

ABSTRACT

Title of Dissertation: **ESSAYS IN MACROECONOMICS
AND HOUSEHOLD FINANCE**

**Manuel Esteban Molina Soto
Doctor of Philosophy, 2025**

Dissertation Directed by: **Professor John Shea
Department of Economics**

The historical co-occurrence of major new innovations and asset price booms is well established. In the first chapter of this dissertation, we empirically document this phenomenon for “breakthrough” patents. We then propose a mechanism by which rational bubbles cause (but are not caused by) breakthrough innovation and improve welfare. Our hypothesis is that bubbles facilitate greater diversification in a new, but risky, sector. Bubbles do crowd out saving in capital, but on the other hand, they also enable productive investors to access more resources for investment in the risky sector. This, in turn, enables investment in a larger set of risky ideas, increasing the probability of successful breakthrough innovations. Successful innovations further enable capital accumulation and diversification. An economy with bubbles generates a higher steady-state level of output and consumption and a lower risk of unsuccessful investments.

In the second chapter, we examine how immigration status in the U.S. influences

outcomes in the mortgage market. Using ACS data, we trained machine learning classifiers to predict whether an individual is a U.S. citizen and applied these models to classify HMDA mortgage applications. We find that for Hispanic households, non-U.S. citizens face a 2.1 percentage point lower likelihood of approval, while for Asian households the effect is positive.

We then further categorized applicants as authorized or unauthorized immigrants. We find that, among Hispanic households, likely non-U.S. citizens face a 1.7 percentage point lower probability of approval, while likely unauthorized immigrants face an additional 0.47 percentage point decrease in approval probability. In contrast, for Asian households, likely non-U.S. citizens have a 0.8 percentage point higher probability of approval, while likely unauthorized immigrants experience a 0.37 percentage point decrease.

We also match our data with Fannie Mae and Freddie Mac's performance data and show that likely non-U.S. citizens are not riskier than their U.S. citizen counterparts; they do not have a lower credit score or a higher likelihood of delinquency or default and therefore they should not have a lower likelihood of approval.

ESSAYS IN MACROECONOMICS AND HOUSEHOLD FINANCE

by

Manuel Esteban Molina Soto

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:

Professor John Shea, Chair
Professor Thomas Drechsel
Professor Agustin Hurtado
Professor John Wallis
Professor Min Wu

© Copyright by
Manuel Esteban Molina Soto
2025

Acknowledgments

First, I want to express my sincere thanks to my chair, Professor John Shea, for his mentorship and guidance. He read my papers thoroughly, offered thoughtful feedback, and was always available to meet and discuss every step of my PhD journey.

I am also deeply grateful to Professor Thomas Dreschel, who was always kind and generous with his time. He regularly met with me to discuss my papers and ideas, and his feedback and support were truly valuable throughout the process.

I feel fortunate to have worked with Professor Agustin Hurtado. He was an exceptional mentor and provided invaluable help in writing my dissertation. I learned so much from him. I wished I had met him earlier in my PhD journey—I'm sure we could have written other good papers together.

One of the highlights of my PhD was participating in the history reading group with Professor John Wallis. I enjoyed learning about economic history from such a knowledgeable and smart person. He always had an answer to every question, and I'm confident that everything I learned from him will be extremely valuable in my career.

I also want to thank Professor Min Wu and Dr. Nick Slaughter, who hired me and allowed me to complete my PhD without any funding worries. I always enjoyed working with both of them, and I truly appreciate their support.

In addition, I want to thank Aditya, who has been both my co-author and one of my closest friends.

I want to thank my mother, Edilma, who has always supported me, cheered me on, offered advice, and encouraged me to pursue my dreams. I am deeply grateful for everything she has taught me.

I also want to thank my aunt Adalgiza, my cousin Paulina, and my second family in the U.S., Caridad and Adam, for their love and support throughout this journey.

Finally, I want to thank my grandparents, Víctor and Edilma for their unconditional love, although they are no longer with us, I know they would be incredibly proud to see the first PhD in the family.

Table of Contents

Acknowledgements	ii
Table of Contents	iv
List of Tables	vi
List of Figures	viii
List of Abbreviations	x
Chapter 1: Breakthrough Innovations, Asset Price Booms, and Diversification (joint with Aditya Pande)	1
1.1 Introduction	1
1.2 Review of Related Literature	3
1.2.1 Empirical Literature	3
1.2.2 Theoretical Literature	6
1.3 Empirics	8
1.3.1 Breakthrough Patents (KPST (2021))	8
1.3.2 Bubbles (Greenwood, Shleifer, Yang (2018))	11
1.3.3 Descriptive Statistics	12
1.3.4 Sectoral Regressions	13
1.3.5 Bubbles, IPO Size and Breakthrough Patents	16
1.4 Model	20
1.4.1 Household	21
1.4.2 Technology	22
1.4.3 Risk and Total Investment	23
1.4.4 Financial Intermediaries	24
1.4.5 Optimal Investment	25
1.4.6 Bubbles	25
1.4.7 Capital	28
1.4.8 Level of Diversification	30
1.4.9 Recursive problem	31
1.4.10 Existence of bubbles	32
1.5 Simulation	37
1.5.1 Always "Lucky" Steady State	38
1.5.2 Fully Stochastic Economy	40

1.6	Conclusion	42
Chapter 2: Immigration Status and Access to Mortgage Financing		43
2.1	Introduction	43
2.2	Data	47
2.3	Machine Learning Model	52
2.3.1	Categorizing between Authorized and Unauthorized	58
2.4	Results in Aggregated Data	60
2.5	Micro-Level Regression Results	64
2.5.1	Alternative Specification using Propensity Score Matching	69
2.5.2	Mortgages for unauthorized immigrants	70
2.6	Default Risk	74
2.6.1	Alternative Specification using propensity score matching	81
2.6.2	Citizenship and loan risk during the financial crisis	81
2.7	Conclusion	84
Appendix A: Proofs and Robustness test in Chapter 1		86
A.0.1	Proof of Bubbles existence	86
A.0.2	Derivation Under No Bubble Creation	87
A.0.3	Interval in Which Bubble Creation Shifts Up $E[x_{t+1}]$	88
A.0.4	Robustness Test: Probit with Fixed Effects	90
Appendix B: Additional Tables and Robustness Tests for Chapter 2		92
B.0.1	Shapley values	92
B.0.2	Withdrawal Rates	93
B.0.3	Interest Rate	97
B.0.4	Risk Regression including Co-applicant	99
B.0.5	Mortgage Data by State	101
B.0.6	Default Matching Details	101
B.0.7	Addressing Selection in our training data with Kernel Density Estimation	102
Bibliography		107

List of Tables

1.1	Selected Top Inventions Per Decade	10
1.2	Breakthrough Patents Per Capita by Group	12
1.3	Panel Regressions	15
1.4	Breakthrough Patents per Capita: Alternative Bubble	16
1.5	CAPE Bubble vs. HHL Bubble	16
1.6	Summary statistics for Offer Size, EBITDA, Patents, and Bubble indicator .	17
1.7	Regression results of a Fama-French bubble on the IPO offer size	18
1.8	Estimates of the Effect of Offer size of breakthrough patents	20
1.9	Parameters	39
2.1	Effective Number of observations in the training dataset.	49
2.2	Income Statistics for Hispanic and Non-Hispanic Households in the HMDA Data set (values in 000s)	51
2.3	Parameter values for the XGBoost model	55
2.4	Estimates of the Effect of Non-U.S. Citizenship on Approval Rates	66
2.5	Collinearity Diagnostics in the Approval Probability Model for Non-U.S. Citizenship	68
2.6	Average Marginal Effect of Non-US Citizenship on Mortgage Approval . .	70
2.7	Estimates of the Effect of being Non Us Citizen and unauthorized on Mortgage Approval	73
2.8	Collinearity Diagnostics	74
2.9	Summary statistics by ethnicity and predicted citizenship status	75
2.10	Estimates of the Effect of Non-U.S. Citizenship on Credit Score	77
2.11	Estimates of the Effect of Non-U.S. Citizenship on default	78
2.12	Average Marginal Effects (AME) on delinquency and default by citizenship status	81
2.13	Estimates of the Effect of Non-U.S. Citizenship on default for mortgages acquired from 2005-2007	83
A.1	Probit estimates of bubbles on Breakthrough patents	90
B.1	Estimates of the Effect of Non-U.S. Citizenship on Withdrawal Rates . . .	95
B.2	Estimates of the Effect of being Non Us Citizen and unauthorized on Withdrawal Rates	96
B.3	Estimates of the Effect of Non-U.S. Citizenship on Interest Rate	97

B.4	Estimates of the Effect of being authorized and unauthorized on Interest Rates	98
B.5	Estimates of the Effect of Non-U.S. Citizenship on default	100
B.6	Mortgage data by state	101
B.7	Estimates of the Effect of Non-U.S. Citizenship on Approval Rates	106

List of Figures

1.1	Breakthrough Patents Per Million Persons	10
1.2	Number of Sectors Experiencing A Bubble	13
1.3	Minimum size requirement and optimal portfolio	26
1.4	Existence	34
1.5	Existence	35
1.6	Path to Steady State	40
1.7	Simulation with bubbles and without bubbles, given same draws of shocks to risky technology.	42
2.1	Income differences between Hispanic and non-Hispanic households, based on immigration status, in the ACS dataset after applying filters to identify mortgage market participants.	50
2.2	Income Densities for ACS and HMDA in 2007. The left panels show the distributions in our raw datasets, while the right panels shows the distributions after filtering out data below the 1st percentile of the HMDA income distribution	52
2.3	Sensitivity and Specificity on the Testing Dataset for the Model Trained with Hispanic Households Data	57
2.4	Sensitivity and Specificity on the Testing Dataset for the Model Trained with Non-Hispanic Households Data	57
2.5	Prediction Metrics for the Machine Learning Model Predicting Unauthorized Status Among Non-U.S. Citizens	59
2.6	Predicted Percentage of Non-U.S. Citizens by Year in HMDA and the Percentage of Non-U.S. Citizens in Our ACS Training Data	61
2.7	Mortgage approval rate by predicted citizenship status	63
2.8	Mean Applicant income by predicted citizenship status	64
2.9	Predicted percentage of unauthorized immigrants among all observations by year in HMDA, compared to the percentage of unauthorized immigrants in our ACS training data.	71
2.10	Delinquency and default rates by citizenship status.	76
2.11	Delinquency and default rates by race.	76
2.12	Delinquency rate, percentage of mortgages held by Hispanics and percentage of mortgages held by likely non-U.S. citizens by State for mortgages originated before the financial crisis (2005-2007)	82

B.1	Shapley Values	93
B.2	Withdrawal Rate by predicted citizenship status	94
B.3	Income distribution for non-Hispanic households in 2007. The left panel displays the distribution in our original ACS training dataset, and the right panel shows the distribution after sampling using the KDE-estimated densities.	104
B.4	Gender distribution for non-Hispanic households in 2007 for our original ACS training dataset, ACS after sampling using the KDE-estimated densities and HMDA	105

List of Abbreviations

ACS	American Community Survey
AME	Average Marginal Effect
CAPE	Cyclically-Adjusted Price-Earnings Ratio
FHA	Federal Housing Administration
FSA	Farm Service Agency
GPT	General Purpose Technologies
GSE	Government-Sponsored Enterprise
HMDA	Home Mortgage Disclosure Act
IPO	Initial Public Offering
ITIN	Individual Taxpayer Identification Number
KDE	Kernel Density Estimation
LEP	Limited English Proficiency
MSA	Metropolitan Statistical Area
NAICS	North American Industry Classification System
PUMA	Public Use Microdata Area
R&D	Research and Development
RHS	Rural Housing Service
TFP	Total Factor Productivity
SIC	Standard Industrial Classification
SSN	Social Security Number
VA	Veterans Administration
VC	Venture Capital
VIF	Variance Inflation Factor

Chapter 1: Breakthrough Innovations, Asset Price Booms, and Diversification

(joint with Aditya Pande)

1.1 Introduction

The process of growth at the technological frontier seems to be punctuated by the arrival of periodic “macroinventions” which are then innovated upon and incorporated throughout the economy.¹ These General Purpose Technologies (GPTs),² such as the steam engine, electricity, the semiconductor, and artificial intelligence arguably emerge “ab nihilo”.³ As Crafts (1995) points out, “technological history suggests that seeking for socio-economic explanations of macroinventions is likely to be a fruitless pursuit.” [5] The subsequent process of generating “microinventions”, however, is undertaken by firms that respond to the economic (and importantly for us, financial) environment.

Indeed, many authors have documented large movements in asset prices in sectors that experience such technological revolutions.⁴ To the best of our knowledge, there are still advances to be made in assessing the *causal* effect of these asset price movements

¹Gordon (2016) [1], Mokyr (1990) [2], Philippon (2023) [3], Bloom et. al. (2023) [4]

²We will use GPT and macroinvention interchangeably throughout. We avoid the word “breakthrough” to describe these innovations given that it already has a definition in the empirical literature, used below.

³Mokyr (1990). [2]

⁴Pastor & Veronesi (2009) [6], Lamoreaux et. al. (2009) [7]

on the process of diversification and innovation following the arrival of a GPT. This gap exists, in our view, because of two difficulties that have only recently been alleviated in the literature.

The first challenge is to identify the critical microinventions which build upon the GPT. Recent advances in text analysis allow us to identify more or less novel patents based on the relative frequency of new terms in a patent's text. Using recently available data from Kogan, Papanikolaou, Seru, and Taddy (2021) [8], along with sectoral stock market data. In this paper, we empirically show that more novel, higher-value "breakthrough" patents are positively associated with asset price run-ups that predictably crash.

The second challenge is to illustrate theoretically how assets which trade at prices above fundamental value can cause increased innovation rather than the reverse. Martin and Ventura (2012) [9] develop a tractable model of rational bubbles, drawing on the seminal works of Samuelson (1958) [10] and Tirole (1985) [11]. In their work, assets with no fundamental value are traded at positive prices, ease the financial friction between productive and unproductive investors, and facilitate increases in the capital stock and output. Our model builds on this foundation.

To explain our empirical findings, we propose a mechanism in which financial bubbles enable the economy to boost investment and diversification in newly available technologies that are risky but potentially highly productive. We develop a theoretical framework that incorporates a simple financial friction and introduces a new risky technology in which more diversification improves the chances of success, following Acemoglu and Zilibotti (1997). We integrate rational bubbles into this economy and demonstrate how they can contribute to higher capital accumulation and enhanced diversification in the

emerging, risky technology.

