Research Proposal: A Bayesian County-Level Simulation of the U.S. Small Business Economy with Applications to Policy and Disaster Assessment

Edward J. Egan, Ph.D.
Petabyte Economics Corp.

December 23, 2024

1 Background

1.1 Summary

This document provides the technical description of a research proposal for "Research on Small Business Topics Using Economic Data" in response to solicitation 73351025Q0172 from the Small Business Administration (SBA) Office of Advocacy. It describes a state-of-the-art, data-driven, evidence-based approach to policy and event assessment and a research proposal that leverages this approach. The proposal covers topics a, f, g, i, and n described in section 3.4 of the Statement of Work, and will provide novel findings concerning, among other things, the effects of federal policies and natural disasters on rural small business trends.

The approach creates a Bayesian spatio-temporal simulation of the U.S. small business economy at the county-year grain. The simulation supports scenario analysis and full-joint distribution forecasts. Econometricians can use it to evaluate the consequences of actions by policymakers or nature and project the effects of future shocks to the U.S. small business economy. The simulation will also shed light on America's small business dynamics.

The research proposal will use this simulation to i) assess the impact of the SBA HUBZone program, which aims to award at 3% of federal contract dollars to small businesses in designated areas, and the Opportunity Zones program created under the Tax Cuts and Jobs Act of 2017, which provides tax incentives for investment in designated areas; ii) assess the effects of wildfires, hurricanes, flooding, drought, and other natural disasters on small businesses at the county-year level; and iii) provide descriptive statistics on specific trends and patterns of small business dynamism.

1.2 Research Team

Dr. Edward J. Egan will be the principal investigator and sole team member. Dr. Egan contracts through Petabyte Economics Corp. ("Petabyte", UEI: VRNNE19YME58-0000). Petabyte is a small, private R&D lab working on broad-purpose cloud-scalable economic models, including the model described in this proposal. The company aims to reduce the barriers to real-world use of normative economic approaches to make optimal data-driven and evidence-based strategic or policy decisions. Dr. Egan specializes in entrepreneurship and innovation economics, so he is particularly keen to see normative economic approaches adopted by public agencies, like the SBA, with mandates in these areas.

Dr. Egan received his Ph.D. from the Haas School, U.C. Berkeley in 2012, and subsequently held positions as the Innovation Policy Fellow at the National Bureau of Economic Research, an assistant professor of entrepreneurship and innovation at Imperial College Business School, a fellow of the James A. Baker Institute for Public Policy at Rice University, and a visiting professor of entrepreneurship at Georgetown University's McDonough School of Business. From 2015 to 2018, Dr. Egan was the founding director of the McNair Center for Entrepreneurship and Innovation at Rice University. From 2020 to 2023, Dr. Egan was a research economist at Amazon's advanced research projects group, Core AI, where he developed novel machine learning (ML) technologies to support Bayesian simulation, optimization, and inference.

During his academic appointments, Dr. Egan developed and taught courses covering entrepreneurial finance, firm strategy, public policy, technological innovation and commercialization, economic modeling, and research and computational methods. He also engaged with more than 50 public bodies or policymakers' offices, including the SBA, National Science Foundation (NSF), National Security Administration, U.S. Cyber Command, and Council of Economic Advisors, as well as various U.S. House and Senate committees and state and municipal economic development departments. Dr. Egan's research primarily applies computational economics to policy-relevant questions in entrepreneurship and innovation, including those relating to small business. His research has been published in leading journals, including Research Policy and the Journal of Economic Geography, and his C.V. is provided in the submission package.

1.3 A State-of-the-Art Approach to Policy

This research proposal embodies a normative, objective, data-driven, and evidence-based approach to policymaking and evaluation grounded in economics. Recent efforts to move the U.S. towards using such an approach include the Evidence-Based Policymaking Act of 2018 (H.R.4174, 2018) and President Biden's "Memorandum on Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policy Making" (Memorandum, 2021).

