Information Waves and Firm Investment[*]

Feng Chi[†]

November 2023

Abstract

This paper measures the impact of information quality on the success of firms' investment decisions using the U.S. census as an empirical context. Over the course of a decade, information from the decennial census snapshot likely deviates from the evolving market condition, thereby making the data less relevant. I find that on average, outdated census information increases establishment failure rate by 1.6% per year. The effects are stronger for geographic areas that experience large changes in demographics, for industries that rely on precise information in small trade areas, and for independent retailers that lack alternative sources of demographic information.

Keywords: Economics of Data; Firm Investment; Industry Dynamics

[*] I thank Panle Jia Barwick, Nicholas Bloom, Ryan Chahrour, Allan Collard-Wexler, Steven Davis, Brett Hollenbeck, Byoung-Hyoun Hwang, Ben Leyden, Shanjun Li, Justin Murfin, David Ng, Robert Seamans, and Nathan Yang, as well as seminar participants and discussants at the 2023 Stanford SITE Conference on the Macroeconomics of Uncertainty and Volatility, the 5th Future of Financial Information Conference and Cornell University for helpful comments and suggestions.

[†] Cornell University. Email: fc354@cornell.edu.

1 Introduction

Information is an essential input in the firm decision-making process. Yet empirically assessing how information affects investment outcomes is challenging: firms make endogenous decisions to acquire information, and the quality of acquired information is hard to observe. To quantify the link between information and firm investment outcomes, this paper uses the predetermined release schedule of the U.S. census data as an exogenous source of variation in the information quality.

The U.S. census is a key source of demographic information for various business decisions, such as site selection, product offering, advertising and inventory management.[1] A 1990 lead article from Washington Post reports that census has become "the private sector's most comprehensive planning and marketing tool" and that "[t]housands of other consumer databases are available to businesses, but none can touch the breadth and depth of the census as a roadmap to who and where consumers are" (Farhi, 1990). Even when firms are not seeking information directly from census, census demographic data are incorporated by many intermediaries[2] that help investors large and small make investment decisions, so that investors may be influenced by census information without being aware of its original source.

However, the information quality of census data might vary over time, as the U.S. census is only conducted once every 10 years. No matter how accurate it is at the time of collection, the data loses currency over time, especially in small geographic areas that are more sensitive to demographic changes due to migration. In other words, census data might better reflect the true market condition initially but it becomes stale over the course of a decade, until the next census arrives with refreshed numbers.[3]

Thus, for firms that make investment decisions using census data, this information component arrives in a tidal wave fashion. I conjecture that, as a result, firms make worse investment decisions over time as quality of the information derived from census data deteriorates; and the pattern reverses when new census data are released.

[1] See report by the National Research Council panel for examples of business uses of census data in a variety of industries, including retail and restaurant, banks and other financial institutions, media and advertising, insurance, utilities, health care, and others. A specific case study of retail expansion involves predicting potential revenue for each market. "Some data were from business sources, but census data provided an essential component for analysis" (National Research Council, 1995).

[2] Armas (2001) and Thau (2014) provide examples of intermediaries, such as market research company Claritas and location intelligence company Esri, that integrate census demographic data into software systems that allows for data manipulation and market analysis, mapping, and the preparation of reports to assist businesses in site selection decisions.

[3] During periods with outdated census data, firms have few alternative data sources of comparable quality in granularity and scope. According to Adair (1991) "Telephone surveys by market research companies often provide the same information, but the census is usually more accurate because the bureau tries to find everyone in the nation. That, in turn, provides amazing detail about every city block."

To test this hypothesis, I examine failure patterns of establishments in the retail and restaurant industries. Firms in these industries are organized by geographic locations and serve localized markets, thereby making site selection decisions crucial to the success of their investment (Berry and Compiani, 2021; Mian and Sufi, 2014; Adelino, Ma and Robinson, 2017). More importantly, this setting allows me to observe investment decisions (*entry*) and the associated outcomes (*exit*) at a granular investment-by-investment level, which can be difficult in other industries.

This granular level of analysis is also helpful for circumventing the issue that there are limited numbers of information wave cycles to observe. By studying investment-by-investment firm decisions and outcomes, I can make full use of variations across geographic markets and time. To account for the influence of contemporaneous economic conditions, I focus my analysis on excess failure rates. These failure rates are calculated using the difference between the actual failure rate of an entry-year cohort with the average failure rate across all existing establishments from the same geographic market in the same calendar year.

Using the sample of Retail trade (NAICS 44-45) and Accommodation and Food Services (NAICS 72) businesses located within the state of New York during the 30 years from 1985 to 2014, I find that excess failure rates across entry-year cohorts follow the wave pattern as I conjectured. The failure rate increases by a statistically significant rate of 1.6% each intercensual year during my sample period. A 10-year gap between two decennial census would result in a 16% increase in failure rate due to outdated information.

As a placebo test, I create hypothetical census schedules by randomly assigning a year in each decade as the census year. Out of the 10,000 randomly created schedules, only 193 of them produce estimates greater than or equal to the original estimate using the true census schedule. Therefore, the probability of observing by random chance an effect as large as the original estimate is only 1.9%.

I then explore heterogeneous effects of outdated census data on failure rates. These results help confirm that this main effect operates through the census information channel.

First, I find stronger effects in geographic areas that experience sizable demographic shifts between two censuses. It is in these areas that the disparity between static census data and evolving market conditions becomes most pronounced. In terms of magnitude, the absolute value of the moderating effect ranges from 3% to 12%, equivalent to 19% to 75% of the main effect. Conversely, in areas with stable demographics, the effect of outdated census data on failure rates is negligible.

Second, the effects are most pronounced in industries offering highly localized products and services, notably restaurants and grocery stores. These industries tend to operate within small geographic areas, which might undergo significant demographic shifts rapidly. In contrast, industries that cater to large trade areas, such as motor vehicle dealers and furniture stores, show diminished sensitivity to variations in census data quality. Adding to this narrative, non-store retailers—those without a physical storefront, such as Electronic Shopping and Mail-Order Houses—appear to be largely unaffected by changing census data quality.

Third, when assessing the influence of census data among various firm sizes, I find that census data has the strongest impact on smaller firms, which include independent establishments and small chains. In contrast, large chains have been less influenced by the declining reliability of census data, especially in the latter half of the sample period. With the advent of the digitization era, large chains increasingly have access to emerging technologies and alternative data sources, often inaccessible to their smaller counterparts.

I conduct additional analyses to rule out alternative explanations. First, I demonstrate that my findings are not driven by business cycles. In the cross-industry analysis, both restaurant and grocery industries exhibit similar responses to outdated census data, despite their differences in cyclicality. Moreover, the impact of outdated census data on failure rate is comparable across local areas that are more/less impacted by the early 1990s and early 2000s recessions. Second, I examine whether the results are driven by government policies associated with census data. For government policies to affect the results, they would need to have differential impact on new entrants and existing establishments. Therefore, I focus on place based economic development programs that target new entrants. Using eligibility criteria of these programs, I show that the results are robust for markets ineligible for these initiatives.

Lastly, I explore the potential for firms to strategically delay entry in anticipation of new census data. However, given the specific attributes of the industries analyzed, this is unlikely to significantly influence the empirical findings.

This paper makes the following contributions to the literature. First, it empirically quantifies the value of information to firm investment decisions. Earlier work often assumes that firms form rational expectations about the market conditions, without explicitly taking into account information availability. Different from cases where firms endogenously acquire information (Leisten, 2021), the census setting creates exogenous and pre-scheduled fluctuations in information. Unlike real uncertainties about *future* states (Bloom, Bond and Van Reenen, 2007; Bloom, 2009; Jeon, 2022; Julio and Yook, 2012; Kellogg, 2014; Kumar and Zhang, 2019; Jens, 2017), outdated data creates informational uncertainty about the *current* state.[4]

Second, my work is related to the broader literature that assesses the value of information gathered and distributed by government agencies as public good (Craft, 1998; Gao and

[4] Some other work that studies firm responses to information is Goldfarb and Xiao (2016), where they show that firms might overreact to transitory shocks (e.g., weather) and make wrong inferences.

Huang, 2020; Binz, Mayew and Nallareddy, 2021; Nagaraj, 2022). To the best of my knowledge, this is the first study to examine the economic value of census in the context of firm investment. Compared with many other countries, the U.S. Census is unique in the sense that it makes the data widely available and accessible, creating an ecosystem of census data users (Donnelly, 2019). Despite its original purpose in determining political outcomes, the census data also has implications on the economic efficiency, as it informs firms on optimal investment allocations.

Finally, my research contributes to research about firm turnover. Past theoretical and empirical studies have largely focused on uncovering cross-sectional differences in firm turnover rates (Asplund and Nocke, 2006; Collard-Wexler, 2013; Dunne, Roberts and Samuelson, 1989; Fan and Xiao, 2015; Hopenhayn, 1992; Jovanovic, 1982). I continue this discussion by offering a novel information-specific mechanism for which small firms (i.e., independents) are more likely to fail than large firms (i.e., chains). This finding is in-line with that of Collard-Wexler (2013), which shows that uncertainty reduction can have a material impact on firm turnover. There are broader implications of my findings on the dynamics of market structure, as census data is open access and non-rival (Jones and Tonetti, 2020), benefiting anyone who utilizes it, including small establishments, while exclusiveness of private (big) data can be used to reinforce the competitive advantage of large firms (De Loecker and Eeckhout, 2018; Farboodi et al., 2019; Farboodi and Veldkamp, 2021; Farboodi et al., 2022).

The rest of the paper is organized as follows. The institutional background about the U.S. decennial census is provided in section 2, and a description of data is in section 3. My empirical strategy is discussed in section 4, followed by the main findings about the effects of outdated census data on establishment failure in section 5. Furthermore, I highlight some variations in failure rate patterns across geographical, industry, and firm dimensions in section 6. Finally, I discuss potential alternative explanations in section 7 and the potential impact of delayed entry in section 8. A concluding summary are provided in section 9.

2 Institutional Background

2.1 The Decennial Census

The decennial censuses from 1970 to 2000 use fairly consistent collection methods and variables throughout this time period. The short form collects basic demographic variables from 100% of the population, while the long form captures a wide range of socioeconomic and housing variables from a 1-in-6 sample of the population. The 1-in-6 sample is large enough that the estimates are often treated as if they were exact counts (Donnelly, 2019).

Table 1: Decennial Census Data Release Timeline

Data File	Content	1990 Census	2000 Census
Redistricting Summary File	Population counts used for redistricting	March 1991	March 2001
Summary File 1	Population and housing characteristics	March 1991	June 2001
Summary File 2	Cross tabulations of SF1 by racial groups	August 1991	September 2001
Summary File 3	Detailed socioeconomic characteristics	May 1992	June 2002
Summary File 4	Cross tabulations of SF3 by racial groups	March 1993	October 2002
PUMS	Samples of individual responses	July 1993	April 2003

Sources: https://www.census.gov/programs-surveys/decennial-census/technical-documentation/complete-technical-documents.1990.html
https://www.ibrc.indiana.edu/ibr/2001/spring01/05.pdf

As summarized in Table 1, tabulated data are released on a flow basis after the census was conducted.[5] Basic population counts are published in March the next year, followed by more detailed population, housing and socioeconomic characteristics. Most of the data that appeal to broad business interests, including Summary File 1 and Summary File 3, is made available within 18 months since the initial release.

2.2 Decreased Accuracy In-between Censuses

Despite the high accuracy at the time of data collection, reliability declines over the course of a decade. Small areas, in particular, can experience rapid population and economic changes during intercensual years.

Various efforts have been made to impute demographic information in between two decennial censuses. The Census Bureau provides annual population estimates utilizing administrative records,[6] although the estimates only go down to the county level.[7] Commercial data vendors also make their own estimates, with varying degrees of success (Cropper et al., 2012).

To produce more timely estimates, by 2005 the Census Bureau started publishing the American Community Surveys based on a rolling sample methodology that surveys a much smaller number of households but more frequently, in the hope of achieving the same level of precision as the data from the long-form sample. However, the margins of error can be

[5] Since 1930 "Census Day" has been April 1 in the first year of each decade (ending in zero). Census attempts to capture a snapshot of this specific reference date, while actual census-taking begins before this date and extends for months thereafter. See https://en.wikipedia.org/wiki/United_States_census.

[6] For example, birth and death statistics are from health departments; domestic migration data are from the IRS and the Centers for Medicare and Medicaid.

[7] The 2000s estimates are fairly accurate at the county level, as the records were collected at the county level; the 1990s estimates are relatively poor because results are imputed from the state level numbers. A large portion of the error of closure is concentrated in the 5-34 age group. https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/intercensal/intercensal-nat-meth.pdf

high for small geographic areas (National Research Council, 2015; Donnelly, 2019).

2.3 Usage of Census Data

Since the launch of the Census Bureau's first website in the mid-1990s, anyone can easily look up census information online (Donnelly, 2019).[8] Before that, the public accessed the census data products in print, on computer tape, or on CD-ROM, through a network of affiliated organizations (state executive departments, chambers of commerce, councils of governments, university research departments or libraries).[9] The South Carolina Census State Data Center estimated that 35 percent of the annual requests received for census data are from businesses (National Research Council, 1995).[10]

Even without deliberately making the request, large and small businesses can still be influenced by census data when they base their investment decisions on inputs provided by intermediaries. Marketing data providers such as Claritas, Esri, and SafeGraph incorporate census demographic data into their databases, enhanced by other proprietary data sources and mapping tools to help clients make site selection decisions (Armas, 2001; Thau, 2014).

Commercial real estate agents are trained to help potential buyers and tenants find the right location for their business needs. Demographic data of the local trade area is an important component of their knowledge base. Figure A1 shows an example of the databases used by commercial real estate brokers, featuring demographic characteristics of various neighborhoods for easy comparison.[11] The underlying data of such databases is often directly obtained from census.

