Introduction

Open Storefront Directory is the first longitudinal database of retail businesses in New York City built on publicly available data. The directory is 'open' in the sense that it captures businesses that have been open for business in NYC in the past decade. But it is also 'open' in another sense: the database is built on open data, made accessible by city and state regulatory agencies. Our ambition is to use this database, which is built on public data, to study public issues related to retail businesses.

A brief description of the process by which we construct the database follows. First, we collect government records pertaining to NYC retail businesses, including operating licenses and inspection results. Then, we extract dates, addresses and other identifying information from the records and link information referring to the same entity. Finally, we estimate the start date, end date, and industry for each business. The result is a database of NYC retail businesses covering 2010 to the present.

Motivation

Storefront vacancy in New York has received considerable attention in the press. Blogs, newspapers, and magazines have, for many years, gone back and forth on the implications of the phenomenon.[1] For some, widespread commercial vacancy constitutes a crisis, symptomatic of the city's increasing inhospitality to small businesses. For others, shuttered stores are natural, the inevitable product of a dynamic urban economy. Besides disagreement over the consequences of commercial vacancy, there exists a more elemental disagreement over the precise nature of the phenomenon. The question of whether vacancy has, in fact, risen in the past two decades is unresolved.

Two city agencies have conducted major studies to provide tentative answers to this question. The Department of City Planning found in August 2019 that although storefront vacancy had increased in some commercial corridors between 2007 and 2017, there was no consistent trend throughout the city (see Carling et al. [2019]). A month later, the City Comptroller reached the opposite conclusion, claiming that retail vacancy had increased in the same period (see McWilliam et al. [2019]).[2]

Importantly, each study used data that was limited in scope. The Department of City Planning's data was limited to a handful of commercial corridors, and was observed for two points in time. The Comptroller, meanwhile, estimated that their data was missing about 25% of commercial properties.[3]

While the magnitude of the problem remains obscured by the lack of dispositive research, the pandemic has evidently altered the retail landscape, triggering an unprecedented wave of retail business closures.[4] As a result, accurate data on retail businesses is at once more needed and harder to construct. More needed, because the effect of the pandemic must be disambiguated from secular changes to the retail landscape. Harder to construct, because the pandemic has resulted in simultaneous widespread closures of businesses. In order to understand how city or state policy can intervene effectively, we can't just rely on cross-sectional data.

Current Storefront Data

To understand how the retail business landscape has changed, we need data that goes back many years, is updated frequently, is accessible, and has information relevant to vacancy. Four companies - Data Axle, LiveXYZ, SafeGraph, and Dun & Bradstreet, - appear to meet at least some of those requirements. While these companies keep their data acquisition methods confidential, we have been able to discern some aspects of the process.

Data Axle has millions of U.S. businesses going back to 1997. To verify its information, it tries to call every business at least once a year. Because of this, the data is updated infrequently.[5]

LiveXYZ, which came out in 2018, meanwhile, mapped storefronts by walking every block in New York. They now update their business information by relying on help from their neighborhood partners. Importantly, LiveXYZ tracks storefronts, as opposed to businesses, and as a result can indicate when a storefront is vacant.[6]

SafeGraph, which released its first dataset in 2018, seems to have a mostly automated process of acquiring data, using web crawlers and third-party sources.[7]

Finally, we have access to the Dun and Bradstreet database through Harvard's library, though we are not given any information about how D & B collects data. Dun Bradstreet has records for New York City going back to the 80s.

While there exist other retail business datasets which, anecdotally, are more comprehensive than those mentioned above (e.g. Google and Foursquare), we do not have access to them.

Our Data

In 2019, the data scientist Lindsay Poirier published a blogpost on the website of BetaNYC (a civic technology non-profit) demonstrating that barbershop licenses, which include the name, address, date of license issuance, and date of license expiration, could be used to create a longitudinal directory of barbershops in NYC (see Poirier [2018]).

We extend this paradigm by acquiring all available regulatory records, including everything from liquor licenses to health inspections. Unlike LiveXYZ, Dun & Bradstreet, DataAxle, or SafeGraph, these records are accessible, go back many years, and per state law, have to be updated regularly. Importantly, regulatory records can be used to keep track of retail businesses over time at no cost. Instead of hiring someone to verify a restaurant's existence, we can let a health inspector verify it for us.

Methodology

Our first task is to acquire the records. To accomplish this, we gather data in three ways: querying public data APIs (such as those on Open Data NYC and its state counterpart, NY Open Data), scraping government databases, and filing Freedom of Information Law (FOIL) Requests with state and local agencies.