This research proposes using the CatFish model, described in Egan (2024), to construct a Bayesian simulation of the U.S. small business economy. CatFish is a Categorical variable multiplicative mixed Poisson model. Appendix 4.2 provides a sketch of its underlying math and a brief description of its compute requirements. Its software is available from Petabyte's GitHub page[1] and is written in MATLAB. Scripts to specify and run CatFish models are readily sharable.

CatFish provides scalable, multidimensional simulation and forecasting of count data outcomes grounded in Bayesian theory. It infers the full joint distribution of outcomes,

[1] https://github.com/petabyteeconomics/CatFish

rather than point estimates, such as means or specific percentiles, and its rate parameters and outcomes are aggregably consistent so it can support policy actions at appropriate grains. For example, in CatFish, county forecasts sum to state forecasts, which sum to regional and U.S. forecasts, allowing policy actions at all levels. Most alternative approaches must make separate, incoherent forecasts at different grains for different purposes. Catfish also has easy-to-interpret effects: When one factor is omitted from each category, absent multicollinearity, the omitted factor has a value of one, and the values of the other categories indicate their relative effects. One can also test whether the value of a factor is statistically different from one, which indicates no effect.

Using a Bayesian simulation like CatFish furthers the SBA's goal of moving beyond the current generation of evidence-based policy techniques and incorporating risk. Specifically, the SBA uses the language of statistical decision theory (see section 4.1) in its objective to "strategically manage resources by integrating quality data, evidence, and risk in decision-making processes" (see objective 3.1 in SBA, 2022). The two crucial barriers to using statistical decision theory in practice are that one must specify an appropriate loss function and then determine the action that minimizes it in expectation. This expectation minimization requires a conditional posterior distribution of outcomes on actions, which can be computed from a full joint distribution. A well-specified CatFish model can also provide findings relevant to an appropriate loss function.[2]

2 Research Proposal

This research proposes creating a Bayesian simulation of the U.S. small business economy. The simulation will have three dimensions: small business size (1-9 or 20-499 employees), U.S. county (3,144 counties), and year (1986 to 2022 = 45 years for the training period).[3]

[2] A full decision-theoretic model is beyond the scope of this research proposal. However, Petabyte is currently prototyping a cloud-scalable, multidimensional, non-parametric, automated risk minimization method. We would be delighted to use this technology to attempt a full decision-theoretic model of federal small business policy beginning in May 2025.

[3] Therefore, the models' space will be $(2 \times 3,144 \times 45 =)$ 282,960 elements large.

The outcomes will be counts of small businesses. The simulation will assess the effects of HUBZones, Opportunity Zones, and natural disasters on the number of small businesses in a county. It will also provide trends and stylized facts about the levels of small business activity in counties over time. A particular focus will be given to rural small businesses, using the SBA's definition of a county as rural versus metropolitan (see, for example, SBA, 2023).

Outcome data will be taken from the Business Dynamics Statistics (BDS) data provided by the U.S. Census. Censoring of the data by the Census to preserve anonymity limits the choice of grain. An initial exploration suggests that using two course-grained groups of small business sizes at the county-year level is optimal using public data.[4] Moreover, HUBZones are either counties or areas within counties, and Opportunity Zones are census tracts, so their economic effects should be apparent at the county level.

2.1 Measures and Data

The dimensions of the model and its categorical factors define the characteristics of the simulation. Suppose the model includes factors that capture whether or not counties experienced a specific type of disaster in the previous year, whether they are rural or urban, whether their residents make less than 80% of the non-metropolitan state household income, whether they are designated as a HUBZone, and so forth. Then, the Poisson rate for a poor, rural, flood-stricken HUBZone county is the product of the values for those factors.

I, therefore, propose including some or all of the following county-year-level factors in the model:

- Indicators for specific types of natural disasters, including hurricanes, wildfires, flooding, and drought, available from NOAA Storm Events Database (NOAA) from 1950 to the present.