[8] The internet helps make census data more directly available to small businesses. According to Matthew Cunningham, manager of the Texas Business and Industry Data Center "More small businesses would use it just because it's free, especially now that people are in the information age" (Armas, 2001).

[9] See `https://cityoflancasterpa.com/wp-content/uploads/2013/10/Business-Startup-Toolkit-April-2013.pdf` for an example of a local library's business center providing guidance to local entrepreneurs on how demographic data from the Census Bureau's website can be used to conduct market research.

[10] Adair (1991) provides an example of Eckerd drugstore using census data to mange inventory and product offering in markets with different demographic characteristics. Other examples include Volvo North America using census data to find the best location for its dealerships, and Winn-Dixie grocery who "won't build a new store unless the census indicates there are enough households to support it" (Adair, 1991). According to a report by the Council of Economic Advisers (The Council of Economic Advisers, 2000), "Numerous small businesses responded to a request for examples of business uses of census data." Farhi (1990) reports that "A cemetery owner recently asked the Census Bureau to help him determine the number of people of Italian ancestry living near him in order to anticipate the demand for crypts."

[11] See `https://www.apto.com/blog/the-best-of-commercial-real-estate-data-sources-demographics-broker-databases/` for other examples.

3 Data and Summary Statistics

3.1 Establishment Data

3.1.1 Data Source

Establishment data comes from Dun & Bradstreet's National Establishment Time Series (NETS) database. Every year Dun & Bradstreet takes a snapshot of all U.S. business establishments, collecting information on establishment identifier, address, industry classification, among others. This data is ideal, as it allows me to track when exactly each establishment opened (and if relevant, closed). For many of the alternative data sources available to the public (e.g., Longitudinal Business Database and County Business Patterns), the information about entry/exit conveyed is represented by the net change in establishment counts for each geographic market in a given year, which means that such data is unable to group establishments into entry-cohorts, nor capture cohort-specific failure rates.

The NETS database is comparable to the U.S. Census Bureau's Longitudinal Business Database in terms of coverage but offers the advantage of not requiring special access (Rossi-Hansberg, Sarte and Trachter, 2021).[12] I obtain a sample of the NETS data for Retail trade (NAICS 44-45) and Accommodation and Food Services (NAICS 72) businesses located within the state of New York covering 30 years from 1985 to 2014.[13] The sample contains 6,459,013 establishment-year observations across 857,792 unique establishments.

Focusing on the retail and restaurant industries offers two key advantages. Firms in these industries are organized by geographic locations and largely serve a local market, making local demographic statistics crucial to their entry decisions (Berry and Compiani, 2021; Mian and Sufi, 2014; Adelino, Ma and Robinson, 2017). More importantly, I can observe their investment decisions (entry) and associated outcomes (exit) at the individual level, allowing me to link investment outcome directly with the information set available at the time of

[12] One concern about the NETS data is its discrepancy with the census LBD data, primarily attributed to employment imputation for small establishments ((Barnatchez, Crane and Decker, 2017)). Since my analysis exclusively revolves around establishment entry and exit patterns, this imputation does not play a role in my empirical context. Even in terms of employment, the observed disparities between the NETS and LBD datasets predominantly manifest within specific industry sectors. As reported by (Barnatchez, Crane and Decker, 2017), these differences are largely confined to a 2% margin in the Retail trade (NAICS 44-45) and Accommodation and Food Services (NAICS 72) sectors, which are the focus of my analysis. In light of the usefulness of NETS data for capturing establishment entry/exit, this database has appeared in a wide range of empirical studies that span economics, finance, and public policy (Addoum, Ng and Ortiz-Bobea, 2020; Currie et al., 2010; Kolko, 2012; Levine, Toffel and Johnson, 2012; Neumark, Wall and Zhang, 2011; Schuetz, Kolko and Meltzer, 2012; Tsui et al., 2020).

[13] A large proportion of census tracts are located in the densely populated New York City. To verify that my results are not driven by the tracts in New York City, I report in Table A4 results using the subsample of census tracts located within (outside) New York City. The effects are very similar in terms of magnitude across the two subsamples.

decision-making.[14]

3.1.2 Entry, Exit and Failure Rate

The observed establishment entry (exit) is identified as the year its unique identifier first appears (disappears) in the sample. I make the assumption that the site selection decision is made during the same year as the observed entry, given that it only takes a few months to open a retail business,[15] and that the establishment data are December snapshots. This assumption about entry timing is also consistent with the time it takes for an entrant to become active in theoretical models of industry dynamics, which is often referred to as the "time-to-build" assumption (Ericson and Pakes, 1995) and used in empirical studies about retail dynamics (Arcidiacono et al., 2016; Fang and Yang, 2022; Hollenbeck, 2017; Igami and Yang, 2016; Maican and Orth, 2018; Suzuki, 2013).

I measure investment outcome using the failure rate within the first 5 years of entry, to address potential right censoring for the latter years of my sample. The retail and restaurant industries have relatively high turnover: 49% of the establishments that within the entry cohorts fail within their first 5 years. Long-term survival is less likely to be driven by assessment of the market characteristics when the initial entry decision was made.

My sample period does not permit me to identify exits with certainty for establishments that enter after 2009, or time of entry for establishments in the 1985 sample. Therefore, I measure 5-year failure rates for establishments that entered during 1986-2009, leaving me with 24 entry-year cohorts to study.

3.1.3 Market Definition

Establishments in the retail and restaurant industries have relatively small trade areas,[16] which is precisely why the location choice is essential to success in this business. I define markets at the census-tract level and map an establishment's geographic coordinates to corresponding census tract using the 2010 TIGER/Line Shapefiles, so that the physical boundary of the markets remains fixed throughout my sample and can be easily linked to

[14] Other industries also benefit from the census data in making various investment decisions related to product offering, advertising spending, hiring, and logistics (National Research Council, 1995), but for these industries outsiders often at best observe investment spending and profitability aggregated at the firm level.

[15] It generally takes 2-6 months to open a retail business after the site has been selected. See for example, https://www.thebalancesmb.com/how-long-does-it-take-to-start-a-business-3974594, https://thegrocerystoreguy.com/how-long-does-it-take-to-build-a-grocery-store/, https://pos.toasttab.com/blog/on-the-line/how-long-does-it-take-to-open-up-a-restaurant.

[16] For a salient illustration of how small these trade areas might be, please refer to this example by Thomadsen (2005) about the fast food industry.

census demographic information.[17]

3.1.4 Summary Statistics

The state of New York is the third-largest economy in the United States and has diverse industries and geographic areas. Table A1 reports the number of establishments across NAICS 4-digit sub-industries during my sample period.

3.2 Demographic Data

3.2.1 Data Source

Demographic data comes from the U.S. Decennial Census (1980, 1990, 2000 and 2010) and the American Community Survey (ACS 2008-2012 5-year survey).[18] To identify areas that experience large demographic shifts, I calculate changes in demographic variables between two decennial censuses across geographic areas.

3.2.2 Demographic Shift

Census tracts are sometimes divided or combined every 10 years to maintain the optimal threshold for population size. To make consistent comparisons, census tract data from 1980, 1990, and 2000 are all mapped to 2010 geography using the LTDB crosswalk developed by Logan, Xu and Stults (2014). In addition, dollar values are inflation adjusted based on the Consumer Price Index research series using current methods (CPI-U-RS) from St Louis Fed. Table A2 provides details on the construction of each demographic variable.

Table 2 tabulates the distribution of changes in demographic variables between two decennial censuses. Census-tract level demographics exhibit relatively large variation, highlighting that small geographic areas can experience large unexpected shifts in demographics over the course of a decade.

[17] Among the census geographies, county and census tract have relatively stable definitions and ID codes across the relevant censuses in my sample. In contrast, zipcode tabulations are only available since the 2000 census and county subdivisions since the 1990 census; block and block groups may be completely renumbered in the next census; places only cover concentrated settlements; and metropolitan statistical areas update delineations couple of times per decade (Donnelly, 2019).

[18] I obtain data on 2010 education, employment, income, and housing value from the ACS 2008-2012 5-year survey, since these variables are not available in the 2010 decennial census.

Table 2: Changes in Demographic Variables

Demographic Variable	p10	p25	p50	p75	p90
Population	-0.09	-0.04	0.02	0.10	0.21
%Kids (0-17)	-0.20	-0.13	-0.04	0.06	0.16
%Young (18-34)	-0.28	-0.18	-0.07	0.03	0.11
%Middle (35-64)	-0.03	0.02	0.07	0.13	0.20
%Old (65+)	-0.24	-0.11	0.03	0.20	0.40
%White	-0.30	-0.12	-0.03	-0.01	0.04
%Black	-0.29	-0.07	0.19	0.78	1.94
%Asian	-0.23	0.06	0.43	1.06	2.23
%Latino	-0.36	-0.05	0.45	1.46	3.32
%College degree	-0.33	-0.07	0.14	0.35	0.66
Unemployment rate	-0.55	-0.36	-0.08	0.30	0.86
Median income	-0.21	-0.11	-0.01	0.12	0.28
Median house value	-0.25	-0.14	0.06	0.55	1.21
Observations	89691				

Notes: This table describes the distribution of changes in local demographics between two decennial censuses at the census-tract level. Table A2 provides details on the construction of each demographic variable.

4 Empirical Strategy

To measure the impact of census data on investment outcomes, I compare cohorts of establishments based on their entry years. Different entry cohorts are faced with varying levels of census data quality when they make their entry decisions. Establishments entering early in the decade benefit from recent census data that accurately reflects current market conditions. In contrast, those entering later in the decade rely on the same data, which has become progressively outdated, putting them at a potential disadvantage. As a result, the timing of an establishment's entry can be viewed as a measure of the quality of census data at its disposal.

While information quality plays an important role, other time-varying factors can also significantly influence the failure rate of an entry cohort. For instance, if a store opens one year before the financial crisis and failed, its failure is more likely due to the macroeconomic condition than a poor location choice.

To account for the influence of time-varying factors, I use the failure rate of all existing establishments from a local market in a calendar year as the baseline failure rate for that year.[19] Assuming time-varying factors impact failure rates similarly across establishments in

[19] Intuitively, subtracting the average failure rate is analogous to taking out the calendar year fixed effect,

the same market, irrespective of their entry timing, the baseline failure rate helps eliminate potential confounding effects.[20] This approach allows me to disentangle the economic impact of information quality at entry (which affects only new entrants) from changes in underlying economic conditions (which affects all existing firms in the market).

The excess failure rate for an entry cohort in a given calendar year is thus defined as the actual failure rate for the entry cohort subtracted by the baseline failure rate across all existing establishments:

$$\Delta f_{imt} = f_{imt} - F_{mt} = \frac{\text{Exit}_{imt}}{B_{imt}} - \frac{\sum_i \text{Exit}_{imt}}{\sum_i B_{imt}}, \tag{1}$$

where B_{imt} is the number of establishments that enter at year i into a geographic market m and still exist at the beginning of year t. Here, f_{imt} is the actual failure rate for entry cohort i in market m from calendar year t, and F_{mt} is the average failure rate in calendar year t across all existing establishments in market m.

The excess failure rate of an entry-year cohort i from market m within the first 5 years of entry is calculated by taking a summation of the excess failure rate for the entry-year cohort in each of the 5 years:

$$\text{Excess failure rate}_{im} = \sum_{t=i+1}^{i+5} \frac{\Delta f_{imt} * B_{imt}}{B_{im}}. \tag{2}$$

5 Main Results

5.1 Failure Rate Pattern

My central hypothesis is that failure rates across entry-year cohorts follows a tidal wave pattern, characterized by worse investment decisions over time as the quality of information derived from census data deteriorates and a reversal when the new census data are released.

To test this hypothesis, I first examine the pattern of excess failure rates across entry-year cohorts using a flexible specification:

$$\text{Excess failure rate}_{im} = \sum_{i=1986}^{2009} \beta_i \times I(\text{Entry-Year} = i) + \varepsilon_{im}, \tag{3}$$

where i is an index for each entry-year cohort and m is an index for market (i.e., census tract). The outcome variable Excess failure rate$_{im}$ is the 5-year failure rate for establishments that

which ultimately allows me to control for external factors that are common to all establishments in that year.