Once the records are assembled, we perform basic cleansing, deleting unnecessary columns, type casting data, and filtering out records by zip code or postal city. In the final stage of preprocessing, we assign each record to a tax lot using the Geoclient API. This City API takes an address and returns a ten-digit BBL (uniquely identifying each tax lot), as well as the longitude and latitude. BBLs are assigned by the New York City Department of Finance for the purpose of assessing property taxes, but they are also commonly used to aggregate observations to the building-level, as done here.

Data cleansing is followed by record linkage - the process of determining which records pertain to the same latent entity. For more on the record linkage process, see Appendix A. Through this process, we assign a unique identifier, called a 'Location ID' or 'LID', to all of the records that pertain to the same business entity at one particular location. LIDs that share a record identifier (assigned by the agencies) are given the same 'Business ID', or 'BID'.

While records often contain some indication of the regulated entity's industry, this information is not standardized across departments or agencies, and is consequently not in a usable format. Our challenge then is to convert variegated departmental classifications into a standardized industry code, specifically the North American Industry Classification Scheme (NAICS), the standard used by Federal statistical agencies in classifying business establishments.[8] We accomplish this by using a BERT language model, as in Devlin et al. [2018], specified for sentence and paragraph-length embeddings called S-BERT.[9] Briefly, we concatenate the available departmental classifications (e.g. 'Restaurant-Indian') for each BID into a string, embed it using an S-BERT model pre-trained on a large corpus of natural language, and search across NAICS Code description embeddings to find the one that minimizes the cosine similarity measure. This NAICS code is then assigned to all of the observations in the BID.

With linked records and consistent industry classification, we proceeded to our final step in the process: establishing every business' start and end date. As of now, we are taking the first date associated with the business to be the start date, and taking the last date associated with the business to be the end date, with some qualifications.

Applications

On August 23, 2021, in a conversation hosted by BetaNYC, the Manhattan Borough President, Gale Brewer, said that one of the most important areas in which she thought technology could be brought to bear to solve urban problems was in the realm of policy evaluation. The example she cited was a business improvement district (or BID) on 125th Street, which seemed to be thriving, but had yet to be studied comprehensively over time.

In 2020, thanks to BetaNYC's advocacy, New York City started an official vacant storefront tracker. If this is an indication of anything, it is that data is critical to informing policy decisions. In fact, we plan to integrate our database with the data from the vacant storefront tracker to create a highly accurate storefront database.

We plan to use our database for policy evaluation. We are currently looking at proposed City Council legislation, such as Intro (1796), and running an analysis of the Madison Avenue Business Improvement District. More generally, we are hoping to use our database to study issues related to gentrification.

Immediately, we are in the process of creating an interactive online interface built on the database. Users will be presented with a visualization of the retail businesses that existed in the city at a point in time, and will be given the ability to move forward and backward along this dimension (by months). The website also include the ability to view aggregate measures, such as vacancy and business turnover rates, displayed at the building level. Eventually, we plan to crowdsource data, augmenting the profiles we have created for retail enterprises, past and present.

Finally, one more application of this database we are considering is the creation of a retail site location app. By combining our storefront database with other datasets available on Open Data, such as subway station locations, schools, parks, and demographic information, we can come to an understanding of how geographical features affect vacancy rates and business turnover rates. We can also use our data to create hedonic models that predict rent pricing. Our vision is a widely accessible retail site location app built on publicly available data.

Conclusion

While this project is still in its early stages, we are hopeful that the data will be useful for policy evaluation, and that users will enjoy being able to learn about the history of the businesses in their neighborhoods. If this is successful, we hope to expand our database and website to include other cities besides New York.

A Record Linkage

We start by generating a comparison matrix for the observations, blocking by BBL (for more on the theory and practice of record linkage, refer to Enamorado et al. [2019]). In other words, we generate a matrix comprised of all of the pairwise comparisons in each BBL, generating features based on the relationship between two records on some field. Formally, we define a matrix Γ of size $n \times m$ where $\Gamma_{i,:} = \gamma(x, j) := [\theta_1, ..., \theta_m] \ \forall \ i \in \{1, ..., n\}$ where $x, j \in \{I^\beta\} \mid x \neq j$ and where I indexes the observations from our dataset $\mathcal{D}$ in BBL β. Here, $\theta_y \ \forall y \in \{1, ..., m\}$, refer to the features generated by comparing record x to record j on field y. By way of illustration, at least one of these features (say θ_k) is defined as a comparison between the names of records x and j. Formally,

$$\theta_k := \frac{\text{tf-idf}_{[x,:]} \cdot \text{tf-idf}_{[j,:]}}{||\text{tf-idf}_{[x,:]}|| * ||\text{tf-idf}_{[j,:]}||}$$

where tf-idf refers to a term-frequency inverse document frequency matrix created for $\mathcal{D}$ on the trigrams (or contiguous sequence of 3 characters) of the business name string.[10] θ_k is also called the cosine similarity between $\text{tf-idf}_{[x,:]}$ and $\text{tf-idf}_{[j,:]}$.[11]

