MSSP Final Portfolio

Aidan O'Hara

December 21, 2023

Table of contents

1 Investigation of Social Media Use by Oral and Maxillofacial Surgeon Firms

1.1 Introduction

Oral and Maxillofacial Surgeons (OMS) perform procedures on the mouth, jaw, face, and neck, addressing issues such as tooth extraction, jaw realignment, facial trauma, and tumor removal to improve both function and aesthetics, important for this analysis. Surgeons who maintain effective social media engagement may have better recruitment of new patients and connect with patients beyond the local audience, seeking a specialist. Social media is crucially important to a successful advertisement initiative, especially to reach younger aesthetically motivated audiences. To gauge the utility of social media to Oral and Maxillofacial Surgeons the consulting client surveyed 4621 members of the Association of Maxillofacial-Oral Surgery by email. 420 responses were received, 410 were used in this analysis omitting seven from Canada and three outliers.

1.2 Data and Methods

Data contains responses from OMS offices in the United States about advertising, social media, and newly recruited patients. Specific general information about each OMS firm gives the name and location of each office, number of surgeons, and newly recruited patients. Primary means of advertisement and a general advertising budget inform the OMS offices' status quo. Extensive features are collected to describe the application of social media: The use of a dedicated account, which platforms the office maintains an account on, the purpose of using social media, how many followers the office has, if the social media has paid promotion or a manager, frequency of posting, and the perceived impact on the business.

Missingness not at random is likely present because practices with dedicated staff, larger advertising budgets, and active social media are inherently more likely to respond to a survey about social media use than their counterparts. The measured impact of social media is probably an overestimation. Offices with dedicated media staff are especially likely to respond positively as drawing new patients to the practice is a measure of their job performance. Because we cannot control for missingness, not at random, the reader is cautioned against using these results for inferential purposes.

The construction of a Bayesian network informs another complication, While new patients were surveyed by "New patients primarily from social media in the past 12 months", follower counts were current at the time of the survey. Linear regression demonstrates a strong positive relationship but there is no way to verify from this sample if new patients incur higher follower counts or vice-versa.

Only basic modeling was conducted to further investigate the survey response feature relationships. Base ten log transformations were applied to account for the diverse magnitude of responses and maintain some response interpretability.

1.3 Results and Discussions

Including effects from each important aspect of a practice's social media efforts, no clear contribution is established about the effect of the number of accounts, dedicated social media staff, and the presence of a dedicated account. This is not to say these variables have no effect, more so that they are outweighed by the effect of budget size, the number of followers reported, and oddly the perceived impact social media has had on the practice.

The magnitude of OMS firms' budget, the magnitude of their total followers, and the perceived impact of social media on the business are shown to have a significant effect on the magnitude of each respondent's total newly recruited patients. Curious if there was any particular social media platform that incurred a higher incidence of new patients another model was constructed. The use of each social media platform is measured as a binary effect and the resulting coefficients dictate the average increase of new patients expected from the use of each specific platform. Facebook, YouTube, and Instagram, in order of most to least, are the only platforms with any significant effect and certainly positive impact.

1.4 Conclusion

Careful consideration of the data generation process has led to our previous statements about the missingness mechanisms present in this analysis and we would again caution the reader from drawing inferences that do not acknowledge this potential bias. However, conditioned on that bias it does appear that increased social media planning and usage is associated with an increase in the number of patients an office generates from social media.

2 Chicago Mosquitoes Compared with Community Features

2.1 Introduction

Summer's warm weather and rainy climate bring on the perfect opportunity for humanity's ultimate nemesis to make itself known: Mosquitoes. While highly adaptable, Mosquitoes still require specific circumstances to propagate, standing water sources, food sources, and other environmental factors. Using purpose-built devices placed at an assortment of locations in a given environment, Mosquitoes are regularly trapped, collected, counted, and tested for West Nile virus by Municipal or Academic institutions to better understand the potential impact on public health. This analysis inquires whether local median income and population density influence the number of trapped mosquitoes.

2.2 Data and Methods

The primary data set used in this analysis details trapped mosquitoes. Sourced from the City of Chicago originally and then found on Kaggle, the data contains records about the mosquitoes' response to a West Nile Virus test, and the location of each collected trap, for seasonal weeks between 2007 and 2022. The estimates of trapped Mosquitoes were spatially joined with median income and population totals from the American Community Survey about census tracts in the Chicago Urban area. The population density was calculated using the area computed from the included shape files.