The rest of the paper proceeds as follows. Section 1.2 reviews the existing literature. Section 1.3 documents a positive association between asset price run-ups and breakthrough patenting. Sections 1.3.1 and 1.3.2 review our data sources for breakthrough patenting and asset price run-ups respectively, and Section 1.3.4 provides results from a sectoral panel regression. Section 1.3.5 shows that bubbles are associated with larger IPO size, and larger IPOs are associated with more breakthrough patenting in the future. Section 1.4 sets up and solves an overlapping generations model featuring bubbles and investors who can invest in risky breakthrough innovations or a safe asset. In this model, bubbles are welfare-improving: they increase investment in risky sector, which in turn increases diversification across risky breakthrough innovations, raising steady-state output and consumption.

1.2 Review of Related Literature

1.2.1 Empirical Literature

The literature examining asset price run-ups is vast, as is the literature on innovation, but the explicit intersection of the two is small. For example, in his 2021 Annual Reviews of Economics article, Simsek [12] cites only Haddad, Ho, & Loualiche (2022) [13] and Xu & Dang (2018) [14] as examples. We use similar data to HHL, while Xu & Dang are closer in empirical spirit. Nanda & Rhodes-Kropf (2013) [15] ask a similar question to ours but focus explicitly on venture capital. They find that firms funded in “hot” periods are not of lower quality, but rather riskier and more innovative. Dong, Hirshleifer, & Teoh

(2018) [16] similarly work at the firm level, identifying “misvaluation” in firm’s stock prices and linking it to the production of more novel patents. They focus on a subset of Compustat firms from 1976-2012 and use a measure of patent novelty, while we examine innovation of all firms from 1963-2002 using an alternative measure of patent novelty.

1.2.1.1 Xu & Dang (2018)

Xu & Dang find that higher equity prices due to increased investor optimism (“market sentiment”, in their terms) increase R&D investment and patent production by firms. They measure market sentiment as an aggregate time series using Shiller’s S&P 500 cyclically-adjusted price-earnings ratio (CAPE), and use a variety of measure of financial constraints.⁵ They find that financially constrained firms invest more in R&D when optimism is high, while financially unconstrained firms do not. Moreover, financially constrained firms are more likely to issue equity in boom times, whereas financially unconstrained firms are not. This “financing channel” is one of the mechanisms we also believe stimulates breakthrough innovation in booms. However, Xu & Dang do not explicitly address whether or why more innovative firms receive disproportionately more funding in optimistic times.

Xu & Dang find that booms improve innovation outcomes, as measured by quantity and quality of patents filed—that is, the number of patent filings and the number of citations per patent improve along with market sentiment. Compared to our work, they

⁵Specifically, they use “orthogonalized CAPE”, regressing CAPE on six macroeconomic variables used by Baker & Wurgler (2006) [17] and using the residuals of this regression as a measure of equity market sentiment. They also use Baker & Wurgler’s own investor sentiment index as an alternative. Constraints are proxied by size, projected free cash flow, and the dividend payout ratio, among others.

do not use newly available measures of patent quality, only focus on a short sample (1985-2004) of 6,139 publicly traded (Compustat) firms, and measure market sentiment as a single aggregate time-series. By contrast, we are able to tag asset price booms at the sectoral level. This allows us to control for both firm-fixed and time-fixed effects at the sector level to more cleanly measure the association between booms and innovation.

1.2.1.2 Haddad, Ho, & Loualiche (2022)

Our study of the link between breakthrough innovations and financial markets closely follows the approach of Haddad, Ho, & Loualiche (2022). [13] HHL investigate how financial markets value innovations, finding that during financial bubbles, patents issued by (Compustat) firms are valued 40% more by equity investors than observably equivalent patents issued during non-bubble periods. To this end, they combine data from Greenwood, Shleifer, Yang (2018) tagging equity bubbles at the Fama-French 49-industry level with the Kogan, Papanikolau, Seru, & Stoffman (2017) data on patent values. HHL find that the quantity of innovation as measured by patents filed per patent class moderately increases during bubbles, but they do not directly examine the effect of bubbles on patent quality. By contrast, we measure patent quality using the Kelly, Papanikolau, Seru, & Taddy (2021) data based on textual similarity. We then directly measure the association between patent quality and bubble periods at the 6-digit NAICS industry-level for all patenting firms, while HHL restrict themselves to patenting by Compustat firms.

1.2.1.3 Nanda & Rhodes-Kropf (2013)

Nanda & Rhodes-Kropf (2013) also provide evidence that “hot” financing periods in venture capital markets affect the quality distribution of funded firms. They operationalize “hot” periods simply as those when more firms receive VC funding in a given quarter. This variable is associated with higher failure rates, but also higher valuation conditional on an IPO as well as more (and more highly-cited) patenting. This suggests that venture capitalists engage in more risk-taking and experimentation during “hot” times (increased variance of outcomes) rather than simply funding worse firms (a leftward shift of the distribution). They confirm the robustness of this result using an instrumental variable that identifies exogenous increases in VC funding.⁶ This finding nicely illustrates one of the mechanisms we have in mind, but is restricted to 12,000 firms receiving first-stage funding between 1985 and 2004 (from the Dow Jones Venture Source database).

1.2.2 Theoretical Literature

This paper aims to contribute to the theoretical literature on financing new technologies, equity price booms and busts, and rational bubbles.

Nanda & Rhodes-Kropf (2013) [15] propose a model in which startups face two types of risk: fundamental risk and financing risk. The former is the probability that the technology or process fails, while the latter refers to a startup that would otherwise have a positive NPV but fails because not enough investors provide financing. This

⁶The instrument is fundraising by leveraged buyout funds, which are often invested in by institutional investors along with VCs, and hence are associated with increased capital for VC firms. But leveraged buyout funds plausibly have no direct effect on the outcomes of startups invested in by VC firms, as they generally invest in established firms.

self-fulfilling fear is most pronounced for the riskiest firms. The key implication is that it can be optimal to invest “with the crowd”, particularly for very innovative projects. Nanda & Rhodes-Kropf assume that rational expectations equilibria are selected by a commonly observable signal which follows a Markov process. The model which we explore below features a similar indivisibility of projects.

This paper also aims to contribute to the theory of rational bubbles, starting with the seminal work of Samuelson (1958) [10] and Tirole (1985) [11], who developed a model with deterministic bubbles that increase consumption but decrease capital accumulation. More recently, Martin and Ventura (2012) [9] documented the co-occurrence of financial bubbles and periods of high growth, and developed an overlapping generations model that allows for bubbles to appear and disappear stochastically. In contrast to the Samuelson-Tirole model, their approach allows for higher capital accumulation and output during bubble episodes. Our model builds on their theoretical framework.

Our main theoretical contribution is an extension of Martin and Ventura’s model to include risky investment decisions and diversification as in Acemoglu and Zilibotti [18]. The latter model features a safe and risky technology and a minimum size requirement function that prevents investors from fully diversifying risk. The model below incorporates risky investments in new technologies and demonstrates how bubbles can positively influence diversification. Pastor and Veronesi, 2009 [6] also show how the introduction of a new technology in an economy can generate a boom and bust in the new technology’s stock prices. Their model illustrates how bubbles can emerge when the economy experiments with new technologies, so that bubbles are caused by innovation but do not cause innovation. In contrast, our paper adopts a different perspective, aiming to show that financial

bubbles can increase investment in new technologies and diversification, thus raising the likelihood of successful investments.

1.3 Empirics

Throughout the empirical part of this paper, we adopt the Greenwood et. al. (2018) [19] method to tag rapid asset price run-ups which predict an increased probability of a crash; they (and we) call these episodes bubbles. We document the co-occurrence between bubbles and breakthrough innovation at the sectoral (6-digit NAICS) level. To do so, we combine a text-based measure of patent novelty from Kelly et. al (2021) with bubble tags at the Fama-French 49 industry-level.

1.3.1 Breakthrough Patents (KPST (2021))

Kelly et. al. (2021) leverage the entire text corpus of the universe of US patents (9 million patents from 1840 - 2010) to distinguish those innovations that are truly radical from the mass of prosaic improvements. Until now, the literature has relied on the number of citations a patent receives as a proxy for its contribution to the generation of further knowledge.⁷ Recent research has also calculated patent value using event studies of stock prices around patent approval announcements; this value dataset was used as a dependent variable in Haddad, Ho, & Loualiche's analysis, for example, KPST propose that a patent's contribution to knowledge can be measured by the ratio of its textual similarity to patents in the future (forward similarity, i.e. "impact") relative to its

⁷Hall, Jaffe, & Trajtenberg's NBER Patent Data Project (2001) represents perhaps the most prominent example of a tradition stretching back to at least Zvi Griliches, but one which has always acknowledged its limitations

backward similarity (“novelty”). In this way, Nikola Tesla’s 1888 patent was one of the first to use the phrase “alternating current”, which was then used widely in subsequent filings—a novel and impactful patent. KPST show that their measure of importance, the ratio of 10-year-forward to 5-year-backward similarity, is broadly correlated with citation-based measures and value-based measures.⁸ Moreover, it is available for all patents (rather than just patents of public firms, or patents after 1947 when citations began to be consistently recorded) and is less noisy than citation-based measures.

KPST provide aggregate and sectoral (up to NAICS 6-digit level) indices of breakthrough patents. They tag as breakthrough patents those in the top 10% of their importance measure. This aggregate series, as shown in Figure 1.1, highlights at least three main innovation waves, which accord with our preexisting intuition.⁹ 1880-1890 aligns with the Second Industrial Revolution and advances in communications and electricity. 1920-1935, broadly the “Roaring Twenties”, involves improvements in plastic and chemical manufacturing, while the IT and biotech revolution can be seen in the spike after 1985.

⁸This is true after removing issue-year fixed-effects, which capture mechanical changes like changes in language usage.

⁹In Appendix Table A.1, KPST provide a list of 250 top inventions as compiled by the USPTO and other online sources. They demonstrate that these align well with their quality measure. In Table 1.1, we display the highest quality invention for each decade to provide a flavor of the timing of the most important inventions.

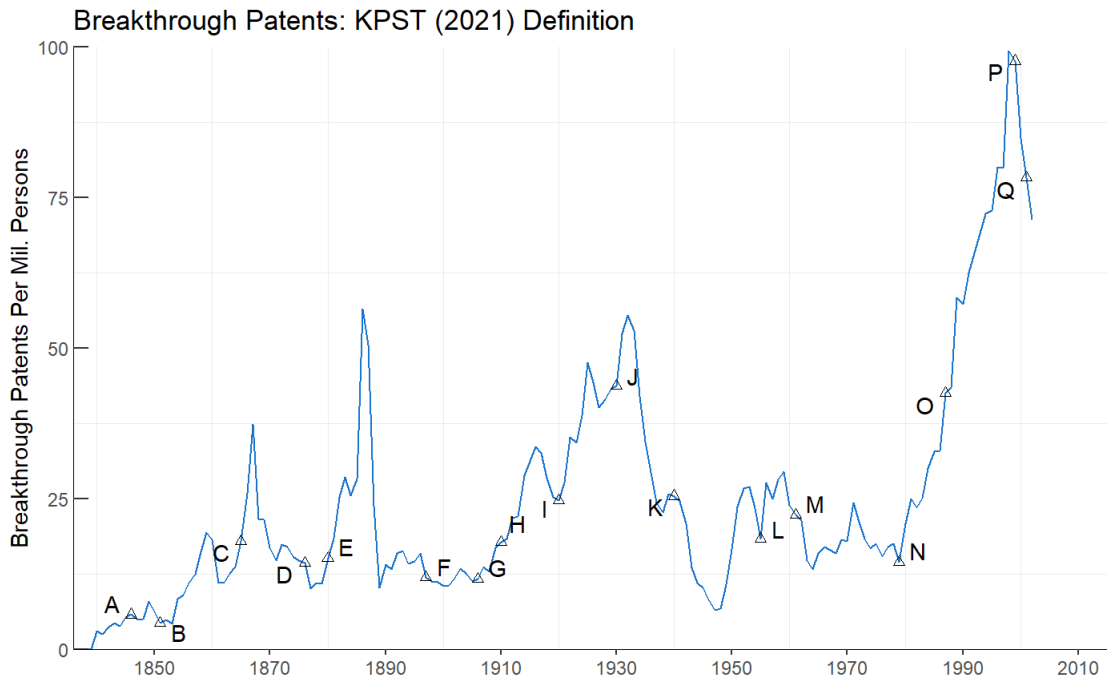


Figure 1.1: Breakthrough Patents Per Million Persons

Note: We highlight the top invention per decade of those compiled in KPST Appendix Table A.1.

	Patent #	Year	Inventor	Invention
A	4453	1846	Morse	Telegraph
B	8294	1851	Singer	Sewing Machine
C	46454	1865	Deere	Plow
D	174465	1876	Bell	Telephone
E	223898	1880	Edison	Incandescent Light
F	586193	1897	Marconi	Radio
G	821393	1906	Wright	Airplane
H	971501	1910	Haber	Ammonia Production
I	1360168	1920	Alexanderson	Antenna
J	1773980	1930	Farnsworth	Television
K	2188396	1940	Semon	Rubber
L	2708656	1955	Fermi	Atomic Reactor
M	2981877	1961	Noyce	Semiconductor Device
N	4136359	1979	Wozniak	Microcomputer with Video Display
O	4683195	1987	Mullis	Polymerase Chain Reaction
P	5960411	1999	Bezos	1-Click Buying
Q	6285999	2001	Page	PageRank

Table 1.1: Selected Top Inventions Per Decade

KPST further show that their indices are indeed associated with productivity growth: a one-standard deviation increase in the innovation index in a sector implies a 1 percent increase in sectorial TFP growth over the following five years.

1.3.2 Bubbles (Greenwood, Shleifer, Yang (2018))

We identify bubble episodes at the Fama-French 49-industry level using the GSY (2018) criterion [19], also used by Haddad, Ho, & Loualiche (2022). [13] GSY show that rapid equity price run-ups predict an increased probability of a crash (defined as a 40% fall in price within two years after the run-up). Moreover, other variables such as volatility and turnover can also help predict crashes.

For GSY, a bubble exists in period t in an industry if all the following three criteria are satisfied:

1. A value-weighted portfolio of the industry had a return of $\geq 100\%$ over the previous two years.
2. A value-weighted portfolio of the industry exceeded the market return by at least 100% over the past two years.
3. The value-weighted industry return over the past 5 years is larger than 50%.

A bubble is considered as having crashed if the value-weighted portfolio experiences a 40% fall in its price within a 2 year period. Perhaps episodes tagged as bubbles that do not crash should not be considered as bubbles, therefore in our empirical analysis we used all episodes tagged as "bubbles" and also only the ones that have crashed.