[4] At finer grains, including using seven small business sizes, zip codes as geographic areas, or adding sector or other NAICS groups, the data are either likely too censored or not publicly available. The availability of data sets covering weather, disasters, socio-economics, and other measures reinforces the county-level choice.

- Indicators for disasters from FEMA's Disaster Declarations for States and Counties (DDS) Summaries File.
- Categories for federal contracting goals that relate to HUBZones.
- Indicators that counties contain a census tract designated as an Opportunity Zone.
- Quantiles for median income, poverty level, household wealth (c.f., Fairlie, 2012), race (c.f., Fairlie, 2018 and Howell, 2021), education (c.f., Moutray, 2007), age (c.f., Zissimopoulos, 2007), and internet access from the American Community Survey (ACS).
- Quantiles for women, minority, veteran, and other business owner characteristics from the Annual Business Survey (ABS).
- Quantiles for agglomeration of established firms, possibly by sector, from BDS (see Egan, 2023).

I also propose including some or all of the following other factors in the model:

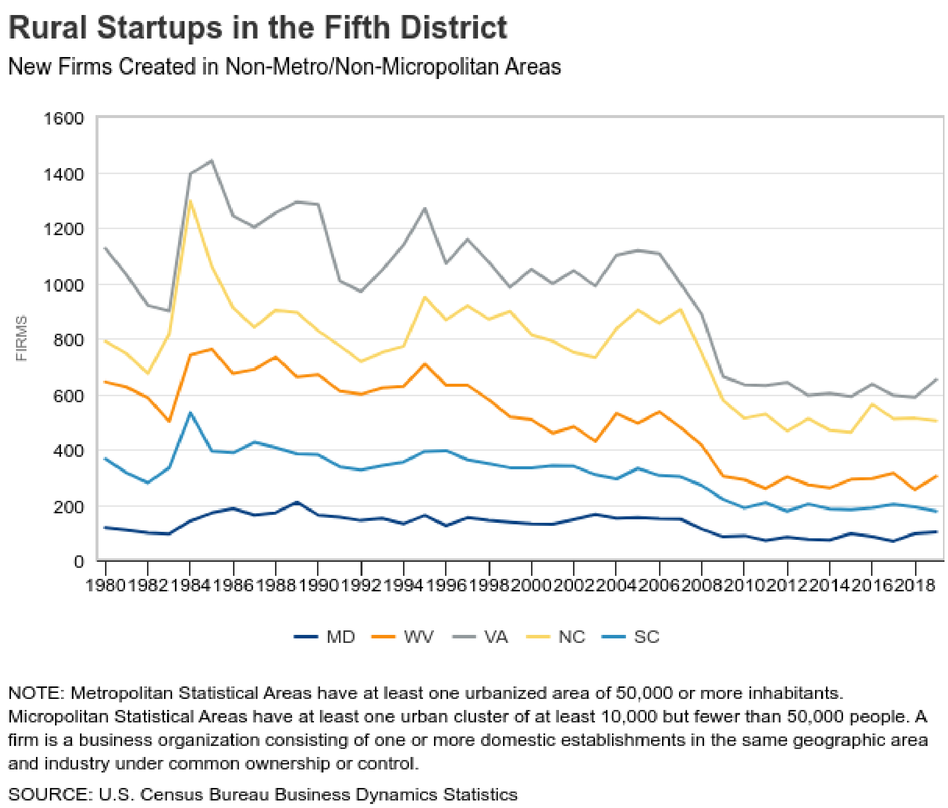
- Categories of risk scores from the FEMA National Risk Index (NRI) described in FEMA (2024).
- Indicators for COVID-19 periods. See Fairlie (2020), Kim (2020), Kobe (2022), Richards (2023), and others, for the effects of COVID-19 on small businesses.
- Indicators for recessions, quantiles for inflation, and other macroeconomic factors from the Bureau of Economic Analysis.
- Categories for geographic areas or types or geographies (i.e., coastal, mountains, etc.)