Once we have composed our comparison matrix Γ for our dataset $\mathcal{D}$, we proceed by classifying each $\gamma(x, j)$ as a match or a non-match, i.e. giving it a label in $\{0, 1\}$, accordingly. We test two different strategies to accomplish this: a deterministic method, and a probabilistic method. The deterministic method uses a hard-coded procedure based on the features θ_y to determine label assignment. This procedure was developed heuristically, after observing matching behaviors over several iterations. The probabilistic method uses an unsupervised clustering algorithm, specifically the Mixed Deep Gaussian Mixture Model (MDGMM) of Fuchs et al. [2020] with two components (match, and non-match), a DGMM specified to handle mixed data types (such as continuous, binary, and ordinal values). Ultimately, the probabilistic method did not achieve observably better results than the deterministic method, so we proceed with the latter. Finally, connected components in our dataset $\mathcal{D}$ are recovered using the labeled comparison matrix $\Gamma^{'}$, and are assigned a unique identifier, called a 'Location ID' or 'LID'. LIDs that share a record identifier (assigned by the agencies) are given the same 'Business ID', or 'BID'.

Notes

1. The New York Times, for instance, published in 2018 this striking array of images documenting commercial blight. There is a website, VacantNewYork.com, which purports to display in red all the vacant storefronts in Manhattan. There is even a blog dedicated to commemorating as many NYC businesses that have closed down in recent years as possible. Vanishing New York. But there are just as many articles that are skeptical of the crisis, such as this 2018 New York Magazine article which, among other things, argues that most claims about retail vacancy fail because "Comprehensive data about retail vacancy in Manhattan doesn't exist."

2. These studies may also have reached opposite conclusions because of the way they measured storefront vacancy. The City Comptroller measured vacancy in terms of vacant square footage as a percentage of the total square footage, while the Department of City Planning measured vacancy in terms of vacant storefront units as a percentage of total storefront units.

3. The Comptroller acquired vacant retail square footage data through tax filings. But as it turns out, only certain commercial properties—properties of a certain size and certain value—have to state on their tax filings what portion of their property is vacant. Hence, the comptroller estimated that it was missing about 25% of commercial properties in New York City.

   The Department of City Planning, on the other hand, compared survey data it had collected on a select number of neighborhoods from 2007/2008 with vacancy data for 2017/2018 supplied by LiveXYZ, a technology company that has mapped storefronts in New York City.

4. According to a study done in July 2020 by the small business advocacy group Partnership for New York City, as of July 2020 nearly one third of NYC's small businesses had closed and would likely never reopen. See, also, for instance, this NYT article from September, 2021.

5. Data Axle shares its methodology on its website

6. Live XYZ's CEO described the company in an AM NY article.

7. Safegraph describes its data acquisition process on its website

8. For more on NAICS codes

9. See Reimers and Gurevych [2019] for an in-depth explanation of S-BERT

10. For more on the tf-idf algorithm, refer to this helpful Medium explainer.

11. Here, cosine similarity is explained with some depth

References

Sulin Carling, Samuel Levy, and Kristina Schmidt. Assessing storefront vacancy in nyc: 24 neighborhood case studies, August 2019. URL https://www1.nyc.gov/assets/planning/download/pdf/planning-level/housing-economy/assessing-storefront-vacancy-nyc.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ted Enamorado, Benjamin Fifield, and Kosuke Imai. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2):353–371, 2019.

Robin Fuchs, Denys Pommeret, and Cinzia Viroli. Mixed data deep gaussian mixture model: A clustering model for mixed datasets. *arXiv preprint arXiv:2010.06661*, 2020.

Andrew McWilliam, Lawrence Mielnicki, and Preston Niblack. Retail vacancy in new york city: Trends and causes, 2007-2017, 2019. URL https://comptroller.nyc.gov/reports/retail-vacancy-in-new-york-city/#Endnotes.

Lindsay Poirier. Scraping data about manhattan's licensed beauty salons and barbershops, May 2018. URL https://beta.nyc/2018/05/15/scraping-nys-beauty-salon-and-barbershop-data/.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL http://arxiv.org/abs/1908.10084.