1.3.3 Descriptive Statistics

We map bubbles at the Fama-French 49 industry level to patent data at the NAICS-6 level using the Fama-French-to-SIC correspondence provided by Kenneth French, combined with the SIC-NAICS correspondence provided by the Census Bureau. The final dataset tracks breakthrough patents in 529 NAICS 6-digit sectors issued over the 40 years from 1963-2002.

Bubble?	Crashed?	Mean	Median	S.D.	Min	Max	N
No	No	0.06	0	0.42	0	15.94	20873
Yes	No	0.31	0	1.49	0	14.42	130
Yes	Yes	0.46	0	1.95	0	14.72	157
# NAICS	-	-	-	-	-	-	529
# Years	-	-	-	-	-	-	33

Table 1.2: Breakthrough Patents Per Capita by Group

As can be seen in Table 1.2, this procedure tags 287 (1.36%) of all sector-year observations as experiencing bubbles. Indeed, 30% of 6-digit NAICS experienced at least one bubble year. Of those sectors that did experience bubbles, the median experienced two years in a bubble, while some sectors experienced six total bubble years over the sample period.¹⁰ Figure 1.2 shows the time series of the number of sectors experiencing bubbles. Per-capita breakthrough patents in bubble sector-years are over 6 times as high relative to non-bubble times, albeit with a higher standard deviation.¹¹ Interestingly, they are higher still if we restrict ourselves to bubbles that eventually crashed. We explore this difference further in the next section.

¹⁰Note that these six years are not necessarily in one continuous bubble episode.

¹¹To be exact, we measure breakthrough patents per million persons.

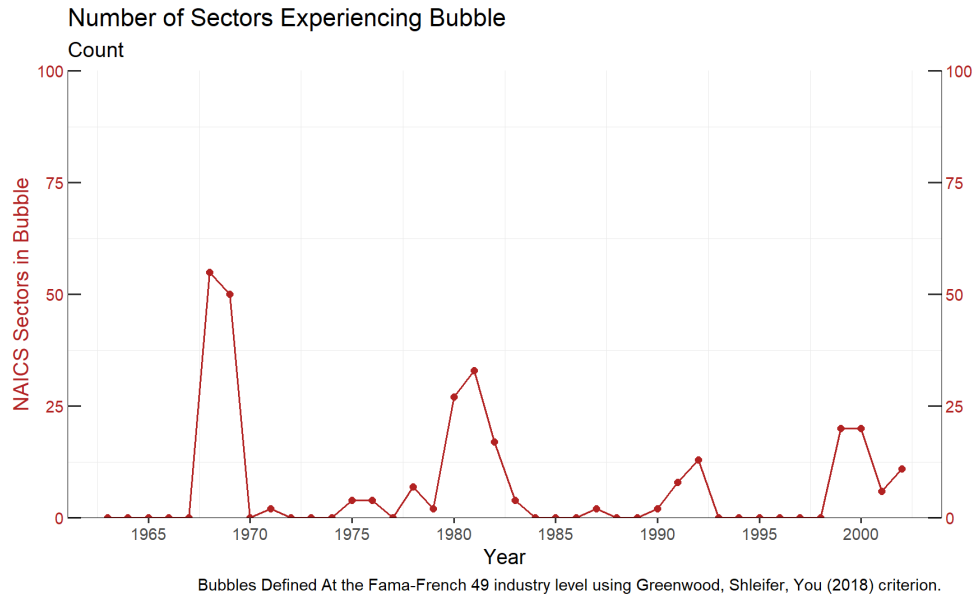


Figure 1.2: Number of Sectors Experiencing A Bubble

1.3.4 Sectoral Regressions

We run simple panel regressions, modeled on those run by HHL (Equation 1.1). bp_{jt} is breakthrough patents per million persons in sector j at time t . $B_{j,t-k}$ is the bubble indicator from Greenwood, Shleifer, Yang (2018), tagged at the (higher) Fama-French 49 industry level and mapped to the NAICS-6 level, which we lag to account for the fact that patenting ought to lag financing, as well as the fact that the KPST series uses issue year rather than application year. We also include lags of the dependent variable to account for the natural autocorrelation of breakthrough patenting. All regressions include time-fixed (λ_t) and sector-fixed (θ_t) effects and use Poisson maximum-likelihood estimation to account for the large number of zeros often found in patent data (at the six-digit NAICS level, many sector-years will have no patents).

$$bp_{jt} = \sum_{k=0}^{k=5} \beta_{t-k} B_{j,t-k} + \sum_{k=1}^{k=5} \psi_{t-k} bp_{j,t-k} + \theta_j + \lambda_t + \epsilon_{jt} \quad (1.1)$$

In the first column of Table 1.3, we regress the number of breakthrough patents per million persons against the occurrence of bubbles between zero and five years prior, as well as lagged breakthrough patents.¹² We observe a statistically significant effect of experiencing a bubble shortly preceding the occurrence of breakthrough innovations. A bubble in the preceding three years, for example, implies 0.5 more breakthrough patents per million persons in a given year (summing the relevant coefficients). This translates to roughly 150 more breakthrough patents in a 6-digit NAICS sector, given a total US population of 330 million today. This represents a roughly one standard deviation increase in breakthrough patents per sector-year. This is non-trivial if, as KPST suggest, a one s.d. increase in breakthrough patents implies a one percent increase in the sectoral TFP growth rate over the following five years.

Column 2 restricts the set of bubbles to those which subsequently crashed (157 sector-years out of 287). Here we observe significant effects at longer horizons prior to treatment—between two and five periods prior—but of similar magnitudes.

¹²In Appendix Section A.0.4 we run the same specification as a probit, where the outcome variable is having any breakthrough patenting at all in a given sector-year and the results are qualitatively similar (bubbles in year t-1 and t-2 have a positive and statistically significant effect on breakthrough patents)

Table 1.3: Panel Regressions

VARIABLES	Regression Type		
	(1) Bubble, HHL: All Bubbles	(2) Bubble, HHL: Crashes Only	(3) Bubble: CAPE Definition
	<i>Brkt. Per Mil. Persons</i>	<i>Brkt. Per Mil. Persons</i>	<i>Brkt. Per Mil. Persons</i>
Bubble (0)	0.111 (0.138)	-0.0922 (0.155)	0.197** (0.0813)
Bubble (-1)	0.102** (0.0427)	0.0120 (0.111)	0.0555 (0.0650)
Bubble (-2)	0.231*** (0.0851)	0.233*** (0.0871)	0.266*** (0.0889)
Bubble (-3)	0.124** (0.0581)	-0.000468 (0.0690)	0.111* (0.0616)
Bubble (-4)	-0.0587 (0.0568)	0.154** (0.0713)	0.137 (0.106)
Bubble (-5)	0.0596 (0.0921)	0.190* (0.105)	-0.366** (0.159)
Brkt. Pats. (-1)	0.136*** (0.0197)	0.135*** (0.0242)	0.122*** (0.0239)
Brkt. Pats. (-2)	0.0377*** (0.00777)	0.0539*** (0.0132)	0.0457*** (0.00884)
Brkt. Pats. (-3)	0.0732** (0.0335)	0.0558** (0.0276)	0.0859*** (0.0241)
Brkt. Pats. (-4)	0.0404** (0.0160)	0.0365*** (0.0131)	0.0425*** (0.00925)
Brkt. Pats. (-5)	-0.246*** (0.0713)	-0.224*** (0.0553)	-0.239*** (0.0581)
Observations	14,756	14,756	14,756

Note:

Robust standard errors in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

In the third column, we conduct a similar regression, but this time defining a sectoral bubble as a sector-year in which the median cyclically-adjusted price-earnings ratio (CAPE) in a given sector is above 26.1, placing it in the top percentile of sector-years in our data.¹³ Here we also observe significant effects of a bubble on breakthrough patenting, although the positive contemporaneous effect is troubling if bubbles can only cause increased innovation with a lag.

Tables 1.4 & 1.5 show that this measure of bubbles differs quite a bit from the HHL measure. For example, only 25 sector-years simultaneously experience a bubble and

¹³The sectoral CAPE data are available in the Wharton Research Data Services database at the Fama-French 49 level.

CAPE > 26	Mean	Median	S.D.	Min	Max	N
NO	0.07	0	0.49	0	15.94	17000
YES	0.04	0	0.15	0	1.35	180
N/A	0.32	0	1.44	0	14.72	277

Table 1.4: Breakthrough Patents per Capita: Alternative Bubble

CAPE > 26	HHL Bubble	N
NO	NO	16856
NO	YES	144
YES	NO	155
YES	YES	25
N/A	NO	264
N/A	YES	13

Table 1.5: CAPE Bubble vs. HHL Bubble

have CAPE > 26. Even more surprisingly, breakthrough patenting is actually *lower* on average in sector-years with CAPE > 26 than those without.

However, when controlling for lagged terms, the results are quite similar to those using the HHL bubble definition—an elevated CAPE at time t produces more breakthrough patenting in the near future.

1.3.5 Bubbles, IPO Size and Breakthrough Patents

In this section, we first show how bubbles in a sector can lead to larger Initial Public Offerings (IPOs). We then show the positive relationship between a larger offer size and the number of breakthrough patents produced by these companies.

We use the Firm Database of Emerging Growth Initial Public Offerings from Kenney and Patton (2014) [20], which includes all emerging growth IPOs between January 1990 and December 2010. They define emerging companies as “newly established firms, or

firms that are not based on older firms by being a spinoff or subsidiary operation". We merge this dataset with the 49 sectors bubbles dataset, and then fit the following regression:

$$OfferSize_{it} = \beta_0 + \beta_1 bubble_{j,t-1} + \beta_2 EBITDA_{i,t-1} + \theta_j + \lambda_t + \epsilon_{it} \quad (1.2)$$

where the *OfferSize* is the total amount offered in millions (number of shares times offer price) for company *i* in sector *j*, *bubble* is a dummy coded as 1 if there was a bubble in sector *j* at time *t* – 1, and *EBITDA*_{*i,t-1*} is lagged earnings before interest, taxes, depreciation and amortization in millions of dollars for company *i* taken from Compustat. θ_j and λ_t are sector and time-fixed effects respectively. We use lagged EBITDA as current end-of-the-fiscal-year EBITDA would not be available by the time of the IPO.

The summary statistics for our variables are presented in Table 1.6 and the regression results are presented in Table 1.7, indicating a significant positive association of offer size with bubbles, controlling for EBITDA. In bubble years, firms raise \$24 million more than observably similar firms would in non-bubble years. Of course, EBITDA does not control for unobserved firm quality not captured by earnings, so we cannot claim causality here.

Variable	Mean	P10	P50	P90	# Obs
Offer Size (in millions)	70.19	7.02	37.16	135.26	3210
EBITDA (in millions)	12.24	-11.14	1.30	28.44	3210
Breakthrough Patents	1.18	0	0	3	3210
Bubble (dummy)	0.041	–	–	–	–

Table 1.6: Summary statistics for Offer Size, EBITDA, Patents, and Bubble indicator

Table 1.7: Regression results of a Fama-French bubble on the IPO offer size

	<i>Dependent variable:</i>
	OFFER_SIZE_1000
$Bubble_{t-1}$	24.468** (11.647)
$EBITDA_{t-1}$	1.034*** (0.017)
Constant	-13.113 (61.971)
Observations	3,210
R ²	0.532
Adjusted R ²	0.532
Residual Std. Error	104.295 (df = 3207)
F Statistic	1,824.135*** (df = 2; 3207)
Sector Fixed Effects	YES
Year Fixed Effects	YES
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Next, to ask whether higher offer size leads to more breakthrough innovation, we match the IPO data to the breakthrough patenting data for each firm. We match half of the companies in our dataset and estimate the following model:

$$BRKTPatents_{i,t} = \alpha + \beta_1 PostIPO_{i,t} + \beta_2 (PostIPO_{i,t} \times OfferSize_i) + \gamma X + \tau_t + \epsilon_{i,t} \quad (1.3)$$

where $BRKTPatents_{i,t}$ is the number of breakthrough patents of company i in year t , $Post_{IPO}$ is an indicator variable equal to 1 if the company is in the post-IPO period in that year, $OfferSize$ denotes the company's IPO offer size in millions, X includes sector and firm fixed effects, and τ_t represents time fixed effects.

Table 1.8 presents the results, showing a statistically significant statistical effect of both offer size and the IPO on breakthrough innovation. After an IPO, companies produce an additional 0.167 breakthrough patents on average, and each additional million in funds raised is associated with an additional increase of 0.00004 breakthrough patents filed afterwards.

As we discuss in the model section below, larger offer sizes resulting from investors' willingness to hold the risky securities of newly floated firms during bubbles can be welfare-improving, by easing financial frictions in allocating capital to new, risky technologies and increasing diversification in these new sectors.

Table 1.8: Estimates of the Effect of Offer size of breakthrough patents

Dependent Variable:	Breakthrough
<i>Variables</i>	
postIPO	0.1673*** (0.0605)
postIPO \times Offer Size (millions)	0.0004*** (6.65×10^{-5})
<i>Fixed-effects</i>	
ff49	Yes
Filing year	Yes
Firm	Yes
<i>Fit statistics</i>	
Observations	30,198
R ²	0.34907
Within R ²	0.00170
<i>Signif. Codes: ***: 0.01, **: 0.05, *: 0.1</i>	

1.4 Model

In this section, we develop an overlapping generations model that allows for the stochastic creation and bursting of bubbles, which agents optimally hold. We also examine risky investments in innovative technologies, and how the level of diversification is endogenously determined. Analytically, we demonstrate that bubbles lead to a higher capital stock, higher wages, and increased levels of savings, which in turn lead to greater diversification.

In the model, we interpret the arrival of a GPT as the introduction of a new technology that transforms output into capital at a higher rate than the currently available technology. However, this new technology is risky, and sufficient investment is necessary to diversify the risk. For example, consider the discovery of quantum computing. With low investment,

only a few firms can be funded, making the success of quantum computing as a whole unlikely. In contrast, if more funds are allocated to the sector, more firms can be funded, thereby increasing the likelihood of success.

This model could be applied in an economy following the arrival of a new GPT, which creates the potential for risky new investments. Through bubbles, the amount that can be invested in this sector increases, thereby enhancing the level of diversification and reducing risk. If the sector becomes fully diversified, the technology could potentially become the new standard in the economy and lead to higher output until the next GPT arrives.

1.4.1 Household

Consider an overlapping generations model without population growth. Each generation comprises a unit measure continuum of risk neutral households.¹⁴ Households live for 2 generations and supply 1 unit of labor when young. Young households aim to maximize old-age consumption. As households only care about consumption in old age, they invest all of their labor income.