2.2 Literature Review

A full literature review will be conducted as a part of the research. This section provides a synopsis of relevant work to frame the proposal.

2.2.1 Rural Small Business Trends

Basic data and characterizations of trends for rural small businesses are surprisingly hard to find. SBA (2024), which provides the 2024 Small Business Profile for Rural Areas, is drawn from the Census' Nonemployer Statistics and Statistics of U.S. Businesses and does not contain firm counts. An analysis by the SBA (see Wilmoth, 2019) states that the number of small business establishments (not firms) grew by 7.2% from 2000 to 2019. However, a popular chart from the Census shows that the number of small rural firms in Fifth District states declined by around half from their highs in the mid-1980s to 2019.



There has been a recent resurgence in the dream of tech entrepreneurs moving to rural areas and stimulating local economies, evident in Sablik (2022) and elsewhere. These authors echo sentiments expressed during the dotcom boom (see Henderson, 2002), which came to naught.

2.2.2 HUBZones and Opportunity Zones

The literature generally finds that HUBZones and Opportunity Zones do not spur local economies. For Opportunity Zones, example papers include Snidal (2024)'s "The Nonimpact of Opportunity Zones on Home and Business Lending", Eldar (2022)'s "Opportunity Zones: A Program in Search of a Purpose", andTrivedi (2021)'s "Opportunity Zones Providing Opportunity for Whom?". For HUBZones, the SBA was asked in GAO (2008) to "directly measure the program's effect on firms" or "directly measure the program's effect on the communities in which the firms are located". In response, the SBA commissioned Beale (2008), which found that "the program has a substantial impact in only a very small percentage of HUBZones... however, the impact in two-thirds of all HUBZones is nil". Dilger (2017), a report by the Congressional Research Service, highlights that "debates over the program's effect on economically distressed communities" continued a decade later.

2.2.3 Natural Disasters and Small Business

A county-grain simulation of U.S. small businesses offers a compelling opportunity to study the effects of natural disasters. Press (2024) provides some summary statistics on enterprises that have experienced disasters of various types from 2017 to 2021 by business size. Surprisingly, more than 40% of U.S. small businesses were in a county that experienced one or more disasters. Most research papers, however, study a single type of disaster, often in a specific location. For example, Marshall (2015) studies small business demise and Runyan (2006) identifies barriers to recovery following Hurricane Katrina, while Davlasheridze (2017) looks at the impacts of floods on firms with fewer than 50 employees. A notable exception is Eyre (2020), who use social media activity to study the recovery of small businesses from various natural hazards in Puerto Rico and elsewhere. This research will provide broad findings that can summarize and extrapolate the impact of natural disasters on all U.S. small businesses.

2.3 Hypotheses and Testing

The hypotheses for the three main analyses are as follows:

- I do not advance specific hypotheses concerning rural small business trends.[5] Instead, I intend to report interesting observations. The simulation will contain covariates that can shed light on important rural small business dynamics.

- The stated goal of both HUBZones and Opportunity Zone policies is to create local economic development and benefits. However, the literature suggests that these programs have not had positive effects. As a consequence, I maintain the null hypothesis that neither program will have a statistically significant effect on small business levels. I will use a difference-in-differences approach within the simulation to infer causality.

- Natural disasters are exogenous, so my empirical approach will use lagged factors and controls for covariates that predispose a county to impact. My prior is that, despite mitigation and recovery efforts, natural disasters will lead to decreases in the number of small businesses in counties. I expect these effects to be related to counties' socio-economic and geographic characteristics and the types of disasters.