[20] I test the robustness of this assumption in subsection 7.1

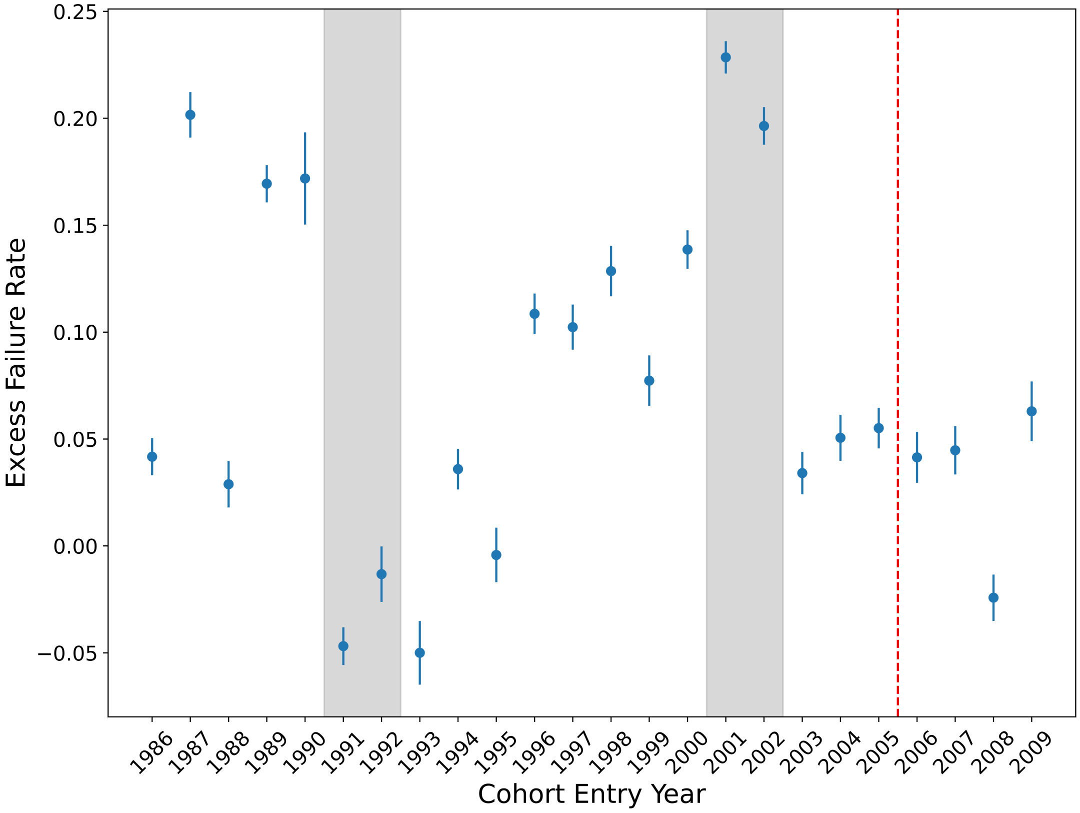


Figure 1: Excess Failure Rate by Entry Cohorts

Notes: The figure plots coefficients and associated 95% confidence intervals of the entry-year indicator variables in Equation 3. The dependent variable is excess failure rate. Standard errors are clustered at the census-tract level. Table A3 presents the regression results in the alternative tabular format.

enter market m in year i, relative to the average failure rate of establishments in the same market m. Standard errors are clustered at the market level to flexibly account for potential serial correlations in the error term.

Figure 1 visualizes the coefficient estimates from Equation 3 across entry cohorts. These estimates reveal a wave-like pattern that aligns with the census data release schedule.[21] Following the release of the 1990 census data, the failure rate drops substantially. For the 1991 and 1992 cohorts, as additional census datasets roll out—indicated by the shaded regions—the failure rate declines. This immediate decline in the failure rate suggests that many firms draw insights from demographic variables in the early release. By the time the 1993 cohort enters, all major census data files have become available. In the following years,

[21] For a detailed schedule, please refer to Table 1.

failure rates of the subsequent cohorts continue to rise, as the 1990 census data becomes outdated. When fresh data from the 2000 census arrives, the failure rate does not fall right away, indicating that firms now benefit more from detailed socioeconomic data released later. Post 2003, once all the new data is assimilated, failure rates move back up again.

In 2005, the Census Bureau introduces the American Community Survey (ACS),[22] which is designed to provide more timely but potentially less precise estimates than the original decennial census. The dotted line marks this transition. Following this point, several ACS datasets become available[23] and the failure rate experiences a more gradual increase.

5.2 Failure Rate and Distance to Census Data Release

To quantify the impact of census information quality on firm investment outcomes, I model the wave pattern in Figure 1 in a parametric specification with two main elements. First, excess failure rate is characterized as a function of how long it has been since the last census is released. This time gap serves as a proxy for the discrepancy between the current state of the market and the snapshot recorded in the previous census. The longer this gap, the more outdated the information becomes. Consequently, when firms base entry decisions on this outdated information, their failure rate tends to be higher.

Secondly, I introduce a breakpoint that delineates the two distinct temporal phases within each information wave cycle. The initial phase includes entry cohorts from the first two years following the census data release. Given the ongoing release of new census data files during this period, the impact on the failure rate is ambiguous; it depends on whether an average firm places more value on early or later data batches. The second phase covers the years following the first two. With all major files already released, census data becomes increasingly outdated during this period, making it easier to isolate the impact of outdated census data on firm failure.[24] The resulting piece-wise linear specification that captures these

[22] The American Community Surveys collects socioeconomic characteristics from a significantly smaller sample of the population on a rolling basis. In larger regions or densely populated areas, annual estimates are derived from a full year's worth of data, whereas in smaller areas, samples are aggregated over 5 years to improve accuracy. While the smaller sample size of the American Community Survey makes it more cost-effective and timely, it is not as precise as the decennial census. In some cases, particularly for smaller geographies or specific data breakdowns, the margin of error can be greater than the actual estimate itself.(Donnelly, 2019)

[23] From 2005, 1-year estimates are provided for areas with a population of at least 65,000; from 2007, 3-year estimates for areas with over 20,000 people; and from 2009, 5-year estimates for all geographies. Like the decennial census, ACS data is released in the subsequent year, e.g., 2005 data is released in 2006.

[24] Focusing on this phase also makes my results less sensitive to the time to build assumption, as firms will have had more time to incorporate updated census data into their investment decisions. For example, an establishment that enters by the end of 2003 has between 1.5 years (since the release of Summary File 3) and 2.75 years (since the release of Summary File 1) to access census data and make investment decisions, depending on which demographic variables are important to these decisions.

information waves can thus be written as follows:

$$\begin{aligned}\text{Excess failure rate}_{im} = {} & \alpha_1 I(\mathrm{S}_i < 2) + \alpha_2 I(\mathrm{S}_i \geq 2) \\ & + \beta_1(\mathrm{S}_i - 2) I(\mathrm{S}_i < 2) + \beta_2(\mathrm{S}_i - 2) I(\mathrm{S}_i \geq 2) + \varepsilon_{im} \quad (4)\end{aligned}$$

where i is an index for the entry-year cohort and m is an index for the market. Standard errors are clustered at the market (census-tract) level to to flexibly account for potential serial correlations in the error term.[25]

S_i is the distance from entry-year i to the initial data release year of the most recent decennial census. This variable captures how outdated census data has become over time. Specifically for my sample period I define S_i to be:

$$\mathrm{S}_i = \begin{cases} i - 1981, & \text{if } 1981 \leq i < 1991 \\ i - 1991, & \text{if } 1991 \leq i < 2001 \\ i - 2001, & \text{if } 2001 \leq i < 2010 \end{cases} \quad (5)$$

The coefficients α_1 and β_1 represent the intercept and the slope of the first phase, when census data files are gradually being released; the coefficients α_2 and β_2 represent the intercept and the slope of the second phase, when all major census data has been released and firms have had the time to incorporate the new information into their entry decisions.

The main coefficient of interest β_2 captures on average how much failure rate increases when entry takes place one year further away from when the census snapshot was taken. The interpretation of β_1 is somewhat nuanced. Given the potential heterogeneity in the informativeness of different data and in the time it takes to utilize the newly released data by different firms, there is no obvious prediction on the effect for the first phase. Given the ambiguities with respect to β_1, I focus my discussion on the interpretation of β_2.

Table 3 reports the coefficient estimates. Column (4) - (6) report coefficients estimated using a more flexible specification Equation 4, where the two segments have separate slopes and intercepts. Columns (1) - (3) present results for an alternative specification with the restriction that the two linear components connect at the break-point.[26]

The coefficient estimates of β_2 are very similar across these two specifications. Columns (1) and (4) report coefficients estimated from the full sample. The distance between the es-

[25] For robustness, Table A5 reports standard errors using alternative ways of clustering to further account for any potential spatial correlation within the same entry cohort. The β_2 coefficient estimates are robust under these alternative ways of clustering.

[26] As an aside, the R^2 values I obtain are within a similar range as the implied pseudo-R^2 values in other empirical studies about retail exit (Kosová and Lafontaine, 2010), which range from 0.01 to 0.06.

Table 3: Excess Failure Rate and Distance to Census Data Release

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-0.046***	0.016***	-0.108***	0.006	0.034***	-0.032***
	(0.002)	(0.003)	(0.003)	(0.005)	(0.008)	(0.006)
β_2	0.015***	0.025***	0.004***	0.016***	0.025***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
α_1				0.105***	0.021	0.164***
				(0.009)	(0.014)	(0.010)
α_2				0.010***	-0.012**	0.027***
				(0.002)	(0.004)	(0.003)
Constant	0.017***	-0.009*	0.038***			
	(0.002)	(0.004)	(0.003)			
Observations	100433	56299	44134	100433	56299	44134
R^2	0.007	0.027	0.033	0.008	0.027	0.036

Notes: This table summarizes the tests for the relationship between excess failure rate and an entry cohort's distance to census data release, using piece-wise linear regressions with the break-point at two years after initial release. The coefficients α_1 and β_1 represent the intercept and the slope of the first phase; the coefficients α_2 and β_2 represent the intercept and the slope of the second phase, respectively. The main coefficient of interest β_2 captures on average how much failure rate increases when entry moves one year further away from when the census snapshot was taken. In Columns (1) - (3), the two segments connect at the break-point. In Columns (4) - (6), the two segments have separate slopes and intercepts. Columns (1) and (3) use the full sample periods of entry cohorts from 1986 to 2009. Column (2) and (4) use the sub-sample of entry cohorts before 2000. Columns (3) and (5) use the sub-sample of entry cohorts starting from 2000. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

tablishment's entry time and the previous census data release has a positive and statistically significant impact on the establishment's cohort based on entry year. The coefficient estimate of 0.016 represents an annual increase of 1.6% in failure rate. Based on this estimate, a 10-year gap between two decennial censuses would result in a 16% increase in failure rate if firms have to rely on the most outdated information.[27] Considering that the average 5-year failure rate in my sample is 49%, the decade old information could increase a firm's baseline failure rate by almost a third.

I further divide the entry cohorts into two groups: those before 2000 (pre-2000) and those after 2000 (post-2000). Columns (2) and (5) present the results for the pre-2000 sample, while

[27] I also consider using establishments other than the entry cohort to measure the benchmark failure rate of the local market in a given calendar year. As shown in Table A6, the results are qualitatively and quantitatively very similar to the main result from the original baseline specification.

Columns (3) and (6) show the results for the post-2000 sample. In both time periods, the coefficient estimates of β_2 are positive and statistically significant. However, these effects are notably larger in the pre-2000 period. One contributing factor to this difference is the introduction of the American Community Survey, which provided more timely data to supplement the decennial census. This transition occurred within the broader context of the information age, where new technologies[28] provided firms greater access to alternative sources of demographic data. Overall, the evidence suggests that the impact of census data on firms is influenced by the presence of alternative information sources.

5.3 Placebo Test

To more closely evaluate whether the pattern in Figure 1 occurs by chance, I use a placebo test that shuffles the treatment assignment. To implement this test, I randomly shuffle the census years and re-estimate the main specification in Equation 4. In particular, I randomly assign a year from each decade in the 1970s, 1980s, 1990s, and 2000s as the census year, resulting in 10,000 possible combinations and thus 10,000 alternative census schedules.[29] For each hypothetical schedule, I then recalculate the distance to the most recent census for each entry-year cohorts and then re-estimate Equation 4.

Out of the 10,000 randomly created schedules, 193 produce β_2 estimates greater than or equal to the original estimate using the true census schedule. The probability of generating by chance a coefficient estimate as large as the original estimate is thus 1.9%. Since this p-value is not based on standard errors at the geographic market level, it helps verify that spatial correlation is unlikely to drive the main patterns I infer.

6 Heterogeneity in Failure Rate Patterns

In this section, I explore potential heterogeneity in the impact of outdated census data on failure rates. I examine whether the main effect documented in the previous section is more pronounced for (1) geographic areas that experience substantial demographic shifts, (2) industries that depend on localized information in small trade areas, and (3) small firms that lack alternative sources of information. My findings suggest that the effect on the failure rate is more pronounced in scenarios where census data plays a more crucial role. Additionally,

[28] Many consumer activities can be tracked by their web browsing history and mobile phone usage, thanks to the growing popularity of the internet and smart phones.

[29] For example, one hypothetical set of census schedule could be 1975, 1988, 1992, and 2005. In this case, an establishment that enters in 1986 has to rely on data from the 1975 census and its distance to the most recent census release is 10 years. I need to draw a census date from the 1970s in this scenario as the hypothetical 1980s census occur after the start of my sample.

I use placebo tests to show no discernible effect in instances where census data is expected to have no significant impact.

6.1 Failure Rate and Shifts in Demographics

6.1.1 Incremental Effect

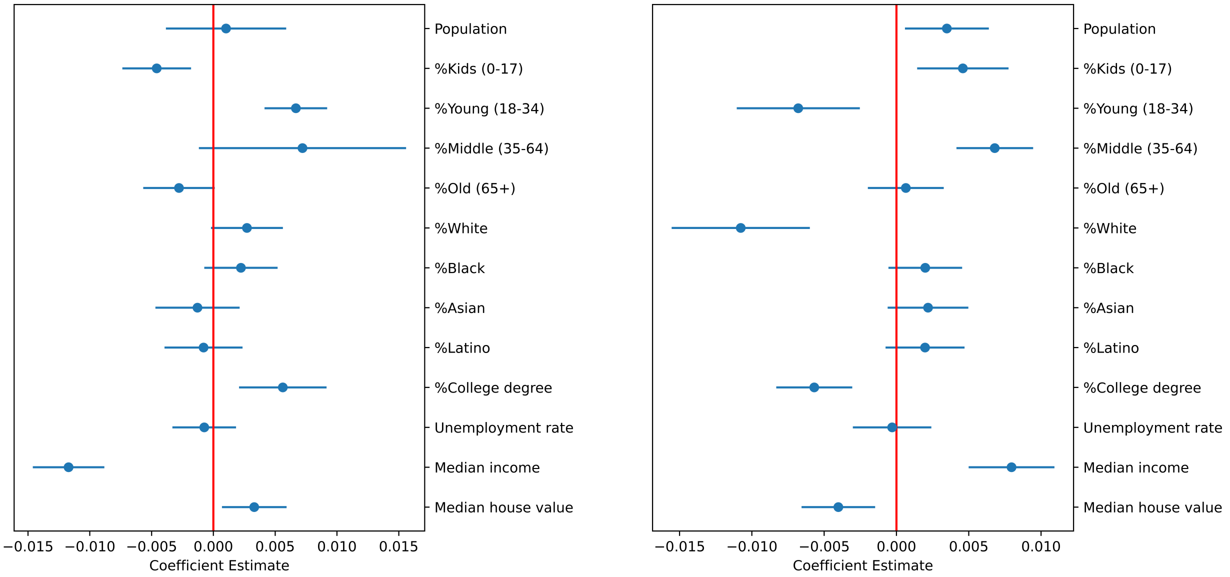
This section explores how the relevance of census data might differ across various geographic areas. Geographic variations create cross-sectional differences in the relevance of census data beyond timeliness. In areas where demographics remain stable over time, census data collected from the past continues to provide insight into current conditions. However, in areas undergoing rapid demographic shifts, the relevance of census data diminishes quickly. Consequently, I expect a more pronounced impact of outdated census data on failure rates in areas experiencing rapid demographic changes.