In the economy, there are two types of households: a fraction ϵ are the innovators, denoted as I , who can only invest in the new risky sector (explained below), while the remaining $1 - \epsilon$, denoted as unproductive investors U , can only invest in the old sector. The old sector yields a positive riskless return on any unit of output invested, denoted as q .

One key assumption, following Martin and Ventura (2012), is that we introduce a

¹⁴This is a simplifying assumption for now.

financial friction that prevents unproductive investors from directly lending money to the innovators for investment on their behalf. We can think of this as representing banks, who are prevented by regulations from investing directly in risky new technologies. As we will show, equity markets indirectly allow unproductive households to circumvent this friction by buying bubbles from the productive households.

1.4.2 Technology

The final goods sector produces output using labor and capital which for simplicity *fully depreciates every period*:

$$Y_t = k_t^\alpha l_t^{1-\alpha} \quad (1.4)$$

As markets are competitive and the young households provide one unit of labor, the factors of production are paid their marginal products:

$$w_t = (1 - \alpha)k_t^\alpha \quad (1.5)$$

$$r_t = \alpha k_t^{\alpha-1} \quad (1.6)$$

With this specification, the total savings rate in the economy is $s = (1 - \alpha)$.

There are two capital-producing technologies which transform savings from period $t - 1$ into capital that will be used at time t . There is the old and safe technology, which is already completely diversified and riskless. It takes 1 unit of final good today and transforms it into $q \geq 1$ units of capital tomorrow. There is also the new risky technology.

Within it, there is a continuum of business ideas indexed by $j \in [0, 1]$, where j represents a particular state of the world. Business idea j has a positive payoff Q if state j is realized and 0 otherwise, where $Q \gg q$.

Note that we assume that only one idea j will succeed at any time t and the rest will fail. This is a simplifying assumption (one could also assume that a positive measure of ideas pay off at any time t). The important feature, explained below, is that innovators must bear some risk to get Q , although this risk does not affect their investment decision as they are restricted to invest in the new technology.

1.4.3 Risk and Total Investment

If an individual invests in only one idea, the probability of success is effectively 0. Individuals can diversify their risk by investing in a subset of ideas \bar{J} over $[0, 1]$, which yields a positive return with probability equal to the measure of \bar{J} .

To avoid full diversification, we assume a minimum size requirement $M(j)$ on *aggregate* investment in idea j . Idea j only pays Q if state j is realized **and** the minimum size requirement is met. If aggregate household investment is less than $M(j)$, the payoff of idea j will be 0. This is similar to the work of Nanda & Rhodes-Kropf [21], in which each investor can only fund a fraction of the total investment that will make a project viable.

$M(j)$ is defined as below, where D is a constant:

$$M(j) = \max \left\{ 0, \frac{D}{(1 - \gamma)}(j - \gamma) \right\} \quad (1.7)$$

For any $j \leq \gamma$, there is no minimum size requirement: the idea is viable at any level of investment. For $j \geq \gamma$ the minimum size requirement increases linearly. We can see the minimum size requirement function depicted in Figure 1.3.

1.4.4 Financial Intermediaries

We assume free entry into intermediation for the investment in the risky technology, where intermediation itself is a costless activity. Each intermediary can only collect funds to invest in a single idea j , issuing Arrow securities which pay Q units of the final good if state j is realized and nothing otherwise.¹⁵ Intermediaries must raise enough money to cover the minimum size requirement $M_t(j)$ for a given idea in order to invest in it. Critically, if there is not enough demand for investment in a particular idea in equilibrium, that security is **not** supplied (that idea is not "open"). This *endogenous* market incompleteness as a function of total investment in risky assets is a key driver of the model's behaviour.

Note that the price of any such security must equal 1 unit of the final good. The reasoning is as follows.

1. Suppose $P_t(i, j) < 1$. The intermediary returns Q units of the final good per security, so this would be loss-making for the intermediary in the event of success.
2. Suppose $P_t(i, j) > 1$. The intermediary would make positive profits in the event of success. Free entry into intermediation implies this cannot be the case—competition drives P to 1.

¹⁵Acemoglu & Zilibotti (1997) show that this assumption can be relaxed.

1.4.5 Optimal Investment

Innovators have to decide how much to invest in each idea j that is currently open. This investment is denoted by $I(j, t)$. Given that all households face the same price and that each state j returns Q , it is optimal for innovators to spread their investment evenly across all open markets, or in other words, to buy a perfectly "balanced portfolio".¹⁶ This portfolio is depicted in figure 1.3. The height of the shaded rectangle is the amount invested for each individual idea j , and the area is the total amount invested, which optimally is equal to the wage that the innovator receives plus the bubbles b_t^I they are able to sell (detailed below).

The set of open investments is denoted by n_t^* and a idea must be open for the innovator to be able to invest in it. The optimal investment in each idea when the entrepreneur takes the number of open markets n_t^* as given is:

$$I^*(j, t) = \frac{\epsilon s k_t^\alpha + b_t^I}{n_t^*} \quad (1.8)$$

1.4.6 Bubbles

We introduce bubbles following Martin and Ventura (2012). Bubbles are akin to pyramid schemes: they are an asset which incurs no cost to produce, and they start and end stochastically. They lack any productive use, and households purchase them only

¹⁶Since for now we assume that households are risk neutral, they are in fact indifferent between funding a small measure of sectors with more money and spreading their funds over all available ideas. Here, we assume that they will choose the second option.

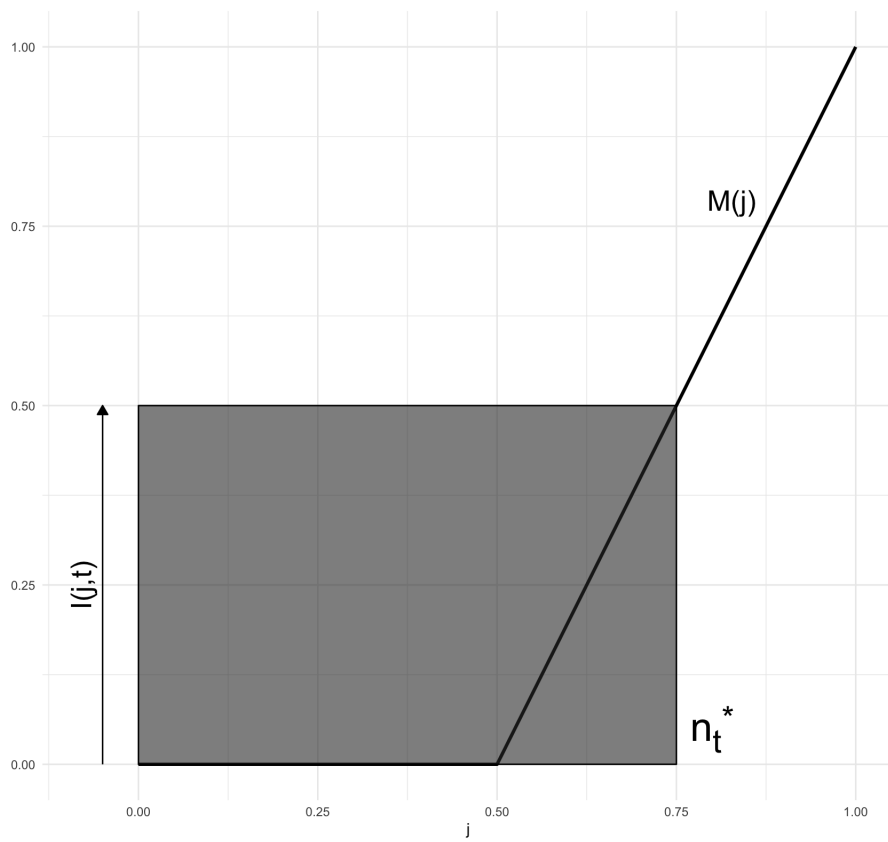


Figure 1.3: Minimum size requirement and optimal portfolio

because they believe that they can be resold in the future at a higher price. The creator of the bubble receives all the proceeds from its initial sale as profit. Consequently, any subsequent buyer essentially purchases the right to sell the asset later and receive the proceeds from the next buyer.

In any given period, there are two possible sellers of bubbles:

1. Old agents who are selling bubbles bought when they were young.
2. Young agents (innovators or unproductive investors) who are able to create new bubbles.

Bubbles are only bought by the young generation, as the old generation simply consumes and dies. As detailed below, either type of young agent can buy bubbles.

Define b_t as the market value of all old bubbles, and b_t^I and b_t^U as the market value of new bubbles created by innovators and unproductive investors, respectively.

We assume that $Q\gamma > q$, ensuring that $Qn_t^* \geq q$ in every period, and that the innovator takes n_t^* as given. In order for bubbles to exist, i.e $b_t + b_t^I + b_t^U > 0$, the following conditions have to hold:

$$E_t \left\{ \frac{b_{t+1}}{b_t + b_t^I + b_t^U} \right\} = \begin{cases} q\alpha k_{t+1}^{\alpha-1} & \text{if } \frac{b_t + b_t^I}{(1-\epsilon)sk_t^\alpha} < 1 \\ \in [q\alpha k_{t+1}^{\alpha-1}, n_t^* Q\alpha k_{t+1}^{\alpha-1}] & \text{if } \frac{b_t + b_t^I}{(1-\epsilon)sk_t^\alpha} = 1, \\ n_t^* Q\alpha k_{t+1}^{\alpha-1} & \text{if } \frac{b_t + b_t^I}{(1-\epsilon)sk_t^\alpha} > 1 \end{cases} \quad (1.9)$$

$$0 \leq b_t \leq sk_t^\alpha \quad (1.10)$$

Where $E_t \left\{ \frac{b_{t+1}}{b_t + b_t^I + b_t^U} \right\}$ represents the expected return of holding a bubble. An investor

must receive a payoff that is at least equivalent to the return on investing in capital for it to be worthwhile to purchase a bubble.

The initial purchasers of bubbles are unproductive young households. They possess savings equal to $(1 - \epsilon)sk_t^\alpha$. Therefore, as long as this value exceeds the market value of existing bubbles plus new bubbles created by the innovators, or in other words if $\frac{b_t + b_t^I}{(1 - \epsilon)sk_t^\alpha} < 1$, the marginal purchaser will be an unproductive household, and the expected return on bubbles must equal the return on investing in the old technology.

Once the market value of bubbles exceeds the savings of unproductive households, i.e., when $\frac{b_t + b_t^I}{(1 - \epsilon)sk_t^\alpha} > 1$, the marginal purchaser will be an innovator, and the expected return on bubbles must equal the return on the new technology. The value of bubbles should be big enough so that unproductive households are able to purchase them from productive ones, but on the other hand, the value of existing bubbles can not be so high that the young generation does not have enough wage income to purchase them, as shown in equation 1.10.

The expected return to investing in capital equals the rental rate times the expected return on the technology. For an unproductive investor, this is $q < 1$. For innovators, the return to holding capital is stochastic; with probability n_t^* —the measure of the open ideas that the innovator invests in—they will receive Q times the rental rate, and zero otherwise.

1.4.7 Capital

Now we can describe how the capital stock evolves.

1.4.7.1 Case 1

If the marginal investor in the bubble is unproductive, the expected next period capital is determined by:

$$E[k_{t+1}] = n_t^* QI^*(j, t) + q[(1 - \epsilon)sk_t^\alpha + b_t^U - b_t - b_t^I - b_t^U] \quad (1.11)$$

With probability n_t^* , innovators will get $QI^*(j, t)$ units of capital. The unproductive investor has income $(1 - \epsilon)sk_t^\alpha$ but on net buys $b_t + b_t^I$ and receives q on the total amount invested.

1.4.7.2 Case 2

When the marginal investor is the innovator, unproductive investors do not build capital. They simply buy bubbles, given that the expected return is higher than what they would get by saving in capital. Therefore, next period's expected capital is given by:

$$E[k_{t+1}] = n_t^* QI(j, t) - \{[b_t + b_t^I + b_t^U] - [(1 - \epsilon)sk_t^\alpha + b_t^U]\} \quad (1.12)$$

This is the expected return on the innovator's investment $n_t^* QI^*(j, t)$ minus the bubbles innovators buy, which equals the total amount of bubbles bought in the economy $[b_t + b_t^I + b_t^U]$ less the bubbles bought by unproductive investors, which in turns equals their total saving $[(1 - \epsilon)sk_t^\alpha + b_t^U]$.

Hence, combining equation 1.8, 1.11 and 1.12, the expected next period's capital

stock is then:

$$E[k_{t+1}] = \begin{cases} [Q\epsilon + q(1 - \epsilon)]sk_t^\alpha + [Q - q]b_t^I - qb_t & \text{if } \frac{b_t + b_t^I}{(1 - \epsilon)sk_t^\alpha} < 1 \\ [Q\epsilon + (1 - \epsilon)]sk_t^\alpha + [Q - 1]b_t^I - b_t & \text{if } \frac{b_t + b_t^I}{(1 - \epsilon)sk_t^\alpha} \geq 1 \end{cases} \quad (1.13)$$

There are two effects of bubbles on the capital dynamics.

First, there is a *crowding out effect*: existing b_t increases current consumption of the old but reduces real investment: the $-b_t$ term. However, existing bubbles crowd out unproductive investment first, increasing average investment efficiency. This can be seen by comparing Case 1 in Equation 1.13 to Case 2. Case 2 only binds when existing bubbles plus new bubbles are large enough to absorb the entire savings of the unproductive investors.

Second, there is a *reallocation effect*: when an innovator sells bubbles to an unproductive household, innovation investment replaces unproductive investment, speeding up diversification and capital accumulation. We can see this by comparing the $[Q - q]b_t^I$ term in Case 1 with $[Q - 1]b_t^I$ in Case 2, where $q < 1$.

1.4.8 Level of Diversification

Finally, the optimal n_t^* is located at the intersection between the investment function and the minimum size requirement. By equating equation 1.8 and equation 1.7 we get that the optimal number of ideas open at any time t is:

$$n_t^* = \frac{\gamma + [\gamma^2 + 4D^{-1}(\epsilon sk_t^\alpha + b_t^I)(1 - \gamma)]^{1/2}}{2} \quad (1.14)$$

This function is increasing in both k_t and b_t^I , demonstrating that the higher the value of new bubbles, the more diversification increases and the higher the likelihood of successful innovation compared to the equilibrium state without bubbles.

1.4.9 Recursive problem

We can transform equation 1.9 into a recursive problem by rewriting it in terms of the fraction of savings used to purchase bubbles b_t from the old, rather than in terms of the return on bubbles. (Note that equation 1.9 does not express variables at time $t + 1$ on the left-hand side in terms of the same variables in time t on the right).