Finally, this research will briefly examine whether the federal policy inadvertently incentivizes small businesses to locate in disaster areas by examining the relative incidence of natural disasters in HUBZone and Opportunity Zone counties. Johnson, Vermont, illustrates the issue. Johnson is part of Census Tract 9532, designated as an Opportunity Zone in March 2018, following its selection by Governor Phil Scott. Johnson flooded in July 2023 and experienced material damage, and flooded again precisely one year later, though the damage was lighter in 2024. In September 2024, 17 properties in Johnson were considering voluntary buyouts using funds from FEMA. Johnson has a long history of flooding, dating

[5] Despite moving to Vermont and founding Petabyte, I find a large-scale flight of tech entrepreneurs from cities to rural areas dubious. I also find it unlikely that small businesses will become extinct in rural counties.

back to the 1927 flood, and experienced significant damage during Tropical Storm Irene in 2011, raising questions about why it was selected as an Opportunity Zone.

3 Conclusion

Small businesses are crucial drivers of U.S. economic growth and prosperity, and recent advances have made a new policy research, design, and evaluation paradigm possible. Bayesian simulations that can increase knowledge about the state of economic entities and perform scenario analyses on actions affecting them are now feasible in many practical contexts.

Accordingly, this research proposes creating a spatio-temporal Bayesian simulation to assess the causal impacts of three very timely small business policy topics: HUBZones, for which there have been recent changes in federal contract dollar targets (the SBA also has ongoing efforts to develop new measures to assess HUBZones' impacts); Opportunity Zones, which appear likely to be extended during President-Elect Trump's second term; and natural disasters, which are currently increasing in frequency. The simulation will also support the Office of Advocacy's mission to develop and disseminate small business statistics by shedding fresh light on rural small business dynamism.

4 Appendix

4.1 Statistical Decision Theory

Statistical decision theory (see, for example, Berger, 1985) is an axiomatic approach to making optimal decisions under uncertainty. Statistical decision theory has four essential elements: actions, outcomes, loss (and so risk), and statistical evidence.

1. A is the set of actions (i.e., policy decisions) a that can be taken.

2. An outcome vector $\omega \in \Omega$ is random at the time of action but determines the implications of the actions, a, *ex post*.

3. The loss (or gain) function $L(a, \omega)$ quantifies the implications of action a when the outcome is ω.

4. Statistical knowledge $\theta \in \Theta$ provides evidence about uncertainties in outcomes ω. In Bayesian approaches (see, for example, Geweke, 2005), information about θ comes from priors and is inferred from data.

In statistical decision theory, one evaluates alternative actions a by minimizing a risk function, $R(a) = \mathrm{E}_{p(\omega|a)}\, L(a, \omega)$, or equivalently maximizing expected welfare, using statistical knowledge to learn about outcomes.

4.2 CatFish

CatFish produces mixed Poisson rates by inferring the magnitude of effects, θ, of latent categorical factors, c. Let $\mathbf{E}$ be the space formed by the cartesian product of the model's dimensions. If some categorical factors $c \in C_E$ apply to an element $Ein\mathbf{E}$, and ω_E denotes the outcome at E, then:

$$p(\omega_E|\lambda_E) \sim \mathrm{Pois}(\lambda_E) \text{ where } \lambda_E = \prod_{c \in C_E} \theta_c \tag{1}$$

As the sampling distribution $p(\omega|\lambda)$ is Poisson (see equation 1), we use a conjugate gamma prior for $p(\theta)$. By Bayes' theorem, the conditional posterior $p(\theta|\omega, \Theta_{-\theta})$ is then gamma distributed and can be Gibbs sampled. Moreover, as each set of Markov Chains are independent, they can be sampled in parallel. Each sample contains different draws of each θ, so different draws of λ create a mixed Poisson distribution. Note that the conditional posterior, $p(\omega|a)$, can be derived from a full joint posterior distribution using Bayes' rule.