To empirically test the incremental effect of outdated information on failure rates in areas with substantial demographic shifts, I interact the slope terms from Equation 4 with an indicator variable that flag these areas and estimate the following regression:

$$\begin{aligned}\text{Excess failure rate}_{im} &= \alpha_1 I(\mathrm{S}_i < 2) + \alpha_2 I(\mathrm{S}_i \geq 2)\\ &\quad + \beta_1(\mathrm{S}_i - 2)I(\mathrm{S}_i < 2) + \beta_2(\mathrm{S}_i - 2)I(\mathrm{S}_i \geq 2) + \beta_3\widetilde{\Delta X}_{im}\\ &\quad + \gamma_1\widetilde{\Delta X}_{im}(\mathrm{S}_i - 2)I(\mathrm{S}_i < 2) + \gamma_2\widetilde{\Delta X}_{im}(\mathrm{S}_i - 2)I(\mathrm{S}_i \geq 2) + \varepsilon_{im}\end{aligned} \tag{6}$$

where $\widetilde{\Delta X}_{im}$ is an indicator variable equal to 1 if the value of a demographic variable X in market m changes above a threshold between the two decennial censuses surrounding entry-year i. Because different demographic variables might have heterogeneous effects[30] on business success, I estimate Equation 6 for each demographic variable X, and for positive and negative change separately. Table A2 provides detailed definition of each demographic variable and Table 2 tabulates changes in demographic variables at the census-tract level. Other variables are defined in Equation 4.

[30] My empirical design differs from an alternative approach to focus only on deviations between interpolated population versus revealed population from the census (Serrato and Wingender, 2016). There are two main reasons I depart from this approach. First, the comparisons between interpolated and actual values might introduce prediction errors. Second, although using only one dimension of the census might be well-suited in Serrato and Wingender (2016) as their research is focused on population-dependent government spending, my empirical context about retail entry requires more flexibility about the set of demographics that retailers might care about. A market's attractiveness is likely a non-trivial function of many different demographic variables. For this reason, my focus here is not solely on surprises in population counts.



(a) Negative Shift

(b) Positive Shift

Figure 2: Coefficient Estimates (γ_2) by Demographic Variable

Notes: This figure reports coefficient estimates γ_2 and associated 95% confidence intervals from the regression Equation 6 for each demographic variable. A negative (positive) shift is a decrease (increase) in the demographic variable larger than 10%. Standard errors are clustered at the census-tract level. The vertical line indicates where the coefficient estimate is 0.

In areas with large changes in demographics, the impact of outdated census data on the failure rate of new entry cohorts over time is represented by $\beta_2 + \gamma_2$. The main coefficient of interest γ_2 captures the incremental effect of stale information. Figure 2 plots the coefficient estimates for γ_2 and the associated 95% confidence intervals where changes in demographic variables exceed the ±10% threshold.[31]

Among the demographic variables I analyze, I find evidence of statistically significant incremental impact for age, education, income and housing value. For example, when an area has a markedly reduced proportion of young individuals (*%Young 18-34*) than firms had expected based on data from the previous census, the increase in failure rate is further amplified, as indicated by the positive γ_2 estimate in Figure 2a. Correspondingly, a favorable surprise moderates the increase in failure rate, as indicated by the negative γ_2 estimate in Figure 2a. Similar effects are observed for higher education (*%College degree*) and wealth *Median house value*, where an increase in these demographic variables is associated with favorable firm outcomes. In contrast, unexpected decreases (increases) in the percentage of

[31] To assess the sensitivity of this threshold, I also consider specifications using alternative cutoffs at ±15% and ±20% in Table A7. The coefficient estimates are similar in magnitude to results from the 10% cutoff.

kids (*%Kids*) and *%Median income* reduce (increase) failure rates.

Businesses have heterogeneous preferences regarding the ideal demographic profile.[32] Given these diverse considerations, the effects outlined above reflect average outcomes across the firms in my sample. Irrespective of the direction of their impact, these large demographic shifts highlight the gap between outdated census data and current market condition, adding another layer to the effect of outdated data on firm failure. In terms of magnitude, the absolute value of the coefficient estimates for the incremental effect γ_2 ranges from 3% to 12%, equivalent to 19% to 75% of the main effect β_2.

On the other hand, I do not find consistent evidence of incremental effects related to demographic variables such as population, race, and unemployment rate. Unlike the other demographic variables, local population and unemployment statistics are available at annual and even monthly frequencies.[33] Businesses can access this information without waiting for the census data release, which explains why population and unemployment data in the decennial census have a limited impact.

6.1.2 Placebo Test

I conduct a placebo test using areas that experience little changes in demographics between two censuses. In these areas, timing of entry relative to census data releases is unlikely to be associated with significant variation of census information quality. Specifically, I construct a sub-sample of census tracts that have less than 10% change in absolute value for *%Young*, *%College degree*, *Median house value*, *%Kids* and *%Median income*, census demographic variables shown to have a significant impact on firm failure in the previous section.[34] I replicate Table 3 on this "no surprise" sub-sample and confirm that distance to census data release has practically no effect on failure rate in scenarios where information is plausibly stable over time. As reported in Table 4, the coefficient estimate on β_2 is -.0000901.

6.2 Failure Rate and Industry

To explore heterogeneity in the impact of census data across industries, I estimate Equation 4 for each NAICS 4-digit industry. In this context, the benchmark failure rate is specific to establishments within the same industry from the same neighborhood.

[32] For example, while fast food restaurants and dollar stores might shy away from high-income neighborhoods, luxury boutiques or gourmet restaurants may target them specifically. Similarly, while bars might find neighborhoods with young children less appealing, toy stores or family entertainment centers would find them ideal

[33] At the county level, annual population can be inferred from birth and death records National Vital Statistics System (NVSS), while monthly unemployment statistics are released by Bureau of Labor Statistics.

[34] Imposing the 10% restriction on all demographic variables leaves no observations in the sample.

Table 4: Census Tracts with Little Change in Demographic Variables

	(1)	(2)	(3)
β	-0.000		
	(0.009)		
β_1		-0.001	0.010
		(0.035)	(0.089)
β_2		-0.000	-0.000
		(0.012)	(0.012)
α_1			0.099
			(0.150)
α_2			0.079
			(0.043)
Constant	0.082*	0.081	
	(0.038)	(0.041)	
Observations	276	276	276
R^2	0.000	0.000	0.000

Notes: This table summarizes the replications of the main specifications using the subsample of census tracts that have less than 10% absolute change in value for *%Young*, *%College degree*, *Median house value*, *%Kids* and *%Median income*. Column (1) is a simple linear specification. Columns (2) - (3) are piece-wise linear regressions with the breakpoint at two years after initial release. In Column (2), the two segments connect at the breakpoint, while in Column (3), the two segments have separate slopes and intercepts. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Figure 3 reports the coefficient estimates of β_2 across 33 NAICS 4-digit industries. Notably, restaurants and grocery stores exhibit the highest sensitivity to such changes. These two industries also have the largest number of establishments in my sample, which underscores the importance of proximity to their customer base.

The relative ranking of industries in Figure 3 hints at a general relationship between the size of an establishment's trade area and its reliance on census data. Businesses that offer highly localized products and services depend heavily on up-to-date demographic data from their immediate surroundings, as small areas can exhibit rapid demographic changes. Conversely, retailers selling specialty goods like motor vehicles and furniture cater to a wider audience. Their consumers are often ready to travel considerable distances to search and compare, providing them with broader trade areas. As a result, localized demographic fluctu-

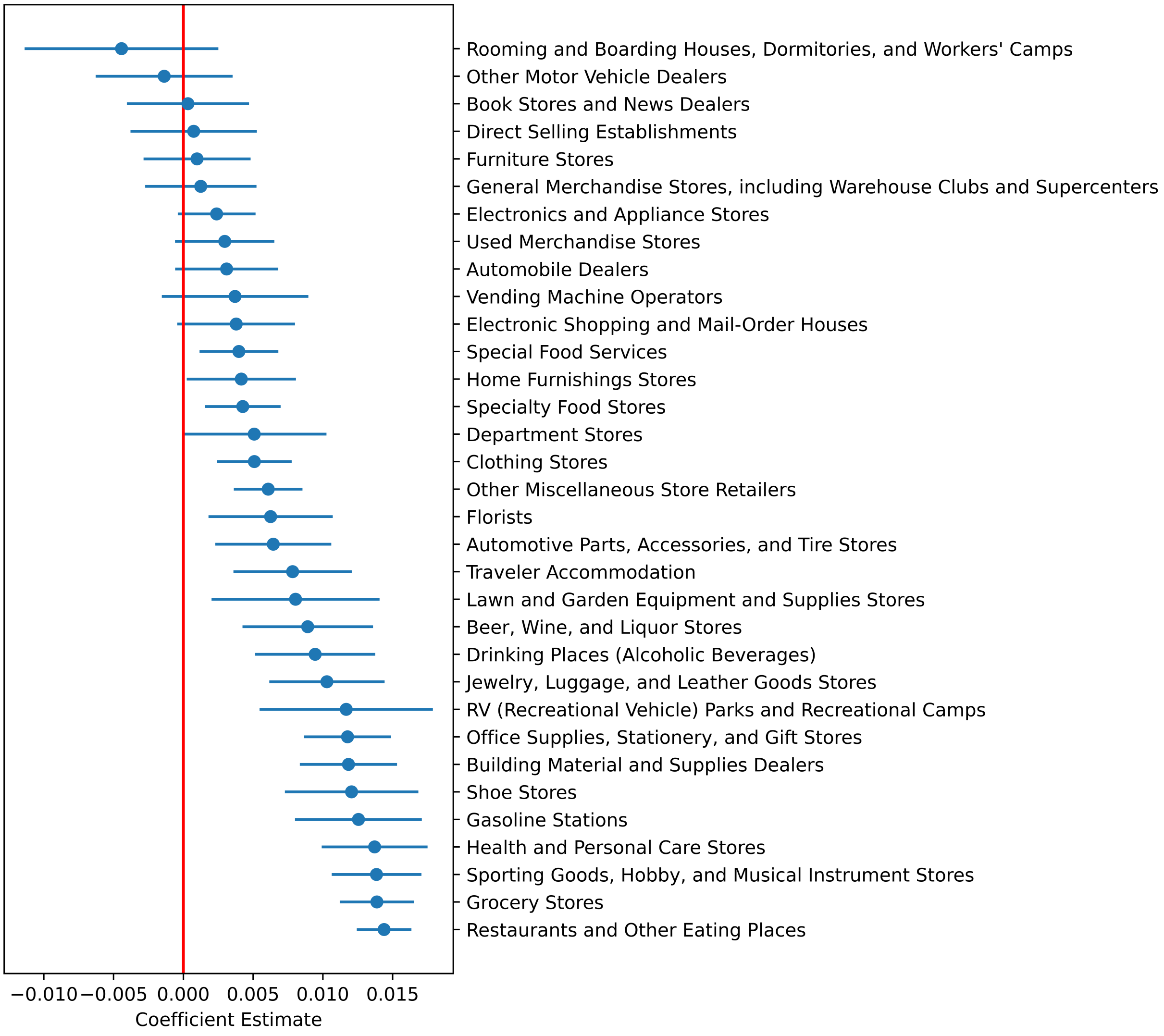


Figure 3: Coefficient Estimates (β_2) by NAICS 4-digit Industry

Notes: This figure plots β_2 coefficients estimated from Equation 4 and the associated 95% confidence intervals for each NAICS 4-digit industry. Standard errors are clustered at the census-tract level. The vertical line indicates where the coefficient estimate is 0.

ations within these expansive regions tend to average out, thereby reducing their dependency on immediate census updates.

To proxy for the size of an establishment's trade area, I categorize retail industries into durable and non-durable goods sectors, in accordance with the Bureau of Economic Analysis' classification of manufacturers of durable and non-durable goods.[35] Durable goods retailers

[35] The Bureau of Economic Analysis (BEA) defines durable goods as those that have a useful life of more than three years. Under this definition, the industries classified into the durable goods sectors include 4411 (Automobile Dealers), 4412 (Other Motor Vehicle Dealers), 4413 (Automotive Parts, Accessories, and Tire

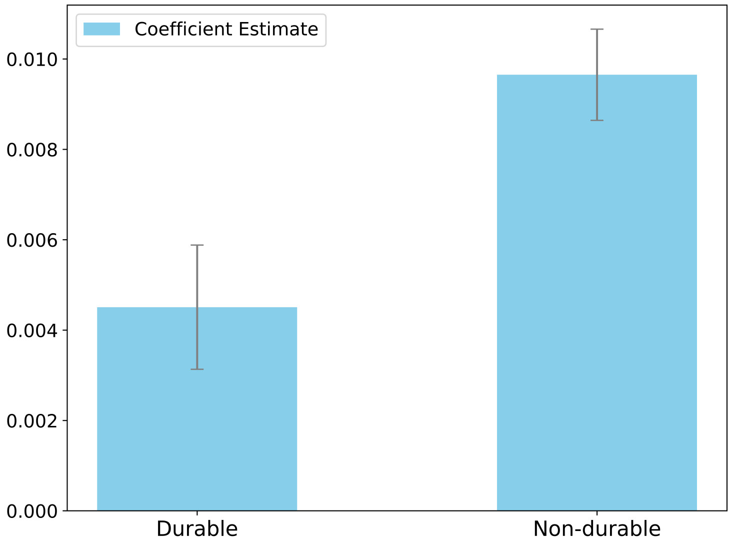


Figure 4: Comparison of Coefficient Estimates (β_2) between Durable and Non-durable Goods Retailers

Notes: This figure plots β_2 coefficient estimates from Equation 4 and associated 95% confidence intervals across durable and non-durable goods retailers. Standard errors are clustered at the census-tract level.

are expected to have larger trade areas, while non-durable goods retailers are likely to have smaller trade areas. As reported in Figure 4, the impact of outdated census data is notably more pronounced for firms in the non-durable goods sector (point estimate 0.0097) than those in the durable goods sector (point estimate 0.0045), which supports my earlier hypothesis regarding trade area size.