Define $x_t = \frac{b_t}{sk_t^\alpha}$, $x_t^I = \frac{b_t^I}{sk_t^\alpha}$ and $x_t^U = \frac{b_t^U}{sk_t^\alpha}$. We can rewrite (1.9) as follows (the full proof is in the appendix).

$$E_t \{x_{t+1}\} = \begin{cases} \frac{\alpha q(x_t + x_t^I + x_t^U)}{s\{[Q\epsilon + q(1-\epsilon)] + [Q-q]x_t^I - qx_t\}} & \text{if } \frac{x_t + x_t^I}{(1-\epsilon)} < 1 \\ \in \left[\frac{\alpha q(x_t + x_t^I + x_t^U)}{s\{[Q\epsilon + q(1-\epsilon)] + [Q-q]x_t^I - qx_t\}}, \frac{n_t^* Q\alpha(x_t + x_t^I + x_t^U)}{s\{[Q\epsilon + 1 - \epsilon] + x_t^I[Q-1] - x_t\}} \right] & \text{if } \frac{x_t + x_t^I}{(1-\epsilon)} = 1, \\ \frac{n_t^* Q\alpha(x_t + x_t^I + x_t^U)}{s\{[Q\epsilon + 1 - \epsilon] + x_t^I[Q-1] - x_t\}} & \text{if } \frac{x_t + x_t^I}{(1-\epsilon)} > 1 \end{cases} \quad (1.15)$$

$$0 \leq x_t \leq 1 \quad (1.16)$$

Similarly, we can rewrite the equation for the evolution of capital as follows:

$$E[k_{t+1}] = \begin{cases} \{[Q\epsilon + q(1 - \epsilon)] + [Q - q]x_t^I - qx_t\}sk_t^\alpha & \text{if } \frac{x_t + x_t^I}{(1-\epsilon)} < 1 \\ \{[Q\epsilon + 1 - \epsilon] + x_t^I[Q - 1] - x_t\}sk_t^\alpha & \text{if } \frac{x_t + x_t^I}{(1-\epsilon)} \geq 1 \end{cases} \quad (1.17)$$

1.4.10 Existence of bubbles

In this section, we derive the conditions on parameters $\{\alpha, \epsilon, q, Q, s\}$ to ensure the existence of bubbles in equilibrium. This ensures that there exists a sequence $\{x_t\}$ s.t. equations 1.15 and 1.16 both hold $\forall t$.

1.4.10.1 Simple Case: No Further Bubble Creation

We start with the simplest case in which there is no new bubble creation after a bubble appears in t_0 , in other words $x_{t_0}^I > 0$, $x_{t_0}^U > 0$ and $x_t^I = x_t^U = 0$ afterwards.

Equation 1.15 then becomes:

$$E_t \{x_{t+1}\} = \begin{cases} \frac{\alpha q(x_t)}{s\{[Q\epsilon + q(1-\epsilon)] - qx_t\}} & \text{if } \frac{x_t}{(1-\epsilon)} < 1 \\ \in \left[\frac{\alpha q(x_t)}{s\{[Q\epsilon + q(1-\epsilon)] - qx_t\}}, \frac{n_t^* Q\alpha(x_t)}{s\{[Q\epsilon + 1 - \epsilon] - x_t\}} \right] & \text{if } \frac{x_t}{(1-\epsilon)} = 1, \\ \frac{n_t^* Q\alpha(x_t)}{s\{[Q\epsilon + 1 - \epsilon] - x_t\}} & \text{if } \frac{x_t}{(1-\epsilon)} > 1 \end{cases} \quad (1.18)$$

We can derive that an initial bubble can be an equilibrium bubble as long as $\alpha < s \frac{Q\epsilon + q(1-\epsilon)}{q}$.¹⁷

Panel 1 of Figure 1.4 plots $E[x_{t+1}]$ in Equation 1.18 for the case $\alpha \geq s \frac{Q\epsilon + q(1-\epsilon)}{q}$,

¹⁷See Appendix Section A.0.2 for the full derivation.

which implies the slope at the origin is greater than 1, so any initial bubble would eventually violate equation 1.16.

In Panel 2, $\alpha < s \frac{Q\epsilon + q(1-\epsilon)}{q}$ implies the slope at the origin is less than 1. For bubbles to exist we need a low enough α that leads to a higher savings rate and along with a high Q/q and ϵ . This condition implies that $\mathbb{E}x_{t+1}$ only crosses the 45 degree line once, as it is a strictly increasing function of x_t . Any initial bubble larger than this crossing point can be ruled out, while any initial bubble below it can be an equilibrium bubble, as it would satisfy equation 1.15 and 1.16 in all periods as x_t tends to zero over time.

1.4.10.2 Adding Bubble Creation

We now allow for bubble creation by the productive investor.¹⁸ We must consider two possibilities in turn: $\frac{Q\epsilon + q(1-\epsilon)}{Q} > 1 - \epsilon$ and $\frac{Q\epsilon + q(1-\epsilon)}{Q} \leq 1 - \epsilon$.

$\frac{Q\epsilon + q(1-\epsilon)}{Q} > 1 - \epsilon$: If $\frac{Q\epsilon + q(1-\epsilon)}{Q} > 1 - \epsilon$, bubble creation merely shifts up $\mathbb{E}x_{t+1}$ for all x_t , therefore it does not relax the conditions of existence as $\mathbb{E}x_{t+1}$ would not cross the 45 degree line. See Appendix Section A.0.3 for the proof.

$\frac{Q\epsilon + q(1-\epsilon)}{Q} \leq 1 - \epsilon$: In this case, bubble creation by the productive investor implies that there can exist a crossing point for $\mathbb{E}x_{t+1}$ even if the slope at the origin is greater than 1.

We proceed by considering separately how bubble creation affects $\mathbb{E}x_{t+1}$ in two regions of the x-axis: $[\frac{Q\epsilon + q(1-\epsilon)}{Q}, 1 - \epsilon]$ (our "region of interest") and its complement.

The proof that bubble creation merely shifts up $\mathbb{E}x_{t+1}$ outside the region of interest is in

¹⁸Bubble creation by unproductive investors would simply shift upwards the curves plotted in Figure 1.4 as X_t^U is only present in the numerator of equation 1.15. Bubbles created by the unproductive investor only have one effect: they compete with the old bubbles, decreasing their return.

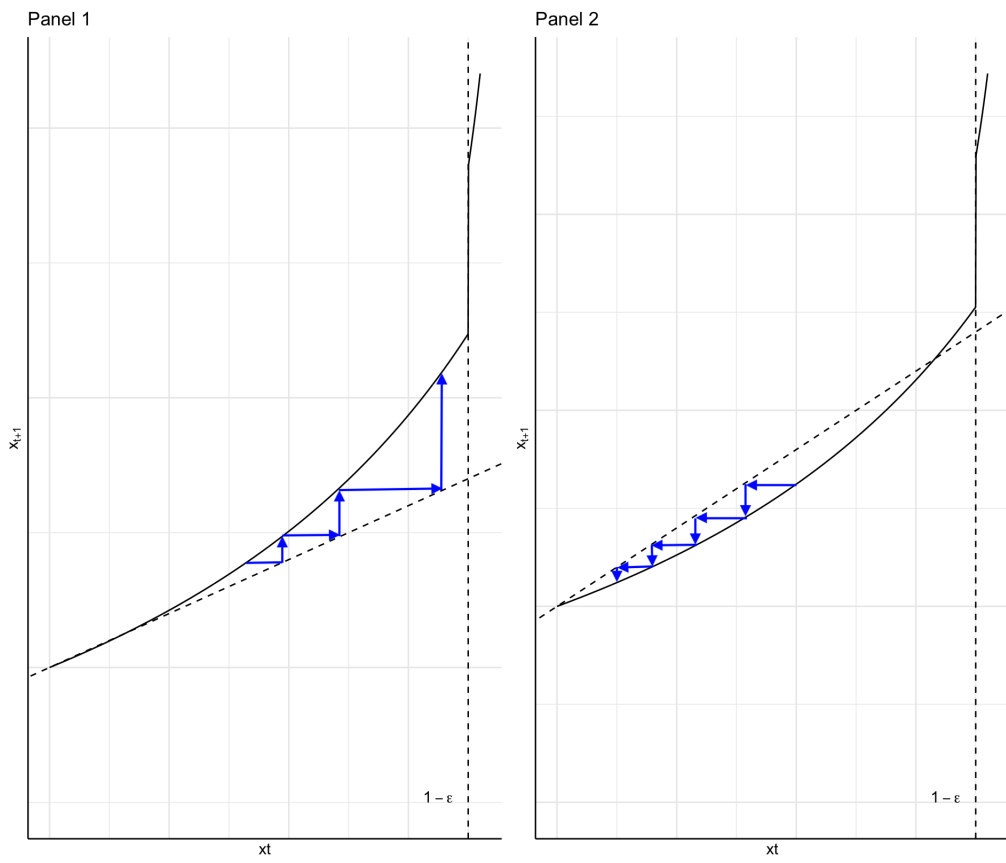


Figure 1.4: Existence

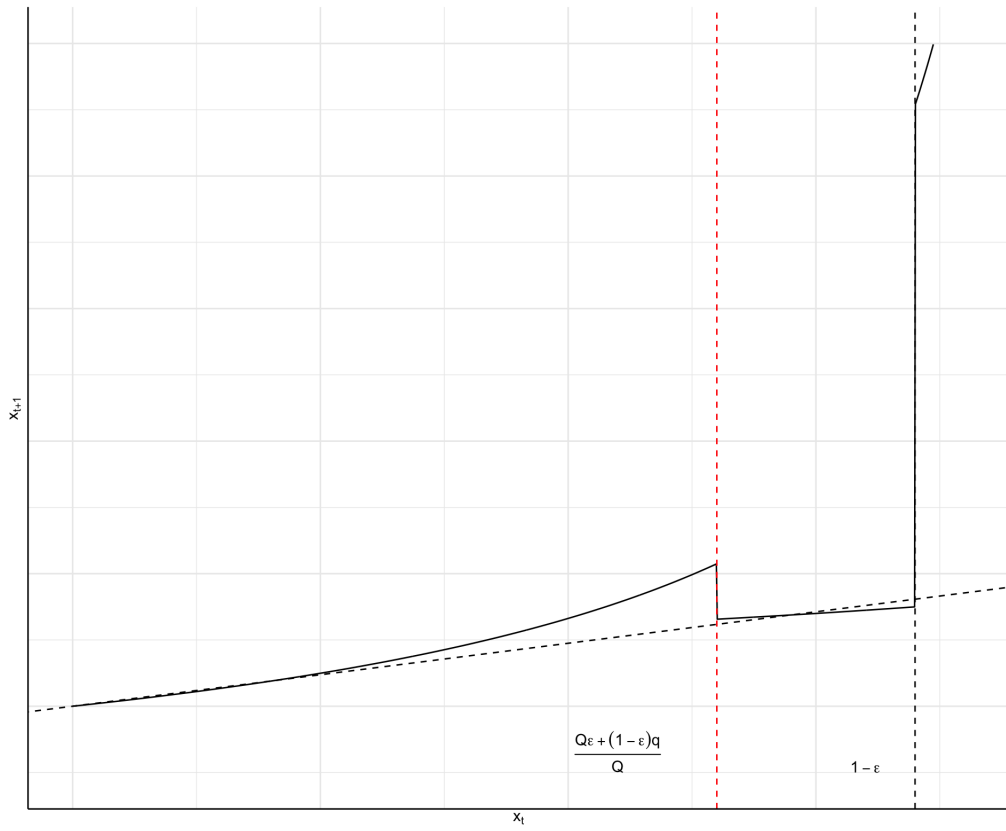


Figure 1.5: Existence

Appendix Section [A.0.3](#).

For simplicity, let $x_t^I = 1 - \epsilon - x_t$: i. e. the innovators sell all their bubbles to the unproductive.

Top line of the right hand side of equation [1.15](#) becomes:

$$\frac{\alpha q(x_t + 1 - \epsilon - x_t)}{s[Q\epsilon + q(1 - \epsilon) + (Q - q)(1 - \epsilon - x_t) - qx_t]} \quad (1.19)$$

which simplifies to

$$\frac{\alpha q(1 - \epsilon)}{s(Q - Qx_t)} \quad (1.20)$$

Now we find the root of this equation, to find the crossing point:

$$\frac{\alpha q(1 - \epsilon)}{s(Q - Qx_t)} = x_t$$

$$Qx_t^2 - x_tQ + \frac{\alpha q(1 - \epsilon)}{s} = 0$$

Define $c = \frac{\alpha q(1 - \epsilon)}{s}$. Then the root exists iff

$$Q^2 - 4Qc > 0$$

which implies that

$$1 > 4 \frac{\alpha q(1 - \epsilon)}{s}$$

or equivalently

$$\alpha < \frac{s}{4q(1 - \epsilon)}$$

Hence, the complete condition for existence of bubbles is as follows, in which bubble creation allows for a relaxation of the restriction on α if $\frac{Q\epsilon+q(1-\epsilon)}{Q} \leq 1 - \epsilon$:

$$\alpha < \begin{cases} s \frac{Q\epsilon+q(1-\epsilon)}{q} & \text{if } \frac{Q\epsilon+q(1-\epsilon)}{Q} > 1 - \epsilon \\ \max\left\{s \frac{Q\epsilon+q(1-\epsilon)}{q}, \frac{s}{4q(1-\epsilon)}\right\} & \text{if } \frac{Q\epsilon+q(1-\epsilon)}{Q} \leq 1 - \epsilon \end{cases} \quad (1.21)$$

As we can see in Figure 1.5, without $\frac{Q\epsilon+q(1-\epsilon)}{Q} \leq 1 - \epsilon$, an equilibrium bubble could not exist (as the slope at the origin is greater than 1).

These conditions highlight the importance of the difference in returns between the risky and safe sectors and the mass of entrepreneurs in the economy. If we take $\alpha = 0.4$ as given, then the savings rate is determined as $s = 0.6$, and the conditions for existence are defined by three parameters: the fraction of entrepreneurs (ϵ), the return on the risky investment Q , and the return on the safe investment q . The numerator $Q\epsilon + q(1 - \epsilon)$ represents the weighted sum of possible returns in the economy. To sustain bubbles in the model, ϵ and Q must be sufficiently large relative to q .

1.5 Simulation

Define B as a state in which the economy has a bubble, and F as a non-bubbly (fundamental) state, i.e. $x_t = x_t^I = x_t^U = 0$.

Let $z_t \in \{F, B\}$ be a random variable with transition probabilities $P[z_{t+1} = B | z_t = F] = p_1$ and $P[z_{t+1} = F | z_t = B] = p_2$.