The mosquito data is formatted such that each row reports the number of collected mosquitoes, of a specific species, where the sample was collected, and the result of the West Nile Virus test. To keep things straightforward concerning time, only data from 2015 was used.

Joining the two tables using their included geometry, each mosquito trap was associated with its specific Census Tract and better still, the ACS estimates about the Census Tract in 2015. Summarizing the resulting join by the Census Tracts, the result of the West Nile Virus test, and the week of each given observation results in a table that provides the total captured mosquitoes by week, by Census Tract, including test positivity.

The changes in the weekly collection of mosquito populations are well represented. The populations registered early in the season are small, before swelling and tapering off. Tracts with low populations have smaller capture totals, while Tracts with high populations have a higher incidence of larger collections.

2.3 Results and Discussions

Challenges encountered in the analysis led to serious doubts about the validity of our findings. Potential issues with Census tract geographies and the need for more nuanced incorporation

of time data are only the beginning. Furthermore, the discussion suggests a curiosity about a deeper exploration into the included Mosquito West Nile virus positivity and the integration of additional community features in future analyses.

2.4 Conclusion

Modeling factors influencing mosquito populations is a complex task, given the intricacies of geographic and temporal variations. Acknowledging these challenges, this analysis sets the stage for future investigations, emphasizing the need for refining methodologies and considering a broader range of community features to unravel the dynamics of mosquito populations.

3 Exome Classification with Different Racial Categories

3.1 Introduction

Exome classification is an important facet of genetic counseling. Exomes, sequences reflecting the protein-coding portion of a genome, can be classified as positive, negative, or indeterminate with respect to the exome's association with genetic diseases. Initial classifications may be inaccurate and potentially updated with another round of testing, per the request of a counselor. As is the case across the whole of the medical industry in the United States, racial disparities are very easily demonstrated. The client requested consulting to investigate the consequences of re-factoring racial categories into bins based on representation. Races with larger representation would be classified as 'A', moderately represented as 'B', and those with the least representation would be classified as 'C'. Mixed races would be classified as the combination of the appropriate representation class, 'AB', 'BC', 'AC', and 'ABC'. The analysis aims to compare the outcomes and rates of different features of exome analysis and classification for each mode of racial categorization.

3.2 Data and Methods

The client provided data detailing 10,000 original exome reports and 3,000 exome reanalysis reports. Each report includes the patient's race, sex, age, diagnostic indication rate, classification type, and two Boolean variables signaling if the report was a reanalysis and whether the report resulted in a reclassification.

During exploratory data analysis proportional comparison between different racial classifications was conducted, enumerating the proportional changes made by the new system. The new system produces mean reanalysis and reclassification rates across racial categorizations that have less variance.

To assess the actual impact of the categorization on those rates logistic and multinomial models using ABC and default race buckets as predictors were used to determine how an exome may be classified, whether an exome is reanalyzed or reclassified, how a reanalysis is initiated, and the evidence used to make a reclassification. Models included sex, age, and diagnostic indication rate as potential confounders.

To statistically evaluate the impact of each categorization on exome reporting, this analysis deployed the likelihood ratio test. The likelihood ratio test, in summary, compares how well two models fit the data by using the ratio of their likelihoods. In this case, there is some probability that a given exome will be evaluated in some fashion based on the age, sex, and indication of the patient based on the data. Calculating the probability of how an exome will be evaluated including the patient categorization, the test, assuming a null hypothesis that categorization does not affect exome evaluation, compares the estimated probabilities. If

the test statistic falls below the threshold the null hypothesis is rejected and categorization is significant to exome evaluation.

3.3 Results and Discussions

Application of the likelihood ratio test demonstrated an overall trend of decreased significance of racial categorizations when the client's new system was implemented. The new system, if implemented, may displace some racial disparities in exome reanalysis and reclassification rates and outcomes.

