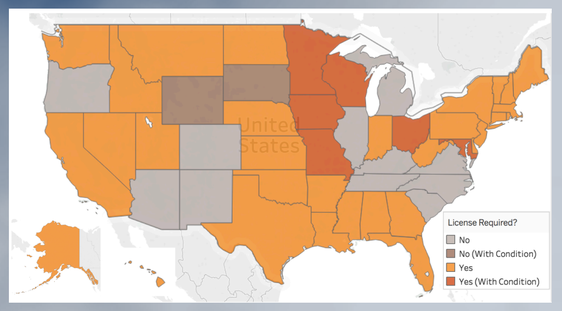
Open Data for Tobacco Retail Mapping

Felicia Chen, Nikhil Pulimood, James Wang | Project Manager: **Mike Dolan Fliss**

Introduction

There is no national database of tobacco retailers.

- Only 37 states require licenses for retailers to sell tobacco. Among those that do, licensing records can be as simple as handwritten records.
- Tobacco products consist of 36% of sales revenue in convenience stores.
- There are weak incentives to obtain proper licensing, with the typical penalty being a monetary fine of $5000.

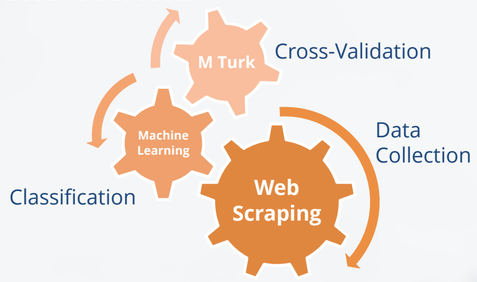


Tobacco Retailer Licensing in the United States

But having the knowledge of tobacco retailers' location is important.

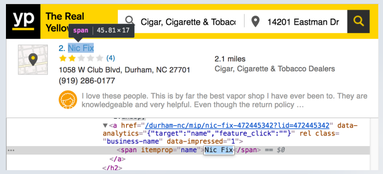
- Youth are more likely to begin smoking in areas with lots of tobacco retailers.
- There is a strong co-variate relationship between density of tobacco retailers and many indicators of social disadvantage, including lack of healthcare coverage.
- Regulations on tobacco sales are often under enforced.

Method Overview



Web Scraping

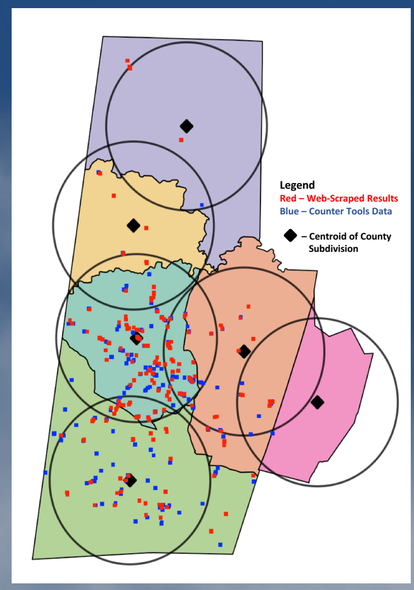
In order to efficiently obtain a list of tobacco retailers, we looked to scrape data from webpages.



HTML script from Yellow Pages

Used R to code an automated web crawler that parses HTML script

- Collected basic store information from Yellow Pages such as the store name, address, and phone number
- Inputs as the following:
  - Two search term keywords: "Convenience Stores Gas Station Cigarette" and "Vape E-Cig Hookah"
  - Search address was the location of the centroids of County Subdivisions
- Sorted results by distance and limited the maximum number of pages to four in order to minimize duplicates

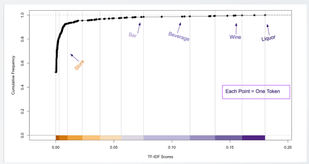


Web-Scraped Stores vs. Counter Tools Dataset in Durham County
Mapping our results and comparing to the Counter Tools dataset.

Machine Learning

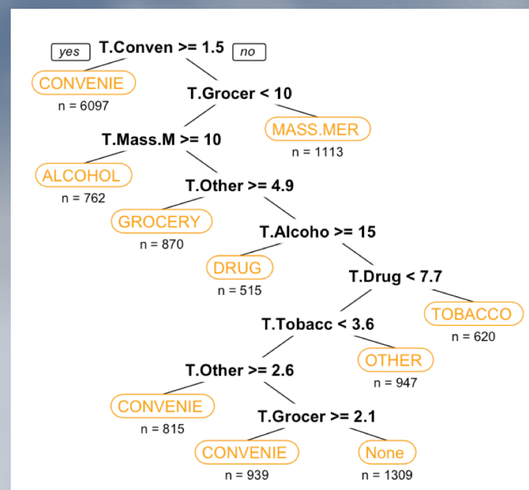
Our aggregated dataset contains many retailers.