Assume $z_t = B$, $x_t^I = \eta_0 + \eta_1 x_t$ and $x_t^U = 0$. Finally, from equation 1.15, we define:

$$x_{t+1} = \frac{\alpha q(x_t + \eta_0 + \eta_1 x_t)}{s(1 - p_2)\{Q\epsilon + q(1 - \epsilon) + [Q - q](\eta_0 + \eta_1 x_t) - qx_t\}} + u_t \quad (1.22)$$

Where u_t is a discrete random variable taking values of $-\sigma$ and σ with equal probability. We introduce this shock following Martin and Ventura. It can be thought of as a small shock to investors' sentiment within a bubbly episode, as distinct from z_t , which is a drastic shock that starts and ends bubbles.

Finally, aggregate consumption is defined by:

$$c = (\alpha + s * x_t)k^\alpha \quad (1.23)$$

1.5.1 Always "Lucky" Steady State

First we analyze the steady state of the economy supposing the economy is always lucky, meaning that in every period, the realized state j corresponds to an idea that was funded. Figure 1.6 shows the path to the steady state of key variables in this example, using the parameters shown in table 1.9:

The parameters ϵ , q , Q were chosen so the steady state value of n_t in the bubbly state is above 0.9, while the rest of parameters are the same as those used by Martin and Ventura.

In Figure 1.6, we compare two steady states: one in which there are always bubbles and there is never a transition towards the fundamental state, and the other in which bubbles can occur but do not. We can see that when bubbles exist in the economy, output, capital, consumption and n_t are significantly higher, which illustrates the prediction

Table 1.9: Parameters

Symbol	Value
α	0.4
ϵ	0.3
q	0.9
Q	12
γ	0.5
D	1
η_0	0.15
η_1	0.18
σ	0.0035
p_1	0.11
p_2	0.11

of the theoretical model that with bubbles, the steady state degree of diversification is significantly higher. This is due to the *reallocation effect* being stronger than the *crowding out effect* for our parameterization. We also plot the path of x_t to its steady state, and as we proved above this equilibrium exists for our set of parameters.

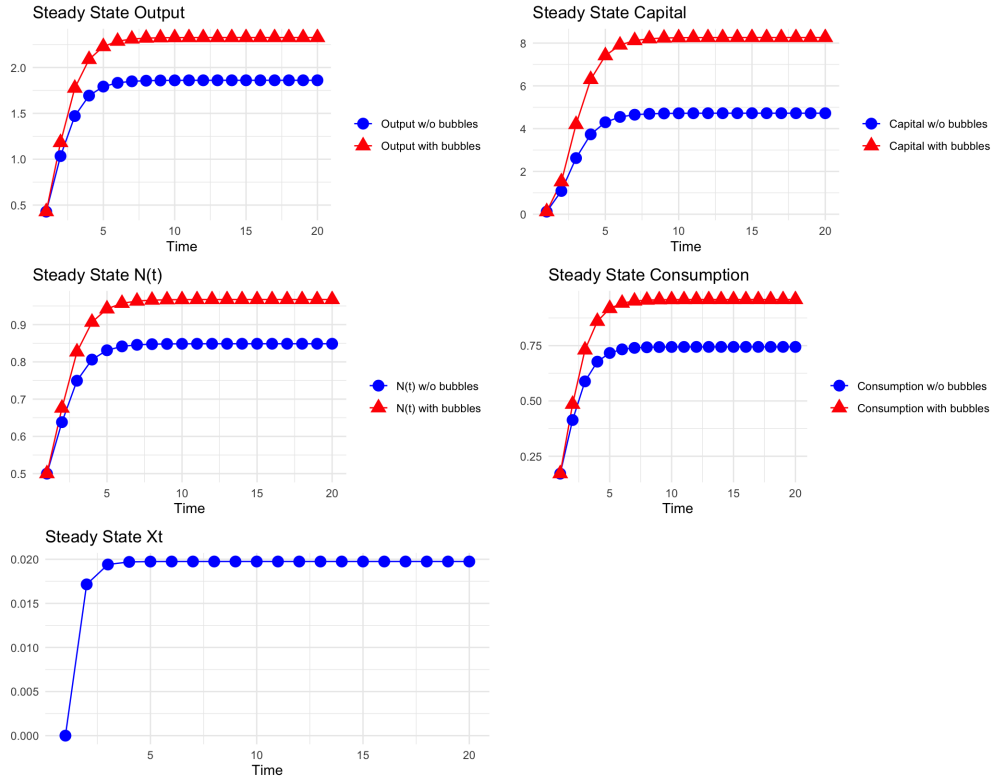


Figure 1.6: Path to Steady State

1.5.2 Fully Stochastic Economy

Next we simulate the economy over time, accommodating two stochastic variables: bubble state transitions and the random realizations of state j . The outcomes of this simulation are depicted in Figure 1.7.

In the first row, we simulate the economy starting with $z_0 = F$ while allowing the generation of bubbles ($p_1 = 0.11$). In the second row, we rule out the generation of bubbles but feed the same series of state realizations j as in the first row.

The economies in the two rows are the same until period 5, when the first economy experiences a bubble. The bubble crowds out saving in capital by the unproductive investors, but reallocates some of that saving to the innovators. This is because unproductive

investors purchase bubbles sold by productive investors. Innovators are able to channel this investment into the risky sector, which yields high returns. These high returns allow for more investment and greater diversification (n_t) in the risky sector, further increasing the probability that any state j will result in a successful innovation. This positive feedback loop continues until the bubble bursts in period 15. The level of investment collapses, and hence so does output (given full depreciation). Lower investment in the risky sector implies lower diversification, which means a larger measure of shocks j are unlucky. One of these unlucky states does hit in period 18, further lowering output and consumption.

Without the bubble but with a series of good draws, the no-bubble economy grows until it experiences a bad draw. Relative to the bubble economy, savings are allocated to capital rather than bubbles. However, a smaller fraction of aggregate saving is invested in the risky sector. With lower investment in the risky sector, an unlucky draw is more likely in the no-bubble economy and a lower level of $n(t)$. This leads to a negative feedback loop of lower investment, lower output, and a higher probability of bad draws given the state realizations.

Bubbles, then, increase welfare in this economy by promoting (aggregate) investment in the risky sector, which makes successful breakthrough innovations more likely.

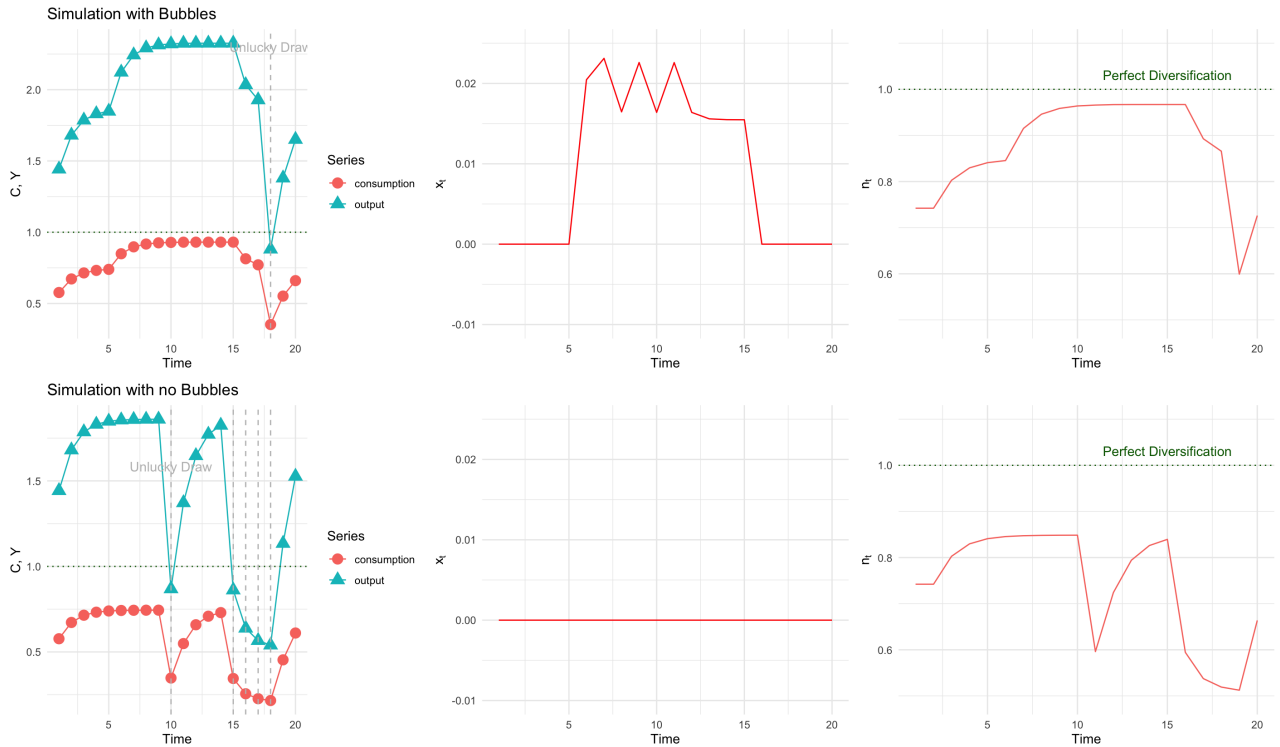


Figure 1.7: Simulation with bubbles and without bubbles, given same draws of shocks to risky technology.

1.6 Conclusion

In this paper, we document the co-occurrence of financial bubbles and breakthrough innovations. We then propose a mechanism by which bubbles cause (although are not caused by) breakthrough innovation, which improves welfare. Bubbles crowd out saving in capital, but also enable productive investors to access more savings to invest in the risky sector. This, in turn, enables investment in more risky ideas, increasing the probability of successful breakthrough innovations in the risky sector. Successful innovations further enable capital accumulation and diversification. An economy with bubbles generates a higher steady-state level of output and consumption and a lower risk of unsuccessful investments.

Chapter 2: Immigration Status and Access to Mortgage Financing

2.1 Introduction

According to the 2022 Census [22], the foreign-born population in the U.S. was 46.2 million, accounting for 13.9% of the total population. Upon arrival, U.S. immigrants typically encounter a more developed financial system than in their home countries, and good access to this system can significantly ease their integration process. Becoming a U.S. citizen greatly improves access to the financial system, as citizens can freely work and obtain a Social Security Number. However, the path to citizenship can take many years or even decades. Unauthorized immigrants can still apply for credit using an ITIN (Individual Taxpayer Identification Number), but it is not clear how much access they have to the financial system or how easily they can obtain credit.

Paulson et al. (2006) [23] show that U.S. immigrants have lower financial access, with only 63% of immigrant households having a checking account compared to 76% of native-born households.

In developing economies, access to the financial sector is extremely limited. For example, in Mexico, only 25% of the population has a bank account, in Venezuela 29%, and in Nicaragua only 4.7% (Beck, Demirguc-Kunt, and Martinez, 2007) [24]. Paulson and Okonkwo (2008) [25] show that immigrants from countries with better institutions

are more likely to participate in financial markets in the U.S. than those from countries with weaker institutions, which significantly impacts their full integration into the U.S. economy.

Being a migrant in the U.S. can impose significant financial barriers to accessing credit markets. For example, migrants' ability to search for credit options may be limited or even nonexistent, and research shows that a lack of search can negatively affect the interest rates they receive (Buttha, Fuster, and Hizmo, 2024 [26]; Alexandrov and Koulayev, 2018 [27]). Additionally, immigrants might be perceived by banks as a riskier population to lend to, either due to the potential risk of deportation or a higher perceived likelihood of default. This is especially problematic for unauthorized immigrants, and most banks do not offer them financing at all.

When an immigrant is documented and has work authorization, they can apply for a Social Security Number (SSN), which can be used for identification purposes when accessing banking and financing. Unauthorized immigrants are not eligible for an SSN, but still have tax obligations under US law. They can apply for an Individual Taxpayer Identification Number (ITIN) to pay taxes on income. Most lenders use SSNs to identify borrowers and access their credit histories, but some lenders also accept ITINs and extend credit to individuals with ITINs. Unauthorized immigrants often face additional challenges in providing income documentation, as they are frequently paid in cash.

Goodman et al. (2024) [28] conducted interviews with ITIN lenders in the mortgage market and estimated that there were a mere 5,000 to 6,000 ITIN mortgages approved in the US in 2023. In 2014, 5.25 million tax returns were filed using ITIN numbers out of 21 million in circulation. 90% of ITIN numbers were issued to people from Mexico

(72%), Guatemala, Honduras, and El Salvador. Surprisingly, the study highlighted that ITIN mortgages had lower delinquency rates compared to the Mortgage Bankers Association's delinquency rates for the mortgage market overall. Lenders identified the lack of a secondary market as the most significant limitation to the expansion of this market. Relatedly, the Federal Housing Administration (FHA) requires non-residents to have a Social Security Number to qualify for insured financing. Similarly, government-sponsored enterprises (GSEs) like Freddie Mac and Fannie Mae do not purchase ITIN mortgages; they only buy mortgages for lawful permanent residents or non-permanent residents.

ITIN holders are a population of particular economic interest, as they make up a significant percentage of the U.S. workforce and contribute to U.S. tax revenue. Filing taxes creates a financial and employment record, which can be advantageous if the immigrant seeks legal status in the future. Evidence of tax compliance demonstrates good moral character and a contribution to society, and it can also serve as proof of income when applying for loans.

A seminal paper by Munnell et al. (1996) [29] investigated discrimination against African Americans in Boston and found that they faced higher loan denial rates, even after controlling for creditworthiness, indicating evidence of discriminatory practices. Discrimination in the mortgage market based on race has remained a persistent focus in the literature (Ladd, 1998 [30]; LaCour-Little, 1999 [31]).

Barriers faced by immigrants in the mortgage market have also been documented in Italy and Spain. Bertocchi et al. (2023) [32] examine the financial behaviors of immigrants in Italy, showing that there is financial discrimination against immigrants, as they are

less likely to have housing and mortgages, increasing their likelihood of financial fragility. Mistrulli, Udin, and Zazzaro (2023) [33] find that immigrants pay a higher interest rate than native Italians and have a 2.7% lower likelihood of being approved for a mortgage. Díaz et al. (2014) [34] found that after immigrants to Spain faced higher interest rates, controlling for creditworthiness and other factors.

Studies have also explored the disadvantages faced by individuals with limited English proficiency in the mortgage market, a particularly significant issue for Hispanic households and even more for unauthorized immigrants (Liu, 2023) [35]. Liu employs machine learning to predict Limited English Proficiency (LEP) status using HMDA data.