CatFish uses a Markov Chain Monte Carlo (MCMC) Gibbs sampling approach to inference.[6] This approach is highly scalable, ergodic, and can be parallelized to take advantage of

[6]Extensions to CatFish, available on GitHub, add Metropolis-within-Gibbs sampling.

modern cloud computing. Petabyte's lead scientists have also developed proprietary methods, called Autotailor, to make MCMC practical in industrial contexts. Autotailor removes warmup sequences and thins parallel Markov Chains to eliminate serial correlation without human-expert intervention.

The computational scale of this project is trivial by modern big data standards: A 282,960-element space equates to about 1.1MB using single-precision floating-point format. So, this model's inference can be performed on a single compute node in seconds at negligible cost, irrespective of the complexity of the model's topology. In general, models with spaces up to around 30 gigabytes can be computed for tens of dollars on a single 8-GPU node in several minutes, and models up to 0.5 terabytes can computed for several hundred dollars on a small CPU-based cluster in under an hour.[7] The SBA may have internal access to unrestricted Business Dynamics Statistics data from the U.S. Census. In this case, a six-dimensional model with seven firm size groups, 324 4-digit NAICS codes, 3,144 counties, 10 firm age groups, and 45 years would be around 13 GB and readily fit into a single node.

4.3 Timeline and Milestones

The timeline and milestones for the project are provided in section 4.3 of the Statement of Work.

References

Beale, Henry and Deas, Nicola (2008), The HUBZone Program Report, Washington, DC: Microeconomic Applications, Inc., prepared for the SBA, Office of Advocacy, May 2008, pp. i-iii.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer.

[7] We support models up to petabyte scale using P4d Ultraclusters.

Davlasheridze, M., Geylani, P.C. (2017), Small Business vulnerability to floods and the effects of disaster loans. Small Bus Econ 49, 865-888.

Dilger, Jay (2017), Small Business Administration HUBZone Program, Senior Specialist in American National Government, Congressional Research Service

Eastman, S. and Kaeding, N. (2019), Opportunity Zones: What We Know and What We Don't. Washington, DC: Tax Foundation.

Egan, Edward J. and Brander, James A. (2023), A New Method for Identifying and Delineating Spatial Agglomerations with Application to Venture-Backed Startups. Journal of Economic Geography, Vol. 23, Iss. 3, May, pp. 485-508.

Egan, Edward J. (2024), The CatFish Handbook, Petabyte Economics Corp. Version 1.5, February.

Eldar, O. and Garber, C. (2022), Opportunity Zones: A Program in Search of a Purpose. BUL Rev., 102, p.1397.

Eyre, R., De Luca, F. and Simini, F. (2020), Social media usage reveals recovery of small businesses after natural hazard events. Nat. Commun. 11, 1629.

Exec. Memorandum (2021), Memorandum on Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking, Joseph R. Biden Jr.

Fairlie, Robert W., and Harry A. Krashinsky. (2012). Liquidity constraints, household wealth, and entrepreneurship revisited. Review of Income and Wealth, 58(2): 279-306.

Fairlie, Robert. (2018). Racial inequality in business ownership and income. Oxford Review of Economic Policy, 34(4): 597-614.

Fairlie, Robert. (2020). The Impact of COVID-19 on Small Business Owners: Evidence from the First Three Months after Widespread Social-distancing Restrictions. Journal of Economics and Management Strategy, 29(4): 727740.

FEMA (2024), The OpenFEMA Dataset: National Risk Index Counties. Federal Emergency Management Agency (FEMA), Retrieved from https://resilience.climate.gov/datasets/FEMA::national-risk-index-counties/about on 12/13/2024. This product uses the FEMA OpenFEMA API, but is not endorsed by FEMA. The Federal Government or FEMA cannot vouch for the data or analyses derived from these data after the data have been retrieved from the Agency's website(s).

GAO (2008), Small Business Administration: Additional Actions are Needed to Certify and Monitor HUBZone Businesses and Assess Program Results, U.S. Government Accountability Office, GAO-08-643, June 17, 2008, p. 5.

Geweke, John (2005), *Contemporary Bayesian econometrics and statistics.* John Wiley & Sons.