But not all may actually sell tobacco products. The next step was predicting such characteristics of a store.



Jenks Natural Breaks for StoreType - Alcohol
Using Jenks classification to cluster our modified tf-idf scores, in order to select cut-offs for tokens that predict StoreType – Alcohol.

- From the Counter Tools dataset, each retailer is listed as being in one of nine categories.
- Tokenized store names by breaking them down into n-grams. Calculated a modified version of the term frequency–inverse document frequency (tf-idf) score for each n-gram within each category.
- Used Jenks Natural Breaks to cluster tokens with similar scores together, and to determine which tokens were the best predictors for a store being in each category.
- Modeled a decision tree through R, where are training set was 70% of our data and our test set the other 30%.
- Predicted categories of retailers and other factors, such as whether tobacco and alcohol are sold in the store.



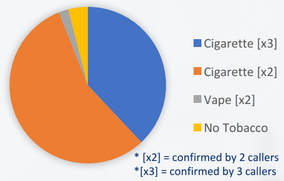
Decision Tree for Store Type Classification
The decision tree model classifies store categories based on the modified tf-idf scores of each store name.

Amazon Mechanical Turk (M Turk)

An optimal way of cross-validating our information

Created two HITs (Human Intelligence Tasks) for the public to complete.

- Call a potential tobacco retailer to find out if the phone number exists, and if the retailer sold various tobacco products.
- Asked people to search online for the coordinates of a potential tobacco retailer.



* [x2] = confirmed by 2 callers
* [x3] = confirmed by 3 callers

Pie Chart for Stores that Sell Tobacco Products
The percentage of stores that is confirmed by M Turk callers to sell cigarettes, vapes, and other tobacco products.

Conclusion

- **Web-scraping** is the most effective method of data collection. It is both cheaper and quicker to implement than traditional methods, such as ground truthing.
  - The legality of scraping is a grey area. Determining the acceptable usage of APIs (Application Program Interface), caching and storing data, and displaying results requires searching for precedent from a multitude of law cases.
- **Machine learning** with text mining is a relatively precise method for classification. Limited factors are a non-issue, and we can easily modify decision trees to determine which characteristics to include. Missing data, however, may be a potential issue.
- **M Turk** is cost-effective for human cross-validation. The median call-task takes around nine minutes, and the geocode-task about two minutes. It only costs $0.10-0.25 per HIT.

Results

- Aggregated 15,502 unique retailers in North Carolina, and 266 unique retailers in Durham County through web-scraping.
- Found that all 266 retailers matched the dataset of a community partner.
- Created and trained a decision tree using 19,619 retailers that were not in North Carolina, to predict the store types of 363 North Carolina retailers with an accuracy of 85.15%.

Original Coding on Store Visit	n stores	% accurately coded by text-mining machine learning methods: Alcohol	Convenience	Drug	Grocery	Hookah	Mass Merch.	None	Other	Tobacco	Vape	
Alcohol	384	50%	30%		4%			8%	7%			Correct
Convenience	2,818	0%	**93%**		1%		1%	2%	2%			Incorrect
Drug	239		4%	82%				14%				% Predicted
Grocery	600	1%	**40%**		**49%**		3%	3%	4%			
Hookah	2		50%					50%				
Mass Merch.	395		10%		1%		88%	1%				
None	1,020	3%	57%	2%	4%		3%	26%	6%			
Other	384	2%	47%		3%		3%	28%	18%			
Tobacco	138	1%	43%					50%	6%			
Vape	15		27%					73%				

Ex. Out of the 384 Alcohol stores, 50% is correctly predicted as Alcohol, while 30% is wrongly predicted as Convenience Store

Confusion Matrix for Store Type Prediction
Convenience stores had the highest accuracy rate of 93%. Tobacco and Vape stores could not be effectively predicted from store names.

Other Applications

Item	Web Scraping	Machine Learning	MTurk
Tobacco	All relevant stores	Classify store types using store names via text analysis	Cross-validate if a store sells tobacco
Produce	Stores that sell organic produce/accept SNAP	Classify farmer markets, co-ops, grocery stores	Validating SNAP availability and food freshness
Overdoses	Surrounding retailers and establishments	Classify to predict areas that may be prone to incidents	

Acknowledgments We are thankful for the funding and support of Counter Tools, Bass Connections, and Duke Data+. We would also like to thank Dr. Paul Bendich, Dr. Ashlee Valente, Ariel Dawn, and Kathy Peterson. We are especially grateful for the guidance and leadership of our project manager: Mike Dolan-Fliss.