In this paper, we investigate whether non-U.S. citizens face greater difficulties than their citizen counterparts in the U.S. mortgage market. We use machine learning models to identify non-U.S. citizens who applied for mortgages and measure the effect of being a non-U.S. citizen on approval probability. Our findings show that being a non-U.S. citizen decreases the probability of mortgage approval for Hispanic households, while for Non-Hispanic households the effect is significantly lower or even positive for Asian households.

We then classify immigrants as either authorized or unauthorized. Among Hispanic households, we find that both groups—authorized and unauthorized—have lower approval probabilities compared to U.S. citizens, with the effect being more pronounced for unauthorized immigrants. In contrast, among Asian households, non-U.S. citizens have a higher probability of approval, but this effect reverses for those classified as unauthorized, who face a lower likelihood of approval.

Finally, we show that non-U.S. citizens are not riskier borrowers: they do not

exhibit lower credit scores nor higher rates of delinquency or default compared to U.S. citizens.

This paper is structured as follows: Section 2 describes the datasets used. Section 3 explains the machine learning models applied, presents performance statistics, and justify our choice of XGBoost as the preferred algorithm. Section 4 presents aggregate results highlighting differences in income and approval rates between non-U.S. citizens and U.S. citizens. In Section 5, we provide our micro-level empirical analysis, estimating a linear probability model for mortgage approval and showing the negative effect of being a non-U.S. citizen. We also explore the distinction between authorized and unauthorized immigrants to assess whether approval likelihood varies between the two groups. Section 6 shows that non-U.S. citizens are not riskier than their U.S. citizen counterparts. Finally, Section 7 concludes.

2.2 Data

We use 3 datasets:

1. The Home Mortgage Disclosure Act (HMDA) dataset is a publicly available resource on mortgage applications in the United States. It provides information on loan applications, approvals, and denials, as well as details about the loans (e.g., loan amount and type) and borrower characteristics (e.g., race, ethnicity, gender, and income). However, this dataset does not include information on the immigration status of applicants, making it necessary to use a supplementary dataset for training our machine learning model. We use HMDA data from 2005 to 2022.

2. The 1-year American Community Survey (ACS) is an annual survey conducted by the U.S. Census Bureau that provides detailed demographic, social, economic, and housing information about the U.S. population. This survey includes a question about citizenship status, enabling us to train a model to classify observations as either U.S. citizens or non-U.S. citizens as a function of other household-level information. We use the Public Use Microdata Sample of the ACS data from 2005 to 2022.
3. Fannie Mae and Freddie Mac performance data, include quarterly information on mortgages purchased by these government-sponsored enterprises. The dataset provides performance details over the life of each mortgage, including whether it was fully repaid, became delinquent, was foreclosed, or was short sold. We use data covering the period from 2005 to 2022.

To ensure that the ACS training data for the machine learning model is representative of people who apply for mortgages in a given year, we restrict the training data to individuals in the ACS who meet the following conditions:

- Heads of household who are at least 18 years old
- Reside in the contiguous U.S. (i.e. excluding Puerto Rico, Hawaii, and Alaska)
- Own their residence with a mortgage or loan
- Moved into their current residence in the last 12 months.

It is also important to point out that, with these filters, we limit our training sample to applicants who successfully obtained mortgage approval. Therefore, we are omitting individuals who applied for a mortgage but were unsuccessful and ended up renting. Unfortunately, the ACS datasets do not provide a way to identify such individuals.¹

¹In the appendix, we do not restrict the sample to individuals with a mortgage. Instead, we address

The ACS data also provides sampling weights for each observation, which we use to expand the sample for training in order to make it more representative of the US population. After applying these weights and filters, the resulting effective number of observations in our sample is shown in table 2.1.

Year	Hispanic (Before applying weights)	Non-Hispanic	Hispanic (After applying weights)	Non-Hispanic	Total
2005	4,911	44,678	634,876	4,606,735	5,241,611
2006	4,848	41,644	590,520	4,188,279	4,778,799
2007	4,162	37,723	489,290	3,753,123	4,242,413
2008	3,193	32,870	356,987	3,240,438	3,597,425
2009	2,846	27,894	316,078	2,681,438	2,997,516
2010	2,666	25,880	307,598	2,573,394	2,880,992
2011	2,222	21,557	276,701	2,261,719	2,538,420
2012	2,377	22,928	280,766	2,320,053	2,600,819
2013	2,440	25,591	290,077	2,516,725	2,806,802
2014	2,652	27,227	310,217	2,646,276	2,956,493
2015	2,887	29,308	336,743	2,887,518	3,224,261
2016	3,297	32,285	379,811	3,153,918	3,533,729
2017	3,759	34,455	460,318	3,380,510	3,840,828
2018	3,715	34,938	440,582	3,405,406	3,845,988
2019	3,795	35,151	446,914	3,384,036	3,830,950
2020	3,135	26,808	517,122	3,297,447	3,814,569
2021	4,686	37,448	601,663	3,915,179	4,516,842
2022	4,728	37,217	577,651	3,754,453	4,332,104

Table 2.1: Effective Number of observations in the training dataset.

Figure 2.1 compares average incomes in the ACS for Hispanic and non-Hispanic households who recently purchased homes with a mortgage, broken down by citizenship. We observe that non-U.S. citizens in Hispanic households have significantly lower incomes compared to their U.S. citizen counterparts. Conversely, among non-Hispanic heads of households, non-U.S. citizens have significantly higher incomes than their U.S. citizen counterparts.

the selection issue by applying kernel density estimation for sampling.

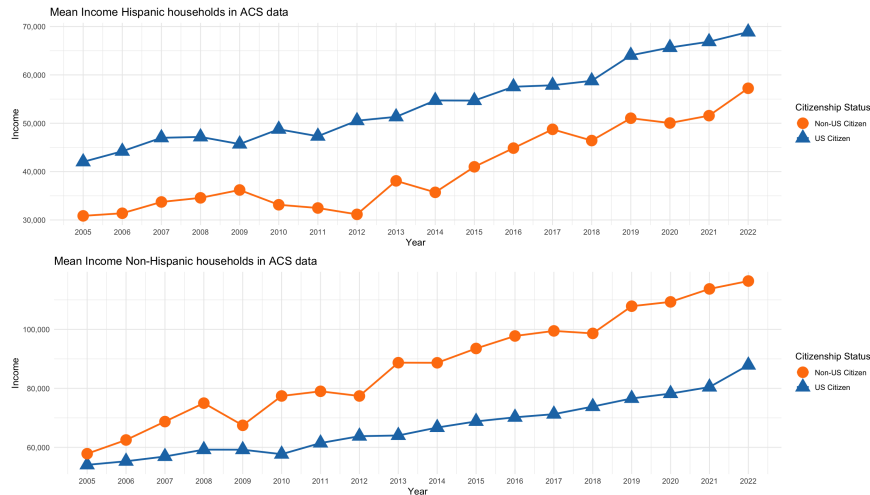


Figure 2.1: Income differences between Hispanic and non-Hispanic households, based on immigration status, in the ACS dataset after applying filters to identify mortgage market participants.

counterparts.

Given the differences observed between Hispanic and non-Hispanic immigrants in the mortgage market, we split our training dataset into Hispanic and non-Hispanic heads of household and train two separate classifiers.

In our training set, the ACS includes multiple heads of households who meet the sample conditions but have extremely low income. Figure 2.2 illustrates the income densities of our ACS sample and the HMDA data for 2007, truncated at \$150,000. The left panels show the raw distributions, with the top panel representing non-Hispanic households and the bottom panel focusing solely on Hispanic households.

The HMDA data shows a higher concentration of income at higher levels compared to the ACS distribution. This difference may result from the fact that the ACS data combines mortgages and home equity loans, while the HMDA dataset focuses specifically on purchase mortgages. Finally, the ACS is a survey and may include misreported income or other measurement error.

It is noticeable that, in the ACS distribution, a significant portion of the income mass is concentrated around 0, which is nonexistent in the HMDA dataset. Therefore, to make the training sample more representative of the HMDA data, we filter out observations from our training set with incomes below the 1st percentile in the HMDA dataset. These percentiles for each year are shown in Table 2.2. For example, in 2007, the 1st percentile was \$19,000 for Hispanic households and \$18,000 for non-Hispanic households. In the right panels of Figure 2.2, we filter out all observations with incomes below these percentiles and re-plot the distributions, which now look more similar comparing the ACS and HMDA data.

Table 2.2: Income Statistics for Hispanic and Non-Hispanic Households in the HMDA Data set (values in 000s)

Year	Hispanic Households		Non-Hispanic Households	
	Mean Income	1st Percentile	Mean Income	1st Percentile
2005	90.06	19	103.87	18
2006	101.98	21	111.89	18
2007	101.90	19	116.66	18
2008	94.83	16	120.53	17
2009	89.05	14	117.15	16
2010	88.15	15	120.45	16
2011	92.21	15	120.78	16
2012	93.25	16	121.87	17
2013	95.04	16	122.74	18
2014	96.48	17	125.01	18
2015	97.53	17	126.66	18
2016	97.06	18	126.92	19
2017	98.72	19	130.56	19
2018	93.91	19	122.47	20
2019	95.76	20	124.40	21
2020	98.94	21	126.45	21
2021	106.61	22	133.29	22
2022	119.88	20	145.31	23

Income distribution in 2007

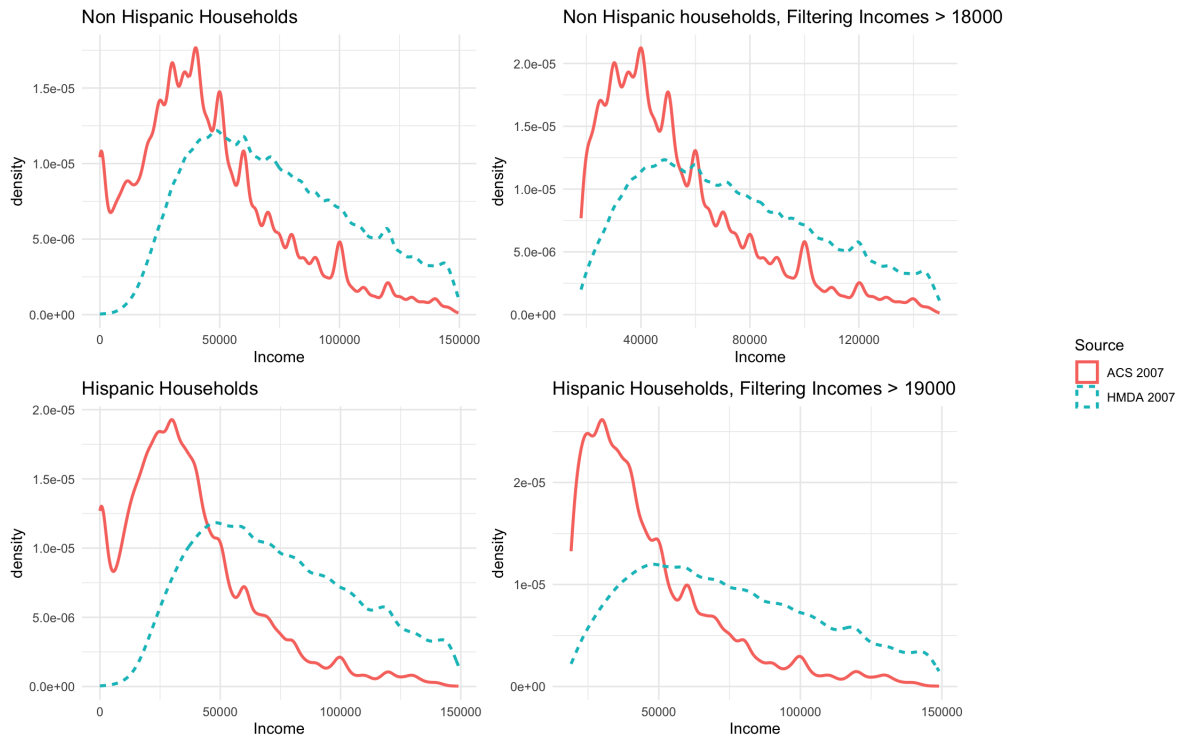


Figure 2.2: Income Densities for ACS and HMDA in 2007. The left panels show the distributions in our raw datasets, while the right panels shows the distributions after filtering out data below the 1st percentile of the HMDA income distribution

2.3 Machine Learning Model

We proceed in two steps: we first train numerous machine learning models to predict whether an individual is a non-US citizen in the ACS data and compare performance metrics to choose the best model. We then apply the trained models to the HMDA dataset to generate a prediction about whether an individual is a citizen or not.

We train classifiers using the ACS data to classify observations either as non-US citizens or as US citizens based on the self-reported citizenship status in the survey. We use the following variables as predictors for citizenship status: Income, Age, State, Race, and Gender. We also incorporate information about each observation's PUMA (Public

Use Microdata Area). For each PUMA, we calculate the percentage of Hispanic people and the percentage of non-U.S. citizens, and we add these variables as predictors to the model. We limit our predictors to these variables because they also appear in the HMDA dataset and can be used to predict citizenship status after training our model.

We use three different algorithms to train our classifiers: a traditional logit model and two commonly used machine learning algorithms: a neural network model using the package developed by Venables and Ripley, 2002 [36]) and the XGBoost algorithm (developed by Chen and Guestrin, 2016 [37]). Both are nonlinear supervised models that can capture complex relationships in the data.

Supervised learning refers to a type of machine learning in which the model is trained on a labeled dataset; in other words, each input is paired with the correct output. The model learns by finding patterns in the input data that correspond to the categories, allowing it to make predictions on new, unseen data based on the relationships it has identified.

Neural networks are a class of machine learning models that consist of one or many layers of interconnected "neurons" or nodes (in this exercise, we use a simple neural network with one hidden layer). Each node processes input data through a weighted summation and applies an activation function to introduce non-linearity. The model "learns" by adjusting the weights during training to minimize the error between its predictions and the actual target values. This weight adjustment aims to minimize the prediction error, and it is performed using a process called "backpropagation", in which the model calculates the gradient of the error with respect to each weight and updates the weights to reduce the error iteratively. Neural networks are particularly powerful in

finding non-linear patterns in the data.

Unlike neural networks, which rely on layers of interconnected nodes and learn through backpropagation, XGBoost takes a tree-based approach. XGBoost stands for “Extreme Gradient Boosting” and is widely used for supervised learning tasks. It builds a group of decision trees sequentially, where each tree attempts to correct the errors made by the previous ones. By iteratively improving the model, XGBoost achieves high predictive accuracy. It also incorporates regularization techniques to prevent overfitting.

