home*.
California Regional Multiple Listing Service. (n.d.). https://go.crmls.org/

[2] *Consumer guide: The appraisal process*. National Association of REALTORS®. (2024a, December 10).
https://www.nar.realtor/the-facts/consumer-guide-the-appraisal-process

[3] Roy E. Lowrance (January, 2015),
Predicting the Market Value of Single-Family Residential Real Estate, 4.2-5.3.1 (62-104),

[4] Khaled, M. D. S. (n.d.).
(PDF) Real Estate Property Valuation Process: Valuation of a hypothetical property.
Real Estate Property Valuation Process: Valuation of a Hypothetical Property.(1-4)
https://www.researchgate.net/publication/344709078_Real_Estate_Property_Valuation_Process_Valuation_of_a_Hypothetical_Property

[5] Cupal, M. (n.d.). (PDF) the comparative approach theory for Real Estate Valuation.
(4.1-6.1,pg1-4)
https://www.researchgate.net/publication/260007864_The_Comparative_Approach_Theory_for_Real_Estate_Valuation

[6] cupal, martin. (n.d.). (PDF) the comparative approach theory for Real Estate Valuation.
(1-5.2) pg1-4
https://www.researchgate.net/publication/260007864_The_Comparative_Approach_Theory_for_Real_Estate_Valuation

[7] Osborne, J. W. (n.d.). (PDF) four assumptions of multiple regression that researchers should always test (1-4).
https://www.researchgate.net/publication/234616195_Four_Assumptions_of_Multiple_Regression_That_Researchers_Should_Always_Test

[8] Shmueli, G. (n.d.). (PDF) to explain or to predict?
https://www.researchgate.net/publication/48178170_To_Explain_or_to_Predict

FRANKLIN EDUCATION
FOUNDATION

Journal of Quantitative and Qualitative Research

2025 (Vol 3)