3.4 Conclusion

In conclusion, the analysis of genetic exome reclassification underscores the potential impact of redefining racial categories on the outcomes and rates of exome analysis. The client's proposed classification system, based on representation, reveals a noteworthy trend of decreased significance in racial categorizations. This suggests that implementing the new system could potentially mitigate some of the observed racial disparities in exome reanalysis and reclassification rates. While acknowledging the complexity of genetic counseling and the multifaceted nature of patient data, this study lays the groundwork for further considerations in refining classification systems to enhance the fairness and accuracy of genetic exome assessments. Future investigations could delve deeper into the nuances of racial categories' influence on exome evaluation and explore avenues for improving the overall equity of genetic counseling practices.

4 Unsupervised Short Document Text Analysis and Topic Modeling

4.1 Introduction

Text analysis involves the computational examination of language-based data to extract meaningful insights. Various techniques are employed to analyze and interpret large volumes of textual data, to understand numerous language-based features. The primary goal of text analysis is to unveil patterns and trends in language that enable more informed decisions, and actionable insights from vast amounts of unstructured data. Working with Data Scientists at Fidelity and other Master's Students at Boston University this project conducted text analysis on large sets of short text commentary in different fields. Unsupervised, the analysis used text analysis and statistical methods to impute the set of undeclared categories present.

4.2 Data and Methods

Spanning two semesters, this analysis explored two different short text document data sets. IMBD movie reviews, sans additional information, allowed the development of the team's methods on highly variable and expressive data. Early exploration and signals made easy to understand by familiar and comfortable media topics. Second semester the team applied methods to proprietary Fidelity Investments data with similar, short text documents. Clouded by technological and corporate jargon these documents required more intensive scrutiny and subject matter expert consultation to investigate.

In both cases, the text analysis roadmap was the same. Before anything could proceed the documents needed to be treated and cleaned of punctuation, extraneous symbols, and any other noise present. Stop words, frequently repeated helper words that do not infer specificity on their own, are removed. Tokenization proceeds to separate documents into their constituent pieces by a given rule. The team explored methods for multiple-word, n-gram tokenization, but ultimately proceeded with single-word tokens. Next on the road to categories, the term-frequency-inverse-document-frequency matrix was constructed. A TF-IDF matrix emphasizes the importance of words within documents while discounting common terms. This approach aids in discriminative feature selection, reduces the overall dimensionality of the task, and enhances tasks such as document clustering, and topic identification by highlighting contextually significant and distinctive terms among the documents.

Latent Dirichlet Allocation (LDA) is applied to the TF-IDF to compute the underlying topics present, constructing a topic probability distribution for each document. Clustering the LDA result provides document collections that may be further studied to infer meaningful topic themes by examining the most representative terms for each cluster and the predominant topics within each cluster.

4.3 Results and Discussions

The whole implementation of the previously described process did result in the described clusters of documents. IMBD movie reviews were clustered in such a way that the associated documents did share similar qualities. Perhaps the most difficult part of this type of analysis, naming the clusters of documents, was never fully explored. Armed with keywords or other search criteria, clusters may be selected and their documents reviewed for additional insights.

4.4 Conclusion

In conclusion, this text analysis endeavor successfully employed TF-IDF matrices and Latent Dirichlet Allocation (LDA) to discern patterns within diverse short text data sets. Through the application of these techniques on IMDB movie reviews and Fidelity Investments data, the team demonstrated the efficacy of discriminative feature selection and dimensionality reduction. Results from and methods for conducting this analysis were compiled and summarized in an R package delivered to Fidelity Investments at the end of the semester and revised to functionality by the author personally over the following summer. This project established a solid foundation for leveraging text analysis and topic modeling techniques to reveal intricate patterns within unstructured data sets across different domains.

5 United States Counties Minimum Carbon Conservation Selection

5.1 Introduction

How do we protect carbon biomass and ecological diversity within the constraints of a real-world budget? Conservation appears to be a simple activity. Find ecological features that should not be harmed, removed, or disturbed, create a policy to protect them, and celebrate; A gross oversimplification. Which ecological features should be preserved? Which ones are most valuable? Environmental valuation is very difficult, what is the cost of a tree? More still, conservation activities incur costs, to the population near the ecological feature and the municipal government tasked with maintaining the effort. In this light, large-scale conservation efforts are not only difficult to plan and coordinate but also expensive.