As further corroboration, I examine the NAICS 454 Non-store Retailers category, which includes Electronic Shopping and Mail-Order Houses, Vending Machine Operators, and Direct Selling Establishments. Unlike other retailers, non-store retailers do not need a physical store front close to their customers, making them much less reliant on local demographic information. As reported in Figure 3, the impact of census data on failure rate is statistically insignificant for all 3 industries in this category.

Stores), 4421 (Furniture Stores), 4422 (Home Furnishings Stores), 4431 (Electronics and Appliance Stores), 4441 (Building Material and Supplies Dealers), and 4442 (Lawn and Garden Equipment and Supplies Stores). The industries categorized under non-durable goods sectors comprise 4451 (Grocery Stores), 4452 (Specialty Food Stores), 4453 (Beer, Wine, and Liquor Stores), 4461 (Health and Personal Care Stores), 4471 (Gasoline Stations), 4481 (Clothing Stores), 4482 (Shoe Stores), 4483 (Jewelry, Luggage, and Leather Goods Stores), 4511 (Sporting Goods, Hobby, and Musical Instrument Stores), 4512 (Book Stores and News Dealers), 4531 (Florists), and 4532 (Office Supplies, Stationery, and Gift Stores). I exclude 4522 (Department Stores), 4523 (General Merchandise Stores, including Warehouse Clubs and Supercenters), 4533 (Used Merchandise Stores), and 4539 (Other Miscellaneous Store Retailers) were omitted due to their ambiguous nature in strictly categorizing as durable or non-durable goods sectors.

6.3 Failure Rate and Firm Size

This section examines whether the impact of census data varies across firms of different sizes. One notable difference between large and small firms relates to their ability to obtain information. Large firms can collect information from their existing customers; they also have the resources to conduct market surveys and purchase proprietary data from vendors. Although census data forms the foundation of their market research, they can readily turn to alternative sources when census data becomes outdated.

Small firms, on the other hand, often lack these resources. In fact, they may not even be aware that some crucial inputs to their decision-making process originate from the census, and that this data is not always up-to-date. For example, when selecting business locations, entrepreneurs often rely on leasing agents, who in turn utilize census-derived metrics to assess whether the local trade area's demographics are suitable for the proposed business.[36] Both the entrepreneur and the agent might accept this data at face value. Consequently, small firms might be more vulnerable to the effects of outdated census data compared to their larger counterparts.

In the retail sector, the size of a firm can be gauged by the number of its locations. To examine the heterogeneity across firms based on their size, I estimate Equation 4 separately for large chains, small chains, and local independents.[37] Large chains[38] are defined as firms with more than 20 locations. This threshold is consistent with existing regulations related to chains,[39] but the patterns are similar using alternative definitions of large chain.[40] The benchmark failure rates are specific to establishments within each size group located in the same census tract, and standard errors are clustered at the census-tract level.

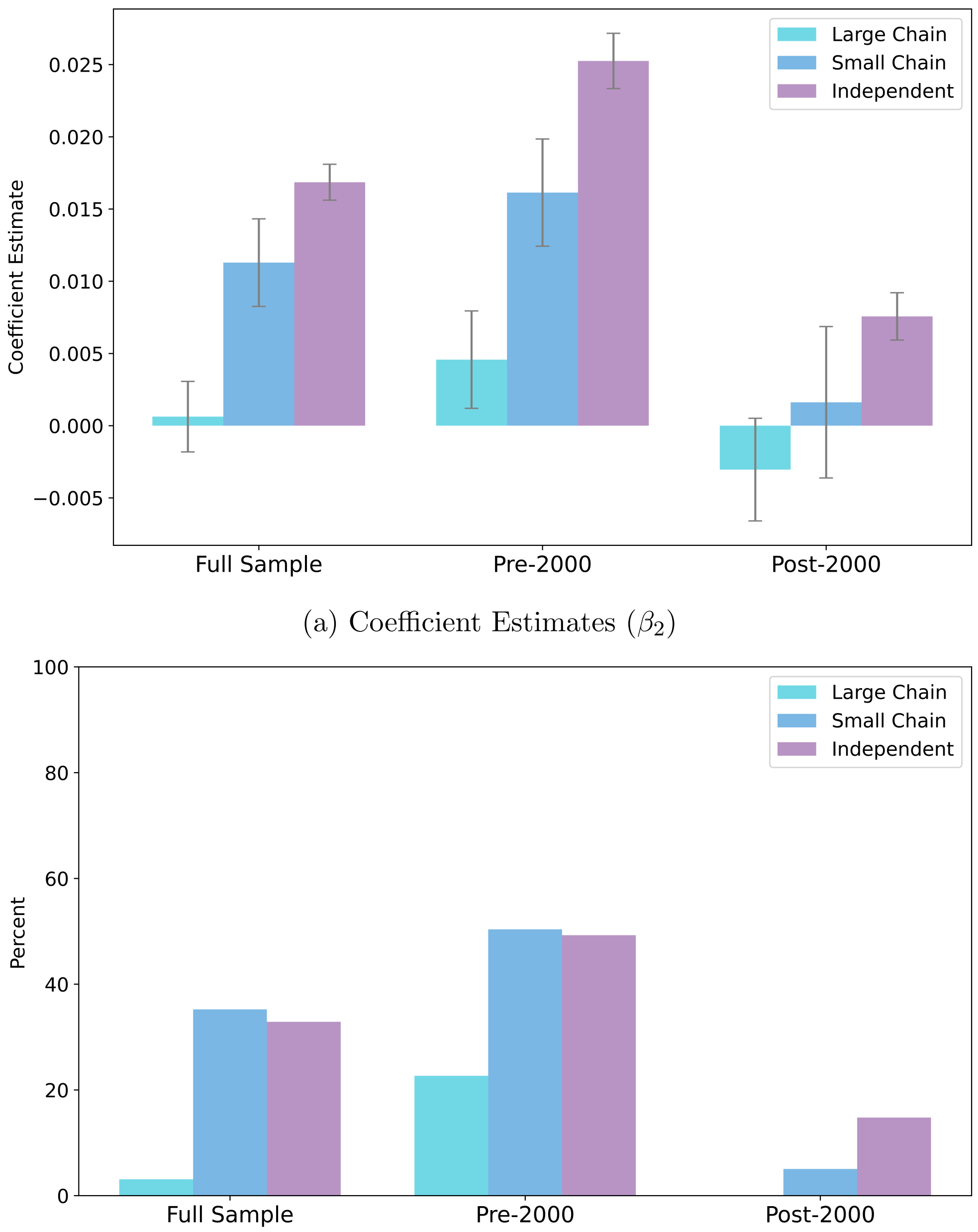
[36] An important factor commercial real estate agents consider when they make recommendations to clients is whether the trade area's demographic profile is suitable to the proposed business. Much like corporate real estate planning departments, agents can narrow down the choices using filters based on demographic information, or use such information to justify the location's value to clients. Figure A1 highlights an example of such databases being directly used by commercial real estate brokers. See https://www.apto.com/blog/the-best-of-commercial-real-estate-data-sources-demographics-broker-databases/ for other examples.

[37] The D&B data groups establishments by ownership. An establishment belonging to a chain might be classified as being owned by a franchisee, yet the parent company (i.e., franchisor) is the one that ultimately makes the site selection decisions (https://www.forbes.com/sites/forbescoachescouncil/2021/06/23/what-franchise-owners-should-know-about-the-site-selection-process/?sh=7e97544732baf). To avoid misclassifying a chain location owned by a franchisee as an independent, I identify chains based on their trade style names in the data set. See Table A8 for details.

[38] Table A1 lists the top 30 chains ranked by the total number of affiliated establishments in the sample.

[39] See, for example, the FDA menu labeling requirements: https://www.ers.usda.gov/amber-waves/2018/october/new-national-menu-labeling-provides-information-consumers-can-use-to-help-manage-their-calorie-intake/.

[40] Please see Table A10 for results setting the threshold at 10 establishments, and Table A11 for results setting the threshold at 50 establishments



(a) Coefficient Estimates (β_2)

(b) Economic Magnitude

Figure 5: The Impact of Outdated Census Data across Firm Size

Notes: This figure reports the impact of outdated census data on failure rate across large chains (more than 20 outlets), small chains (between 2-20 outlets) and independent establishments. Figure 5a reports β_2 coefficients estimated from Equation 4 and the associated 95% confidence intervals. Standard errors are clustered at the census-tract level. Figure 5b illustrates the economic significance of the coefficient estimates, showing how the census data from a decade ago would impact the failure rate as a percentage of each type of firm's average failure rate.

Table A9 reports the full set of coefficient estimates and Figure 5a highlights the β_2 coefficient estimates across firm size. In the full sample, these estimates are positive and statistically significant for small chains and independents. A 10-year gap between two decennial census would result in a 17% increase in failure rate for independents, and 11% for small chains. When adjusted for their respective baseline failure rates, the relative effects are comparable for both groups, as shown in Figure 5b. Large chains, however, are barely affected. This pattern supports the hypothesis that large chains can obtain demographic data from alternative sources in periods distant from census updates, providing them with an advantage over smaller firms.

When comparing estimates across different sub-periods, the patterns become more nuanced. In the pre-2000 period, the observed effects are positive and significant for large chains, small chains and independents. For small chains and independents, increases in failure rate over a 10-year gap are equivalent to 50% of their baseline failure rate. Even for large chains, it's a notable increase of 22%. But in the post-2000 period, these effects moderate for all three groups and become statistically insignificant for chain stores. This shift may be attributed to increased access to alternative data sources for larger firms, facilitated by the advent of new technologies in the 21st century.[41]

7 Alternative Explanations

7.1 Differential Response to Business Cycles

A key assumption of my empirical strategy is that new and existing establishments are influenced by time-varying factors in a similar way, so that the average failure rate of establishments within the same local market can serve as a benchmark to remove potential confounding effects. One might be concerned that the entry cohorts behave differently from existing establishments during different phases of the business cycle.[42] In particular, the early 90s and early 00s recessions overlap with outdated census data in terms of timing, which may drive the pattern. However, the industry breakdown in Figure 3 shows that my results are

[41] One notable example is Synergos Technologies Inc., founded in 2001, that uses postal address data released quarterly, combined with consumer survey data to provide timely and granular demographic data to national and regional companies making strategic location decisions https://synergos-tech.com/popstats/. Its clients include Kroger (grocery), CVS (pharmacy), Chipotle (quick service restaurant), Family Dollar (dollar store) and Simon (real-estate). Mukherjee, Panayotov and Shon (2021) provide a more general example of private data sources substituting less frequently released government macro data.

[42] As the literature (Fort et al., 2013; Pugsley and Șahin, 2019; Sedláček and Sterk, 2017) has documented, employment fluctuations at startups and young firms are pro-cyclical: young firms co-vary more with the overall economy than mature firms. Alternatively, it is also conceivable that startups might be counter-cyclical due to selection: those who still enter despite bad macroeconomic condition are of higher quality.

strongest for restaurants and groceries, despite the marked differences in cyclicality between these two sectors.

To directly test the impact of the early 90s and early 00s recessions, I split the sample based on the severity of the recession in the local area, measured by changes in the county-level unemployment rate before and after the recession.[43] As reported in Table A12, I find similar effects on establishment failure rate (β_2) across sub-samples of local areas that are more/less impacted by the recessions. Therefore, it seems unlikely that these recessions are driving the empirical patterns I document.

7.2 Government Policies

This section explores the extent to which excess failures are impacted by government funding programs tied to census data.[44] Federal funding programs[45] that address the need of households and communities can potentially increase the consumer purchase power in a local market. But this type of program is unlikely to explain differential failure rate between new and existing establishments, especially if the increased spending has similar effects on all businesses in the neighborhood.

More relevant to my empirical context are place-based policies that encourage new business investments with subsidies and tax benefits (Slattery and Zidar, 2020). A prominent policy in the New York State relevant to retail and restaurant businesses is the Economic Development Zones (Empire Zones) Program.[46] To be eligible, a census tract needs to satisfy the following criteria: 1) Poverty rate at or above 20%, 2) Unemployment rate at least 125% of State average, and 3) Population at or above 2,000.[47]

Using these institutional details, I create sub-samples based on a census tract's eligibility for the Economic Development Zones program and estimate Equation 4 separately for these

[43] Based the NBER US recession dates (July 1990 to March 1991 for the early 90s recession and March 2001 to November 2001 for the early 00s recession), I measure the change (percentage change) in same-month unemployment rate before and after the recession (June 1990 to June 1991 for the early 90s recession and February 2001 to February 2002).