Geweke, John and Durham, Garland (2019), Sequentially adaptive Bayesian learning algorithms for inference and optimization, *Journal of Econometrics*, Volume 210, Issue 1, Pages 4-25, ISSN 0304-4076.

Henderson J. (2002). Building the rural economy with high-growth entrepreneurs. Economic Review, Federal Reserve Bank of Kansas City, 87(3), 45-70.

Howell, Sabrina T., Theresa Kuchler, David Snitkof, Johannes Stroebel, and Jun Wong. (2021). Racial Disparities in Access to Small Business Credit: Evidence from the Paycheck Protection Program. National Bureau of Economic Research Working Paper No. w29364.

HR4174 (2018), An act to amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes. 115th Congress

Kim, Olivia S., Jonathan A. Parker and Antoinette Schoar. (2020). Revenue Collapses and the Consumption of Small Business Owners in the Early Stages of the COVID-19 Pandemic. NBER Working Paper No. w28151.

Kobe, Kathryn (2022), Patterns of Small Business and Young Business Hires Pre & Post COVID-19 by Region, ION Economics LLC, Office of Advocacy U.S. Small Business Administration under contract number 73351021R0008

Kugler, Maurice, Marios Michaelides Neha Nanda, and Cassandra Agbayani (2017), Entrepreneurship in Low-Income Areas, IMPAQ International, LLC Columbia, MD 21044, Office of Advocacy U.S. Small Business Administration under contract number SBAHQ-15-M-0150, September

Marshall, M.I., Niehm, L.S., Sydnor, S.B. et al. (2015), Predicting small business demise after a natural disaster: an analysis of pre-existing conditions. Nat Hazards 79, 331-354.

Moutray, C. (2007). Educational attainment and other characteristics of the self-employed: An examination using the Panel Study of Income Dynamics data. Washington, D.C.: U.S. Small Business Administration, Office of Advocacy working paper.

Press, Robert (2024), The Eye of the Storm: Small Businesses and Natural Disasters, U.S. Small Business Administration, Research Spotlight, October

Richards, Tyler (2023), Business Dynamics During The Covid-19 Pandemic, U.S. Small Business Administration, Office Of Advocacy, Issue Brief 17, January

Runyan, R. C. (2006). Small business in the face of crisis: identifying barriers to recovery from a natural disaster. Journal of Contingencies and Crisis Management, 14(1), 12-26.

SBA (2022), Enterprise Learning Agenda Fiscal Years 2022-2026, U.S. Small Business Administration, Office of Program Performance, Analysis, and Evaluation

Wilmoth, Daniel (2019), Growth In Number Of Rural Establishments, Small Business Facts, U.S. Small Business Administration, Office of Office of Advocacy, December

SBA (2024), Rural Areas Profile 2024, U.S. Small Business Administration, Office of Advocacy, August

Sablik, Tim (2022), Growing Rural America Through Startups, Econ Focus, First Quarter 2022

Snidal, M. and Li, G. (2024), The Nonimpact of Opportunity Zones on Home and Business Lending. Housing Policy Debate, 34(3), pp.419-440.

Trivedi, Ruta R. (2021), Opportunity Zones Providing Opportunity for Whom?: How the Current Regulations Are Failing and a Solution to Uplift Communities, 27 Wash. & Lee J. Civ. Rts. & Soc. Just. 745

Webb, G. R., Tierney, K. J., and Dahlhamer, J. M. (2000). Businesses and disasters: empirical patterns and unanswered questions. Natural Hazards Review, 1(2), 83-90.

SBA (2023), Small Business Facts: Small Businesses in Rural Areas, Fact Sheet, U.S. Small Business Administration, Office of Office of Advocacy, Aug 22.

Zissimopoulos, Julie M., and Lynn A. Karoly. (2007). Transitions to self-employment at older ages: The role of wealth, health, health insurance and other factors. Labour Economics 14(2): 269-295.