The continental United States is home to numerous environments and ecological features. Extensive forests have been long-standing ecological retreats and economic resources. The 50 states are disparate enough in size and population that environmental decisions are difficult to enact. The same problem is not as prevalent considering the counties that comprise those states. Counties vary wildly but are easily classified into two groups, urban and rural. Each type of county has a drastically different composition of populations, economic activities, and environmental resources. Assuming an approach to forest conservation that was directed by county governments, this analysis aims to select counties to maximize the protected biomass and minimize the needed costs of implementing a policy protecting the United State's Forests.

5.2 Data and Methods

The cost proxy of conservation for each county was created by combining the Headwater's Economics Rural Capacity Index, estimates of the total annual payroll of environmentally sensitive North American Industry Classification System (NAICS) industries, and US forest service carbon biomass estimates. The Rural Capacity Index, 1-100, represents the capacity of a given county. Headwaters calculated it using poverty indicators, educational levels, and economic features. In this analysis, it serves to enumerate communities that are more capable of enacting a conservation plan. Counties that have low scores have low capacity. This could appear as unstaffed municipal governments, economic strife, or other community maladies. The score is used in the inverse, points below 100 are treated as linearly increasing additional costs. 2021 annual payroll of NAICS industry groups by US county filtered by a list of environmentally sensitive NAICS industries. Aggregate annual payroll of environmentally sensitive industries in US counties will serve as a cost proxy for the economic impact of conservation actions. Estimated forest biomass in each county was estimated with zonal statistics using a raster available from the US Forest Service. The pixel records contained within each county were totaled to generate a gross biomass content.

Cost per county is constructed with the following definition:
Cost = Carbon Counted + Transactions Costs + Industrial Offset
Carbon Counted: biomass density * total forested area * price per mg of carbon
Transaction Costs: minimum transaction cost + (inverse rural capacity * rural capacity modifier)
Industrial Offset: year to pay industry * estimated annual payroll

Besides the cost proxy, the analysis would need to describe different species or environmental features that must be protected. This analysis considers 28 groups of forests described by a different IMG raster available at the US Forest Service. Proportions of forest type contained within each county were calculated and assuming a uniform distribution of carbon biomass within those forests a total biomass per forest type per county was computed. All these features are necessary to power the MARXAN algorithm used to make the optimal selection. MARXAN is a conservation planning algorithm used to identify optimal reserve networks for biodiversity protection. Assigning costs and conservation values to planning units, MARXAN employs a simulated annealing algorithm to iteratively optimize configurations that meet conservation targets while minimizing total costs. Four input tables were constructed. A planning unit file detailing the previously computed cost of selecting any given county. A species file detailing the target proportion of each species that should be selected, as well as a penalty if the algorithm were sufficiently constrained to a budget and unable to meet the goals. A cross table of planning units and species is constructed detailing the species contents of each planning unit. Finally, a boundary length table details the shared boundary of every planning unit, an important feature with an adjustment parameter that allows for selecting network solutions that have more or less spatially proximate selections.

5.3 Results and Discussions

After selecting agreeable model parameters the MARXAN algorithm produces two important tables. Most prominent is the best selection the algorithm made, a binary column detailing which planning units composed the cheapest solution that met the conservation goal. Without such a proximate cost estimation a designated budget seemed frivolous and the MARXAN algorithm was run exclusively to meet the conservation goal with the minimum cost. Beyond the best selection, MARXAN elaborates on the selection process by compiling the selection frequency of each planning unit. This result allows for the inspection of which planning units MARXAN most preferred and which planning units were less distinctly cost-effective.

The results detail areas with forests that may be cheaper to conserve, namely in the Pacific Northwest, the central range of the Rocky Mountains, and numerous regions on the eastern side of the United States. An important caveat to this result, the variance of size and frequency of counties East-West in the United States do not allow MARXAN the same flexibility in protecting the forests therein.

5.4 Conclusion

The MARXAN algorithm's best selection of counties for forest conservation totals a modest $48 billion. Not even a tenth of the annual defense budget, this result underscores the potential feasibility of conservation actions. However, it also raises questions about the accuracy of cost estimations and the estimated potential economic impact of conservation efforts, suggesting a need for a more comprehensive modeling of the true costs involved. Nonetheless, this analysis illuminates regions in the United States where conservation may be economically favorable, offering some insights for decision-makers striving for environmental preservation.