[44] According to National Research Council (2003), over $250 billion federal funds allocated to state and local governments are tied to formulas that involve inputs from various data sources, including the decennial census. With few exceptions, the programs allocate funds to the state level.

[45] US Government Accountability Office (2009) reviewed the 10 largest federal assistance programs that relied at least in part on the decennial census and related data. These 10 programs represent 84 percent of total federal assistance for fiscal year 2009, and Medicaid is the largest among the 10, accounting for over half of the share. These programs all target individuals and communities.

[46] Good Jobs First. 1976–2019. "Subsidy Tracker." https://subsidytracker.goodjobsfirst.org/

[47] These criteria were stipulated by legislature in 1986 when the program was created. Over time, conditions were relaxed to the point that almost any area is eligible. The program was shut down in 2010, due to wide criticisms that the zones no longer correspond to distressed areas and that there is a lack of oversight (New York State Office of the State Comptroller, 2004). If the zones are chosen in ways unrelated to census data, the policy should not affect failure rates of new entrants as census data quality varies over time.

two groups. Table A13 reports the results. The coefficient estimate of β_2 for census tracts that are not eligible for Empire Zones funding is very similar to the main result. Note that the estimated effect is actually smaller in eligible census tracts. If the state directs subsidies to areas that are no longer in economic distress, firm exits would likely decrease.[48] Thus, misallocated government funding based on census data is unlikely to offer an alternative explanation for the increase in establishment failure as census data become more outdated.

8 Discussion

Given the significance of census data and its fixed schedule, firms might strategically time their entry to gain access to fresh data. Such strategic timing can introduce a downward bias in the analysis, understating the true effect of outdated census data on firm failure, as firms facing high uncertainty are more likely to wait for better information. This inclination to delay is likely most pronounced just before new census data is released. Considering that the median lifespan of an establishment in my sample is five years, the opportunity cost of delaying entry by even one year is not trivial.

However, the observed trend of increasing failure rates over time is fairly linear, even though the beginning and end of the second phase may be more influenced by shifts in entry timing than the middle. This observation could be due to the unique attributes of the retail and restaurant sectors that diminish the advantage of waiting for improved information. First, these industries have relatively high investment reversibility, exceeding 70% of other industries.[49] When assets can be repurposed or sold off with less loss of value, the benefits of postponing investment in anticipation of more accurate information is significantly reduced. Second, the high turnover in these industries implies a shorter effective duration for investment, making waiting more costly. Lastly, in these highly competitive industries, the benefit of waiting for better data might be outweighed by the strategic advantage of early entry to secure market share.

[48] I obtain similar results using eligibility for Opportunity Zones. To qualify for Opportunity Zones in New York State, a census tract should have individual poverty rate of at least 20%, and the median family income no more than 80% of the state median. The first Opportunity Zones were designated in 2018, after my sample period. However, the requirements reflect what could be used in other state and local policies if they are based on census variables.

[49] Estimates are based on the redeployability measure from (Kim and Kung, 2017), averaged over the sample period.

9 Conclusion

I empirically measures the impact of information quality on firm investment decision. Using establishment-level entry and exit data on firms from the retail (NAICS 44-45) and Accommodation and Food Services (NAICS 72) industries in the New York State during the 30 year period from 1985 to 2014, I document a 1.6% per year increase in 5-year failure rate due to deterioration in information quality as the ever-evolving current market condition diverges from the census data collected at the beginning of the decade.

This study underscores the value of census data that is part of an ongoing policy debate. In countries where population counting is disconnected from political outcomes, census is seen as a symbolic exercise, and many of these countries have abandoned doing actual counts of the population or limit what is available to the public (Donnelly, 2019). [50] This paper points to a channel where lack of information could hurt the local economy, especially small establishments that are dependent on public data sources. Future studies can take this perspective into account when evaluating the welfare implications of census data policies.

Information waves via census data releases might serve as helpful exogenous drivers of firm exit for research about firm productivity (De Loecker and Syverson, 2021) that relies on production function estimation (Olley and Pakes, 1996). One identification issue in this stream of research is selection bias, stemming from non-random exit (i.e., observed production levels of firms are conditional that they are active). Census data timing might help augment existing selection correction methods, especially for multi-country research about productivity dispersion (Asker, Collard-Wexler and De Loecker, 2014), whereby firms in different countries experience different release schedules from their respective censuses.

My research may also complement the development of models for industry dynamics. The work-horse models of industry dynamics (Ericson and Pakes, 1995) that build on the Markov-Perfect Equilibrium (Maskin and Tirole, 1988) framework often rely on stationarity assumptions (i.e., the time of entry in itself is not a state variable). The presence of information waves might motivate future research to explore the sensitivity of model predictions to such assumptions, as the year of entry likely correlates with the information quality at the time of entry decision. In particular, the variance for uncertainty is unlikely to be constant throughout the industry's evolution, which would impact how researchers implement many of the commonly used estimation approaches that typically rely on stationarity in the model.

[50] For others, there is growing debate about the cost and benefit of collecting and reporting census data. As an example, the Canadian government scrapped the mandatory long-form census in 2011 on the grounds that people should not be forced to give detailed information about themselves. The long-form census was reintroduced in the 2016 census that followed. In another example, the most recent 2020 U.S. census has taken measures to protect the identity of respondents but as a side-effect much data at the block group level became unusable (Wines, 2022; Hotz et al., 2022).

References

Adair, Bill. 1991. “Census helps firms target consumers.” *Tampa Bay Times.*

Addoum, Jawad M, David T Ng, and Ariel Ortiz-Bobea. 2020. “Temperature shocks and establishment sales.” *The Review of Financial Studies*, 33(3): 1331–1366.

Adelino, Manuel, Song Ma, and David Robinson. 2017. “Firm age, investment opportunities, and job creation.” *The Journal of Finance*, 72(3): 999–1038.

Arcidiacono, Peter, Patrick Bayer, Jason R Blevins, and Paul B Ellickson. 2016. “Estimation of dynamic discrete choice models in continuous time with an application to retail competition.” *The Review of Economic Studies*, 83(3): 889–931.

Armas, Genaro C. 2001. “Census Data Big for Businesses.” *Associated Press.*

Asker, John, Allan Collard-Wexler, and Jan De Loecker. 2014. “Dynamic inputs and resource (mis) allocation.” *Journal of Political Economy*, 122(5): 1013–1063.

Asplund, Marcus, and Volker Nocke. 2006. “Firm turnover in imperfectly competitive markets.” *The Review of Economic Studies*, 73(2): 295–327.

Barnatchez, Keith, Leland Dod Crane, and Ryan Decker. 2017. “An assessment of the national establishment time series (nets) database.” FEDS working paper.

Berry, Steven T, and Giovanni Compiani. 2021. “Empirical models of industry dynamics with endogenous market structure.” *Annual Review of Economics*, 13: 309–334.

Binz, Oliver, William J Mayew, and Suresh Nallareddy. 2021. “Firms’ response to macroeconomic estimation errors.” *Journal of Accounting and Economics*, 101454.

Bloom, Nicholas. 2009. “The impact of uncertainty shocks.” *Econometrica*, 77(3): 623–685.

Bloom, Nick, Stephen Bond, and John Van Reenen. 2007. “Uncertainty and investment dynamics.” *The Review of Economic Studies*, 74(2): 391–415.

Collard-Wexler, Allan. 2013. “Demand fluctuations in the ready-mix concrete industry.” *Econometrica*, 81(3): 1003–1037.

Craft, Erik D. 1998. “The value of weather information services for nineteenth-century Great Lakes shipping.” *American Economic Review*, 1059–1076.

Cropper, Matthew, Jerome N. McKibben, David A. Swanson, and Jeff Tayman. 2012. “Vendor Accuracy Study 2010 Estimates versus Census 2010.” Esri.

Currie, Janet, Stefano DellaVigna, Enrico Moretti, and Vikram Pathania. 2010. “The effect of fast food restaurants on obesity and weight gain.” *American Economic Journal: Economic Policy*, 2(3): 32–63.

De Loecker, Jan, and Chad Syverson. 2021. "An industrial organization perspective on productivity." In *Handbook of Industrial Organization.* Vol. 4, 141–223. Elsevier.

De Loecker, Jan, and Jan Eeckhout. 2018. "Global market power." National Bureau of Economic Research.

Donnelly, Frank. 2019. *Exploring the US Census: Your Guide to America's Data.* SAGE Publications.

Dunne, Timothy, Mark J Roberts, and Larry Samuelson. 1989. "The growth and failure of US manufacturing plants." *The Quarterly Journal of Economics*, 104(4): 671–698.

Ericson, Richard, and Ariel Pakes. 1995. "Markov-perfect industry dynamics: A framework for empirical work." *The Review of Economic Studies*, 62(1): 53–82.

Fang, Limin, and Nathan Yang. 2022. "Measuring Deterrence Motives in Dynamic Oligopoly Games." *Management Science*, forthcoming.

Fan, Ying, and Mo Xiao. 2015. "Competition and subsidies in the deregulated US local telephone industry." *the RAND Journal of Economics*, 46(4): 751–776.

Farboodi, Maryam, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran. 2022. "Where has all the data gone?" *The Review of Financial Studies*, 35(7): 3101–3138.

Farboodi, Maryam, and Laura Veldkamp. 2021. "A growth model of the data economy." National Bureau of Economic Research.

Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp. 2019. "Big data and firm dynamics." *AEA Papers and Proceedings*, 109: 38–42.

Farhi, Paul. 1990. "For Business, Census is a Marketing Data Motherlode." *The Washington Post.* (accessed April 19, 2022).

Fort, Teresa C, John Haltiwanger, Ron S Jarmin, and Javier Miranda. 2013. "How firms respond to business cycles: The role of firm age and firm size." *IMF Economic Review*, 61(3): 520–559.

Gao, Meng, and Jiekun Huang. 2020. "Informing the market: The effect of modern information technologies on information production." *The Review of Financial Studies*, 33(4): 1367–1411.

Goldfarb, Avi, and Mo Xiao. 2016. "Transitory shocks, limited attention, and a firm's decision to exit." Working paper.

Hollenbeck, Brett. 2017. "The economic advantages of chain organization." *The RAND Journal of Economics*, 48(4): 1103–1135.

Hopenhayn, Hugo A. 1992. "Entry, exit, and firm dynamics in long run equilibrium." *Econometrica*, 1127–1150.

Hotz, V Joseph, Christopher R Bollinger, Tatiana Komarova, Charles F Manski, Robert A Moffitt, Denis Nekipelov, Aaron Sojourner, and Bruce D Spencer. 2022. "Balancing data privacy and usability in the federal statistical system." *Proceedings of the National Academy of Sciences*, 119(31): e2104906119.

Igami, Mitsuru, and Nathan Yang. 2016. "Unobserved heterogeneity in dynamic games: Cannibalization and preemptive entry of hamburger chains in Canada." *Quantitative Economics*, 7(2): 483–521.

Jens, Candace E. 2017. "Political uncertainty and investment: Causal evidence from US gubernatorial elections." *Journal of Financial Economics*, 124(3): 563–579.

Jeon, Jihye. 2022. "Learning and investment under demand uncertainty in container shipping." *The RAND Journal of Economics*, 53(1): 226–259.

Jones, Charles I, and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review*, 110(9): 2819–58.

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry." *Econometrica*, 649–670.

Julio, Brandon, and Youngsuk Yook. 2012. "Political uncertainty and corporate investment cycles." *The Journal of Finance*, 67(1): 45–83.

Kellogg, Ryan. 2014. "The effect of uncertainty on investment: Evidence from Texas oil drilling." *American Economic Review*, 104(6): 1698–1734.

Kim, Hyunseob, and Howard Kung. 2017. "The asset redeployability channel: How uncertainty affects corporate investment." *The Review of Financial Studies*, 30(1): 245–280.

Kolko, Jed. 2012. "Broadband and local growth." *Journal of Urban Economics*, 71(1): 100–113.

Kosová, Renáta, and Francine Lafontaine. 2010. "Survival and growth in retail and service industries: Evidence from franchised chains." *The Journal of Industrial Economics*, 58(3): 542–578.

Kumar, Pradeep, and Hongsong Zhang. 2019. "Productivity or Unexpected Demand Shocks: What Determines Firms' Investment and Exit Decisions?" *International Economic Review*, 60(1): 303–327.

Leisten, Matthew. 2021. "Information, managerial incentives, and scale: Evidence from hotel pricing."

Levine, David I, Michael W Toffel, and Matthew S Johnson. 2012. "Randomized government safety inspections reduce worker injuries with no detectable job loss." *Science*, 336(6083): 907–911.

Logan, John R, Zengwang Xu, and Brian J Stults. 2014. "Interpolating US decennial census tract data from as early as 1970 to 2010: A longitudinal tract database." *The Professional Geographer*, 66(3): 412–420.

Maican, Florin G, and Matilda Orth. 2018. "Entry regulations, welfare, and determinants of market structure." *International Economic Review*, 59(2): 727–756.

Maskin, Eric, and Jean Tirole. 1988. "A theory of dynamic oligopoly, II: Price competition, kinked demand curves, and Edgeworth cycles." *Econometrica*, 571–599.

Mian, Atif, and Amir Sufi. 2014. "What explains the 2007–2009 drop in employment?" *Econometrica*, 82(6): 2197–2223.

Mukherjee, Abhiroop, George Panayotov, and Janghoon Shon. 2021. "Eye in the sky: Private satellites and government macro data." *Journal of Financial Economics*, 141(1): 234–254.

Nagaraj, Abhishek. 2022. "The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry." *Management Science*, 68(1): 564–582.

National Research Council. 1995. *Modernizing the U.S. Census.* The National Academies Press.

National Research Council. 2003. *Statistical issues in allocating funds by formula.* National Academies Press.

National Research Council. 2015. *Realizing the potential of the American Community Survey: Challenges, tradeoffs, and opportunities.* National Academies Press.

Neumark, David, Brandon Wall, and Junfu Zhang. 2011. "Do small businesses create more jobs? New evidence for the United States from the National Establishment Time Series." *The Review of Economics and Statistics*, 93(1): 16–29.

New York State Office of the State Comptroller. 2004. "Assessing the Empire Zones Program Reforms Needed to Improve Program Evaluation and Effectiveness." *https://web.osc.state.ny.us/osdc/empirezone3-2005.pdf.*

Olley, GS, and A Pakes. 1996. "The dynamics of productivity in the telecommunications equipment industry." *Econometrica*, 64(6): 1263–1297.

Pugsley, Benjamin Wild, and Ayșegül Șahin. 2019. "Grown-up business cycles." *The Review of Financial Studies*, 32(3): 1102–1147.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2021. "Diverging trends in national and local concentration." *NBER Macroeconomics Annual*, 35(1): 115–150.

Schuetz, Jenny, Jed Kolko, and Rachel Meltzer. 2012. "Are poor neighborhoods "retail deserts"?" *Regional Science and Urban Economics*, 42(1-2): 269–285.

Sedláček, Petr, and Vincent Sterk. 2017. "The growth potential of startups over the business cycle." *American Economic Review*, 107(10): 3182–3210.

Serrato, Juan Carlos Suárez, and Philippe Wingender. 2016. "Estimating local fiscal multipliers." National Bureau of Economic Research.

Slattery, Cailin, and Owen Zidar. 2020. "Evaluating state and local business incentives." *Journal of Economic Perspectives*, 34(2): 90–118.

Suzuki, Junichi. 2013. "Land use regulation as a barrier to entry: evidence from the Texas lodging industry." *International Economic Review*, 54(2): 495–523.

Thau, Barbara. 2014. "How Big Data Helps Chains Like Starbucks Pick Store Locations – An (Unsung) Key To Retail Success." *Forbes.* (accessed April 19, 2022).

The Council of Economic Advisers. 2000. "The Uses of Census Data: An Analytical Review." The Council of Economic Advisers.

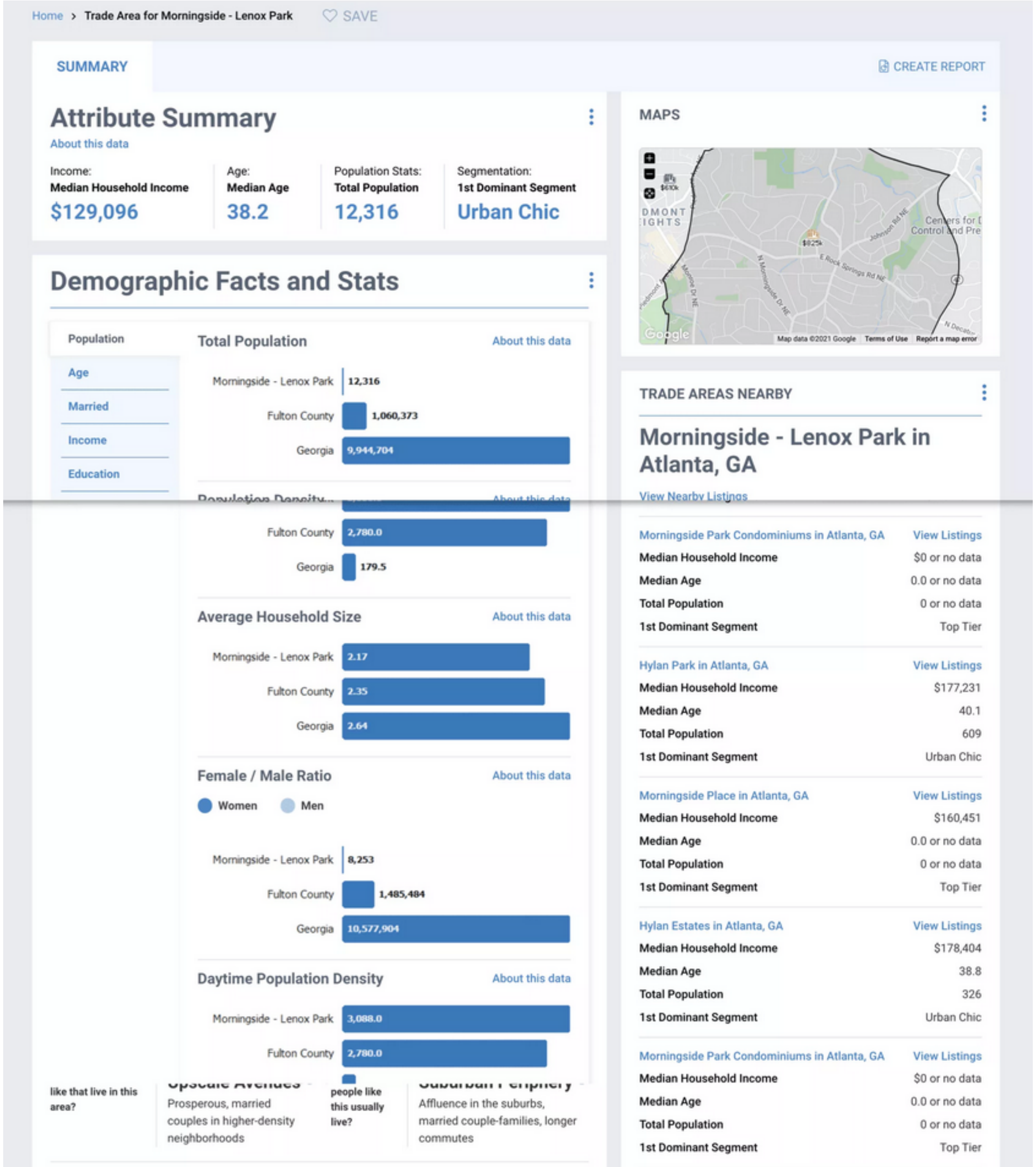
Thomadsen, Raphael. 2005. "The effect of ownership structure on prices in geographically differentiated industries." *RAND Journal of Economics*, 908–929.

Tsui, Jennifer, Jana A Hirsch, Felicia J Bayer, James W Quinn, Jesse Cahill, David Siscovick, and Gina S Lovasi. 2020. "Patterns in geographic access to health care facilities across neighborhoods in the United States based on data from the national establishment time-series between 2000 and 2014." *JAMA network open*, 3(5): e205105–e205105.

US Government Accountability Office. 2009. "Formula Grants: Funding for the Largest Federal Assistance Programs is Based on Census-Related Data and Other Factors. (GAO Publication No. 10-263)." *Washington, D.C.: U.S. Government Printing Office.*

Wines, Michael. 2022. "The 2020 Census Suggests That People Live Underwater. There's a Reason." *The New York Times.* (accessed April 22, 2022).

A Online Appendix



Home > Trade Area for Morningside - Lenox Park ♡ SAVE

SUMMARY

CREATE REPORT

Attribute Summary

About this data

Income: Median Household Income	Age: Median Age	Population Stats: Total Population	Segmentation: 1st Dominant Segment
$129,096	38.2	12,316	Urban Chic

Demographic Facts and Stats

Population

Age

Married

Income

Education

Total Population — About this data

Morningside - Lenox Park	12,316
Fulton County	1,060,373
Georgia	9,944,704

Population Density... — About this data

Fulton County	2,780.0
Georgia	179.5

Average Household Size — About this data

Morningside - Lenox Park	2.17
Fulton County	2.35
Georgia	2.64

Female / Male Ratio — About this data

● Women ● Men

Morningside - Lenox Park	8,253
Fulton County	1,485,484
Georgia	10,577,904

Daytime Population Density — About this data

Morningside - Lenox Park	3,088.0
Fulton County	2,780.0

like that live in this area?

Upscale Avenues

Prosperous, married couples in higher-density neighborhoods

people like this usually live?

Suburban Periphery

Affluence in the suburbs, married couple-families, longer commutes

MAPS

TRADE AREAS NEARBY

Morningside - Lenox Park in Atlanta, GA

View Nearby Listings

Morningside Park Condominiums in Atlanta, GA	View Listings
Median Household Income	$0 or no data
Median Age	0.0 or no data
Total Population	0 or no data
1st Dominant Segment	Top Tier

Hylan Park in Atlanta, GA	View Listings
Median Household Income	$177,231
Median Age	40.1
Total Population	609
1st Dominant Segment	Urban Chic

Morningside Place in Atlanta, GA	View Listings
Median Household Income	$160,451
Median Age	0.0 or no data
Total Population	0 or no data
1st Dominant Segment	Top Tier

Hylan Estates in Atlanta, GA	View Listings
Median Household Income	$178,404
Median Age	38.8
Total Population	326
1st Dominant Segment	Urban Chic

Morningside Park Condominiums in Atlanta, GA	View Listings
Median Household Income	$0 or no data
Median Age	0.0 or no data
Total Population	0 or no data
1st Dominant Segment	Top Tier

Figure A1: Example Trade Area Analysis Report Tool For Commercial Real Estate Brokers

Notes: The figure provides an example of the demographic data used by commercial real estate brokers that market commercial properties to potential tenants. This report is an illustration of the data included in the Realtors Property Resource database provided by the National Association of Realtors. Source: https://blog.narrpr.com/tips/commercial-trade-area-details/.

Table A1: Number of Establishments by Industry

NAICS4		Counts
4411	Automobile Dealers	23,510
4412	Other Motor Vehicle Dealers	7,947
4413	Automotive Parts, Accessories, and Tire Stores	13,945
4421	Furniture Stores	17,528
4422	Home Furnishings Stores	18,590
4431	Electronics and Appliance Stores	36,262
4441	Building Material and Supplies Dealers	26,716
4442	Lawn and Garden Equipment and Supplies Stores	4,710
4451	Grocery Stores	71,505
4452	Specialty Food Stores	43,795
4453	Beer, Wine, and Liquor Stores	10,775
4461	Health and Personal Care Stores	20,675
4471	Gasoline Stations	18,522
4481	Clothing Stores	77,630
4482	Shoe Stores	12,895
4483	Jewelry, Luggage, and Leather Goods Stores	20,778
4511	Sporting Goods, Hobby, and Musical Instrument Stores	33,847
4512	Book Stores and News Dealers	8,905
4522	Department Stores	5,763
4523	General Merchandise Stores, including Warehouse Clubs and Supercenters	13,432
4531	Florists	10,790
4532	Office Supplies, Stationery, and Gift Stores	34,582
4533	Used Merchandise Stores	14,306
4539	Other Miscellaneous Store Retailers	71,371
4541	Electronic Shopping and Mail-Order Houses	7,686
4542	Vending Machine Operators	3,998
4543	Direct Selling Establishments	11,104
7211	Traveler Accommodation	16,679
7212	RV (Recreational Vehicle) Parks and Recreational Camps	3,117
7213	Rooming and Boarding Houses, Dormitories, and Workers' Camps	505
7223	Special Food Services	26,083
7224	Drinking Places (Alcoholic Beverages)	24,693
7225	Restaurants and Other Eating Places	185,588

Table A2: Demographic Variable Definition

Variable	Definition
Population	Total population
%Kids (0-17)	Persons under 18 years old as a percentage of Total population
%Young (18-34)	Persons 18 to 34 years old as a percentage of Total population
%Middle (35-64)	Persons 35 to 64 years old as a percentage of Total population
%Old (65+)	Persons 65 years old and over as a percentage of Total population
%White	White as a percentage of total population
%Black	Black as a percentage of total population
%Asian	Asian or Pacific Islander as a percentage of total population
%Latino	Hispanics not self identified as White, Black, or Asian (Others)
%College degree	Percentage of Bachelor's degree or more among Persons 25 years and over
Unemployment rate	Percentage of Employed among Civilian Population In Labor Force 16 Years And Over
Median income	Median Household Income in 2010 dollars
Median house value	Median House Value for Specified Owner-Occupied Housing Units in 2010 dollars

Table A3: Excess Failure Rate by Entry-Year

	(1)
1986	0.042***
	(0.004)
1987	0.202***
	(0.005)
1988	0.029***
	(0.006)
1989	0.169***
	(0.004)
1990	0.172***
	(0.011)
1991	-0.047***
	(0.004)
1992	-0.013*
	(0.007)
1993	-0.050***
	(0.008)
1994	0.036***
	(0.005)
1995	-0.004
	(0.006)
1996	0.109***
	(0.005)
1997	0.102***
	(0.005)
1998	0.129***
	(0.006)
1999	0.077***
	(0.006)
2000	0.139***
	(0.005)
2001	0.229***
	(0.004)
2002	0.196***
	(0.004)
2003	0.034***
	(0.005)
2004	0.051***
	(0.005)
2005	0.055***
	(0.005)
2006	0.041***
	(0.006)
2007	0.045***
	(0.006)
2008	-0.024***
	(0.006)
2009	0.063***
	(0.007)
Observations	100433
R^2	0.082

Table A4: Excess Failure Rate and Distance to Census Data Release: NYC vs Others

	(1) NYC Census Tracts	(2) Non-NYC Census Tracts
β_1	0.041***	-0.020***
	(0.007)	(0.007)
β_2	0.017***	0.016***
	(0.001)	(0.001)
α_1	0.142***	0.076***
	(0.012)	(0.012)
α_2	-0.017***	0.029***
	(0.004)	(0.003)
Observations	41743	58690
R^2	0.009	0.008

Notes: This table reports coefficient estimates from Equation 4 using the subsamples of census tracts within (outside) of the New York City. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A5: Excess Failure Rate and Distance to Census Data Release using Alternative Clustering Approach

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-0.046***	-0.046***	-0.046	0.006	0.006	0.006
	(0.004)	(0.013)	(0.047)	(0.010)	(0.032)	(0.125)
β_2	0.015***	0.015***	0.015**	0.016***	0.016***	0.016***
	(0.001)	(0.002)	(0.005)	(0.001)	(0.002)	(0.004)
Wild bootstrap p-value			0.020			0.013
α_1				0.095***	0.095**	0.095
				(0.019)	(0.044)	(0.181)
α_2				0.010	0.010*	0.010
				(0.008)	(0.006)	(0.016)
Constant	0.017***	0.017***	0.017			
	(0.006)	(0.006)	(0.021)			
Observations	100433	100433	100433	100433	100433	100433
R^2	0.007	0.007	0.007	0.008	0.008	0.008

Notes: This table summarizes my robustness check of the main test using alternative ways of clustering standard errors. In Columns (1)-(3), the two segments connect at the breakpoint. In Columns (4)-(6), the two segments have separate slopes and intercepts. Parentheses contain standard errors. In Columns (1) and (4) standard errors are clustered at the county level. In Columns (2) and (5) standard errors are two-way clustered at the census-tract and county-and-entry-year level. In Columns (3) and (6) standard errors are two-way clustered at the census-tract and entry-year level. Given the relatively small number of entry-year clusters, p-values from wild bootstrap tests with two-way clustering are reported for β_2. Significance: *** $p<0.001$, ** $p<0.01$, * $p<0.05$.

Table A6: Excess Failure Rate and Distance to Census Data Release using Alternative Benchmark that Excludes the Entry Cohort

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-0.051***	0.026***	-0.129***	0.007	0.053***	-0.051***
	(0.002)	(0.004)	(0.003)	(0.006)	(0.010)	(0.007)
β_2	0.017***	0.027***	0.006***	0.019***	0.028***	0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
α_1				0.118***	0.040**	0.169***
				(0.010)	(0.016)	(0.011)
α_2				0.011***	-0.008**	0.026***
				(0.003)	(0.004)	(0.004)
Constant	0.020***	-0.005	0.038***			
	(0.003)	(0.004)	(0.003)			
Observations	100433	56299	44134	100433	56299	44134
R^2	0.007	0.028	0.038	0.008	0.028	0.041

Notes: This table summarizes my sensitivity analysis of the benchmark failure rate using other establishments (excluding the entry cohort of interest) in the same census tract to measure the market failure rate of a given calendar year. In Columns (1) - (3), the two segments connect at the break-point. In Columns (4) - (6), the two segments have separate slopes and intercepts. Columns (1) and (3) use the full sample periods of entry cohorts from 1986 to 2009. Column (2) and (4) use the sub-sample of entry cohorts before 2000. Columns (3) and (5) use the sub-sample of entry cohorts starting from 2000. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A7: Sensitivity Check for Changes in Demographic Variables Cutoffs

Demographic Variable	Cutoff for Changes in Demographic Variable					
	<-20%	<-15%	<-10%	>10%	>15%	>20%
Population	0.001	0.001	0.001	0.003**	0.003	0.003
	(0.004)	(0.003)	(0.002)	(0.001)	(0.002)	(0.002)
%Kids(0-17)	0.000	-0.002	-0.005***	0.005***	0.004*	0.002
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
%Young (18-34)	0.008***	0.007***	0.007***	-0.007***	-0.006*	-0.006
	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.004)
%Middle (35-64)	-0.002	0.006	0.007*	0.007***	0.009***	0.008***
	(0.008)	(0.006)	(0.004)	(0.001)	(0.002)	(0.002)
%Old (65+)	0.001	-0.002	-0.003*	0.001	0.002	0.002
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
%White	0.001	0.001	0.003*	-0.011***	-0.011***	-0.012***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)	(0.003)
%Black	0.002	0.002	0.002	0.002	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
%Asian	-0.001	-0.001	-0.001	0.002	0.003**	0.004***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
%Latino	0.000	-0.001	-0.001	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
%College	0.005**	0.005***	0.006***	-0.006***	-0.005***	-0.004***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Unemployment rate	-0.001	-0.001	-0.001	-0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Median income	-0.014***	-0.013***	-0.012***	0.008***	0.006***	0.007***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Median house value	0.002	0.004**	0.003**	-0.004***	-0.004***	-0.004***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Notes: This table provides a summary of the sensitivity analysis for the incremental effects of demographic shifts on failure rate using different cutoffs to define a large change. The γ_2 coefficients and associated standard errors are estimated from the regression Equation 6 for each demographic variable with various cutoffs. Standard errors are clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A8: Top 30 Chains by Number of Establishments in the Sample

Chain	Number of establishments
SUBWAY	1586
MCDONALDS	1323
DUNKIN DONUTS	1278
RITE AID	1105
CVS	803
SEVEN ELEVEN	718
BURGER KING	672
RADIO SHACK	633
MOBIL	587
ECKERD	566
SUNOCO	536
STARBUCKS	524
A & P	471
PAYLESS SHOE	430
WENDYS	428
FAMILY DOLLAR	410
STEWARTS	374
BASKINROBBINS	366
KFC	365
DOMINOS PIZZA	344
CARVEL ICE CREAM	342
DOLLAR GENERAL	333
PIZZA HUT	319
GNC	311
TIM HORTONS	308
DUANE READE	295
GETTY	290
WALGREENS	288
EXXON	274
HOLIDAY INN	267

Notes: To identify locations associate with a brand, I first standardize the trade style names by cleaning up the text strings. First, I remove numbers, special characters, and common indicators for a branch such as ”STORE”, ”RESTAURANT”, and ”REST”. Then I manually go through the top 100 brands in terms of the number of locations and use combination of keywords to identify variants of the brand name. KFC, for example, is sometimes recorded under ”KENTUCKY FRIED CHICKEN”, ”K F C”, or ”KENTUCKY FRD CHICKEN”. Finally, I combine locations with various alternative names under the same standardized brand name. This table lists the top 30 brands by number of establishments after cleaning the sample.

Table A9: Chain vs Independent

Panel A: Full Sample

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.017	-0.016	0.005
	(0.013)	(0.014)	(0.005)
β_2	0.001	0.011***	0.017***
	(0.001)	(0.002)	(0.001)
α_1	-0.008	-0.014	0.100***
	(0.022)	(0.024)	(0.009)
α_2	-0.021***	-0.025***	0.009***
	(0.005)	(0.006)	(0.003)
Observations	19045	17912	99081
R^2	0.002	0.003	0.008

Panel B: 1986-1999 Entry Cohorts

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.023	-0.036*	0.037***
	(0.018)	(0.016)	(0.009)
β_2	0.005**	0.016***	0.025***
	(0.002)	(0.002)	(0.001)
α_1	-0.020	-0.061*	0.020
	(0.031)	(0.028)	(0.015)
α_2	-0.046***	-0.041***	-0.004
	(0.007)	(0.008)	(0.004)
Observations	11529	14143	55239
R^2	0.004	0.006	0.028

Panel C: 2000-2009 Entry Cohorts

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.012	0.034	-0.040***
	(0.018)	(0.025)	(0.006)
β_2	-0.003	0.002	0.008***
	(0.002)	(0.003)	(0.001)
α_1	0.005	0.105*	0.151***
	(0.031)	(0.044)	(0.010)
α_2	0.009	0.006	0.018***
	(0.007)	(0.013)	(0.003)
Observations	7516	3769	43842
R^2	0.001	0.002	0.036

Notes: This table summarizes the tests for the relationship between excess failure rate and an entry cohort's distance to census data release separately for large chains (more than 20 outlets), small chains (between 2-20 outlets) and independent establishments. Column (1) reports results on large chains, Column (2) reports results on small chains, and Column (3) reports results on independent establishments. Panel A uses the full sample, Panel B uses the pre-2000 sample and Panel uses the post-2000 sample. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A10: Chain vs Independent using Alternative Definition of Large Chain (10 Locations)

Panel A: Full Sample

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.020	-0.010	0.005
	(0.012)	(0.015)	(0.005)
β_2	0.002	0.011^{***}	0.017^{***}
	(0.001)	(0.002)	(0.001)
α_1	-0.010	-0.006	0.100^{***}
	(0.021)	(0.025)	(0.009)
α_2	-0.026^{***}	-0.021^{**}	0.009^{***}
	(0.005)	(0.007)	(0.003)
Observations	20726	16060	99081
R^2	0.003	0.003	0.008

Panel B: 1986-1999 Entry Cohorts

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.031	-0.030	0.037^{***}
	(0.016)	(0.017)	(0.009)
β_2	0.006^{***}	0.015^{***}	0.025^{***}
	(0.002)	(0.002)	(0.001)
α_1	-0.032	-0.053	0.020
	(0.028)	(0.030)	(0.015)
α_2	-0.050^{***}	-0.036^{***}	-0.004
	(0.006)	(0.008)	(0.004)
Observations	12816	12876	55239
R^2	0.005	0.005	0.028

Panel C: 2000-2009 Entry Cohorts

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.007	0.040	-0.040^{***}
	(0.018)	(0.027)	(0.006)
β_2	-0.002	0.001	0.008^{***}
	(0.002)	(0.003)	(0.001)
α_1	0.014	0.117^{*}	0.151^{***}
	(0.030)	(0.047)	(0.010)
α_2	0.006	0.006	0.018^{***}
	(0.007)	(0.014)	(0.003)
Observations	7910	3184	43842
R^2	0.001	0.003	0.036

Notes: This table summarizes the tests for the relationship between excess failure rate and an entry cohort's distance to census data release separately for large chains (more than 10 outlets), small chains (between 2-10 outlets) and independent establishments. Column (1) reports results on large chains, Column (2) reports results on small chains, and Column (3) reports results on independent establishments. Panel A uses the full sample, Panel B uses the pre-2000 sample and Panel uses the post-2000 sample. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A11: Chain vs Independent using Alternative Definition of Large Chain (50 Locations)

Panel A: Full Sample

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.011	-0.022	0.005
	(0.013)	(0.013)	(0.005)
β_2	0.000	0.011***	0.017***
	(0.001)	(0.001)	(0.001)
α_1	-0.007	-0.017	0.100***
	(0.023)	(0.023)	(0.009)
α_2	-0.017***	-0.033***	0.009***
	(0.005)	(0.006)	(0.003)
Observations	16910	20093	99081
R^2	0.001	0.003	0.008

Panel B: 1986-1999 Entry Cohorts

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.021	-0.032*	0.037***
	(0.019)	(0.016)	(0.009)
β_2	0.004*	0.017***	0.025***
	(0.002)	(0.002)	(0.001)
α_1	-0.020	-0.051	0.020
	(0.032)	(0.027)	(0.015)
α_2	-0.041***	-0.051***	-0.004
	(0.007)	(0.007)	(0.004)
Observations	10153	15312	55239
R^2	0.003	0.006	0.028

Panel C: 2000-2009 Entry Cohorts

	Large Chain (1)	Small Chain (2)	Independent (3)
β_1	-0.001	-0.000	-0.040***
	(0.018)	(0.023)	(0.006)
β_2	-0.003	0.001	0.008***
	(0.002)	(0.002)	(0.001)
α_1	0.008	0.057	0.151***
	(0.031)	(0.039)	(0.010)
α_2	0.012	-0.004	0.018***
	(0.008)	(0.011)	(0.003)
Observations	6757	4781	43842
R^2	0.000	0.004	0.036

Notes: This table summarizes the tests for the relationship between excess failure rate and an entry cohort's distance to census data release separately for large chains (more than 50 outlets), small chains (between 2-50 outlets) and independent establishments. Column (1) reports results on large chains, Column (2) reports results on small chains, and Column (3) reports results on independent establishments. Panel A uses the full sample, Panel B uses the pre-2000 sample and Panel uses the post-2000 sample. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A12: Subsample Analysis: Severity of Local Recessions

	Early 90s Recession		Early 00s Recession	
	Below Median	Above Median	Below Median	Above Median
	(1)	(2)	(3)	(4)
β_1	0.023**	-0.018*	0.008	0.004
	(0.007)	(0.008)	(0.008)	(0.008)
β_2	0.016***	0.016***	0.015***	0.019***
	(0.001)	(0.001)	(0.001)	(0.001)
α_1	0.119***	0.083***	0.103***	0.107***
	(0.012)	(0.013)	(0.012)	(0.013)
α_2	-0.002	0.026***	0.020***	-0.002
	(0.003)	(0.004)	(0.004)	(0.004)
Observations	58594	41839	54805	45628
R^2	0.008	0.009	0.006	0.011

Notes: This table summarizes replication results of the main specifications on subsamples of census tracts with different impact of recessions. The severity of each recession is measured by the percentage change in county-level same-month unemployment rate before and after the recession (June 1990 to June 1991 for the early 90s recession and February 2001 to February 2002). The county-level unemployment rate series are from the Bureau of Labor Statistics (https://www.bls.gov/lau/). The recession dates are from NBER (https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions). Column (1) - (2) are based on the early 1990s recession. Columns (3) - (4) are based on the 2000s recession. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A13: Subsample Analysis: Eligibility for Subsidy

	Eligible for Empire Zones		Eligible for Opportunity Zones	
	No	Yes	No	Yes
	(1)	(2)	(3)	(4)
β_1	-0.001	0.045^{***}	-0.006	0.052^{***}
	(0.006)	(0.013)	(0.006)	(0.011)
β_2	0.017^{***}	0.014^{***}	0.017^{***}	0.013^{***}
	(0.001)	(0.002)	(0.001)	(0.001)
α_1	0.098^{***}	0.140^{***}	0.089^{***}	0.163^{***}
	(0.009)	(0.021)	(0.010)	(0.018)
α_2	0.010^{***}	0.010	0.010^{***}	0.010
	(0.003)	(0.006)	(0.003)	(0.005)
Observations	84139	16294	80183	20250
R^2	0.009	0.006	0.009	0.006

Notes: This table summarizes replication results of the main specifications on subsamples of census tracts with respect to eligibility for government subsidy. Column (1) - (2) are based on eligibility for Empire Zones. Columns (3) - (4) are based on eligibility for Opportunity Zones. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.