To avoid data leakage, in which information that should be unknown to the model is inadvertently included in the feature set (such as future data not available at the time of prediction) and to account for changes in the mortgage market, we train 34 separate models (17 years x 2 ethnicities) using ACS data from 2005 to 2022. Each year’s model is applied to the corresponding year’s HMDA dataset, with separate models created for Hispanic and non-Hispanic households.

During training for each model, we split the data randomly, with 70% allocated to the training set and the remaining 30% to the test set. Since non-U.S. citizens constitute a small fraction of the sample, creating an imbalanced training problem, we also apply upsampling to improve predictive performance. We achieved this by sampling non-U.S. citizens multiple times in our training data until they accounted for 50% of the sample.

For the neural network and XGBoost models, we apply 5-fold cross-validation to select some hyperparameters.² We divide the dataset into five equal parts. In each round of validation, the model is trained on four of these folds and tested on the remaining fold. This process is repeated five times, with each fold serving as the test set exactly

²The logit model does not have any hyperparameters to select.

once. The results from each round are then averaged to provide a more robust estimate of the model’s accuracy and to choose an optimal hyperparameter.

With this method, we select the optimal maximum tree depth for the XGBoost model, considering potential values of 2, 5, and 6, with a tree depth of 6 ultimately chosen. This hyperparameter controls the maximum number of levels each decision tree can have. This parameter determines the model’s capacity to capture complexity in the data. A deeper tree can model more intricate relationships and interactions between features by allowing more splits of the feature space. However, limiting the depth helps prevent the tree from becoming too specific to the training data, therefore reducing the risk of overfitting.

The other parameters we use for XGBoost are:

Parameter	Value	Description
nrounds	100	Number of boosting rounds
eta	0.3	Learning rate
gamma	0	Minimum loss reduction
colsample_bytree	0.8	Subsample ratio of columns for each tree
min_child_weight	1	Minimum sum of instance weight needed in a child
subsample	0.8	Subsample ratio of the training instances

Table 2.3: Parameter values for the XGBoost model

We also use cross-validation to select two hyperparameters for the neural networks model: size (the number of neurons in the hidden layer) and decay (a regularization parameter to prevent overfitting, where higher values simplify the model). Using 5-fold cross-validation, we evaluate two options for the size parameter (5 and 10 neurons), ultimately selecting 10. For regularization, we test decay values of 0.01 and 0.1, with 0.1 yielding the best results.

After training each model for each year, separately for Hispanic and non-Hispanic households, we calculate several prediction accuracy metrics using only the test dataset, which contains data unseen by the model.

Figure 2.3 presents key performance metrics for our models trained on the ACS data for Hispanic households. Among the three algorithms evaluated, the XGBoost algorithm consistently outperforms the others, exhibiting the highest sensitivity (the model's ability to correctly identify true positive cases among all actual positives) and specificity (the model's ability to correctly identify true negative cases among all actual negatives). These results indicate that the XGBoost algorithm is better at classifying both non-US citizens and citizens than the logistic regression and neural network models.

Sensitivity with XGBoost remains consistently high, at approximately 95% or higher each year, compared to values ranging between 85% and 95% for the neural network model and around 70% for the logistic regression model. Similarly, specificity with XGBoost exceeds 95%, significantly outperforming the logistic regression model, which achieves specificity values of approximately 65%, and the neural network model, whose specificity ranges from 75% to 95%.

Figure 2.4 shows the sensitivity and specificity of the three models for non-Hispanic households. Once again, the XGBoost algorithm demonstrates consistently superior performance compared to the neural network and logistic regression models. While the logistic regression model achieves a sensitivity of approximately 70% and a specificity of 85%, the XGBoost model maintains a specificity at around 97% consistently across all years and a sensitivity ranging from 75% to 97%.

These metrics indicate that our models possess strong predictive power, effectively

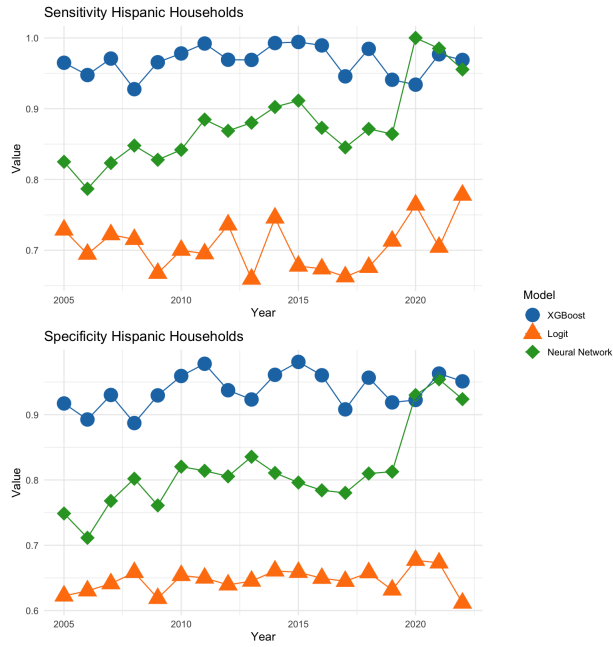


Figure 2.3: Sensitivity and Specificity on the Testing Dataset for the Model Trained with Hispanic Households Data

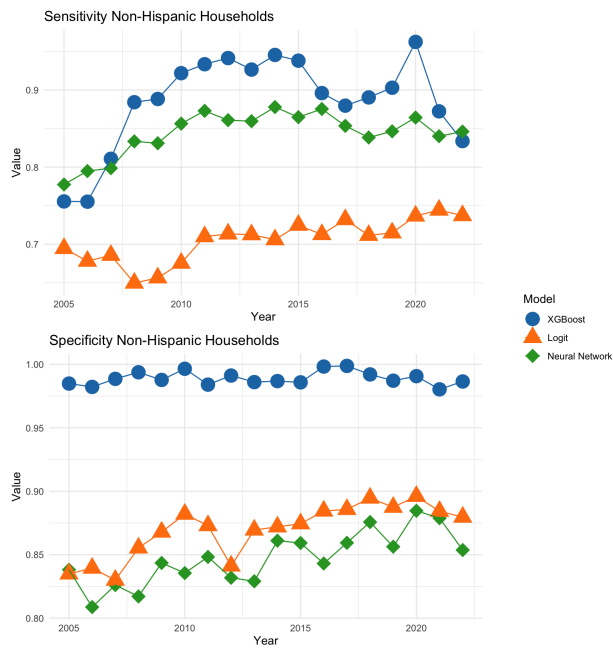


Figure 2.4: Sensitivity and Specificity on the Testing Dataset for the Model Trained with Non-Hispanic Households Data

distinguishing between non US citizens and citizens with high accuracy. The consistently high sensitivity and specificity values suggest that the models are reliable, making them

well-suited for robust and accurate predictions across different years.³

For the HMDA data, we focus on conventional mortgage loans, which exclude Federal Housing Administration (FHA) insured, Veterans Administration (VA) guaranteed and Farm Service Agency or Rural Housing Service (FSA/RHS) loans. We also exclude mortgages for home improvement or refinancing, focusing only on mortgages used for home purchases. Finally we limit our HMDA sample to owner occupied houses, in order to exclude investment properties which might have different dynamics.

HMDA provides more detailed geographical information than the ACS, including the applicant's census tract, whereas in the census data we only have access to the person's PUMA. PUMAs are made up of multiple census tracts and are mostly non-overlapping. Therefore, in HMDA, we map each census tract to its respective PUMA and add information about the percentage of Hispanics and non-U.S. citizens within that PUMA. It is also important to note that census tracts and PUMAs were redefined in 2010. For the earlier years in our dataset, we use the Missouri Census Data Center's Geocorr 2000: Geographic Correspondence Engine tool, and after 2010 we use the Census Bureau's census tract-to-PUMA relationship file.

2.3.1 Categorizing between Authorized and Unauthorized

Our first machine learning models distinguish between non-U.S. citizens and U.S. citizens. However, the non-U.S. citizen category includes both authorized and unauthorized immigrants. To differentiate between these groups, we use the method developed by

³We also calculated Shapley values for Income, Gender, the percentage of non us citizens in a PUMA and the percentage of Hispanics in a PUMA, the graph and details can be found in the appendix

Borjas (2017) [38] and later refined by Borjas and Cassidy (2019) [39] to categorize non-citizen observations in the ACS survey as authorized or unauthorized immigrants. We then retrain our models to predict whether an individual is classified as an unauthorized immigrant.

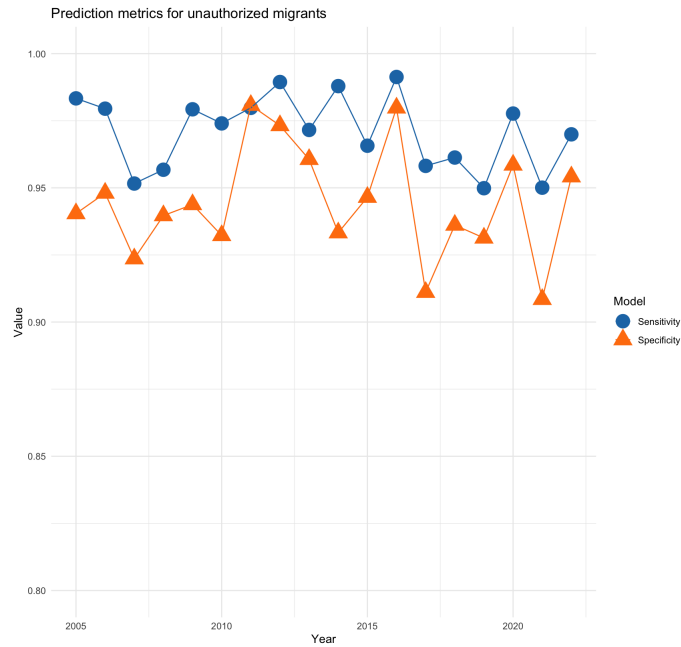


Figure 2.5: Prediction Metrics for the Machine Learning Model Predicting Unauthorized Status Among Non-U.S. Citizens

With this method, we categorize an observation as “likely authorized” in the ACS if any of the following conditions apply:

1. The person is a U.S. citizen.
2. The person arrived before 1980, as the 1986 Immigration Reform and Control Act granted amnesty to 2.7 million people.
3. The person receives Social Security or food stamps, which unauthorized immigrants cannot access.
4. The person is a veteran or is currently serving in the Armed Forces.

5. The person was born in Cuba, which grants them refugee status.
6. The person's occupation is one of the following: physician, registered nurse, air traffic controller, or lawyer, as these occupations require licensing.
7. The person works in the government sector at the local, state, or federal level.
8. To account for highly skilled immigrants likely in the US on an H1B or a similar visa, we categorized as documented those working in the tech sector, such as software developers, database administrators and architects, computer and information research scientists, among others. This method still omits legal immigrants in other sectors.

The rest of observations in our dataset are then classified as unauthorized.

After labeling each ACS observation, we follow a similar procedure as before and train an XGBoost model for each year, using only the subsample of non-U.S. citizens. These models classify non us citizens as either authorized or unauthorized. The prediction metrics for these models are presented in Figure 2.5, which shows that both sensitivity and specificity consistently remain above 95% each year.

2.4 Results in Aggregated Data

We classify each HMDA observation as a US. citizen or non-citizen using each year's ACS model to predict the immigration status of observations in the corresponding year's HMDA dataset, separately for Hispanic and non-Hispanic households (e.g., we use the model trained on 2005 ACS data to predict citizenship status in the 2005 HMDA data).

Figure 2.6 illustrates the predicted percentage of non-U.S. citizen mortgage applicants for each year in our HMDA and ACS datasets, reporting different results for Hispanic and non-Hispanic households.

According to our models, the predicted proportion of non-U.S. citizens among Hispanic households that apply for mortgages ranges from 6.5% to 21%, with the lowest value observed in 2021. Overall, our models predict a lower percentage of non-citizens among Hispanics in the HMDA data than that observed in the ACS data, particularly during the years leading up to the Great Recession, when the percentage of non-U.S. citizens among Hispanic households with recent mortgages was particularly high in the ACS.

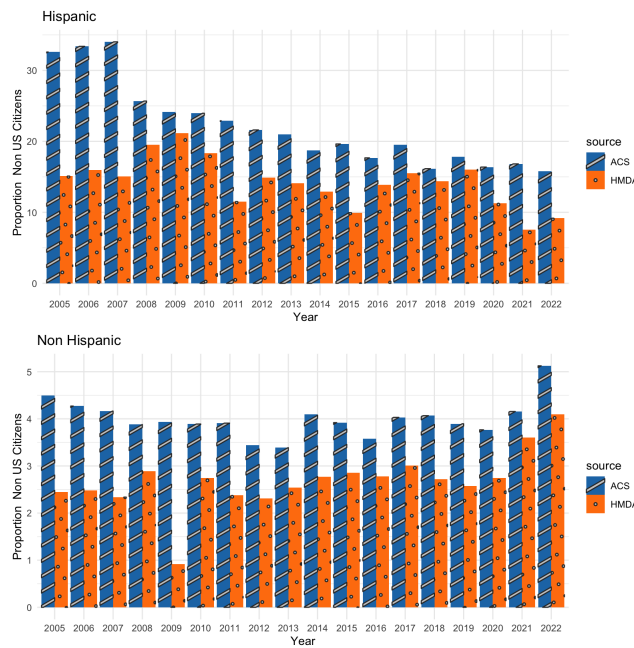


Figure 2.6: Predicted Percentage of Non-U.S. Citizens by Year in HMDA and the Percentage of Non-U.S. Citizens in Our ACS Training Data

For non-Hispanic households, the predicted percentage of non-U.S. citizens in the HMDA dataset is also lower than the percentage in our ACS data, with the predicted

percentage ranging from 1% to 5%, reaching its lowest value in 2009.

These underpredictions can be attributed to differences in the income distribution between the two datasets. From figure 2.2, the ACS data overrepresents low income households relative to the HMDA, and low income households are more likely to be non-citizens in the ACS data.

It is also important to note that for Hispanic households, the proportion of non-US citizens in the ACS data steadily declined over time, from 30% in 2005 to 15% in 2022. In contrast, for non-Hispanic households, the percentage of non-citizens remained relatively constant and even increased during the COVID-19 pandemic.

Figure 2.7 shows mortgage approval rates in our HMDA dataset, based on the predicted citizenship status for both Hispanic and non-Hispanic households. The approval rate is significantly higher for U.S. citizens than for non-U.S. citizens for both Hispanic and non-Hispanic households. However, the gap in approval rates between U.S. citizens and non-U.S. citizens is much more narrow for non-Hispanic households. Additionally, the overall approval rate is significantly lower for Hispanic households compared to non-Hispanic households.

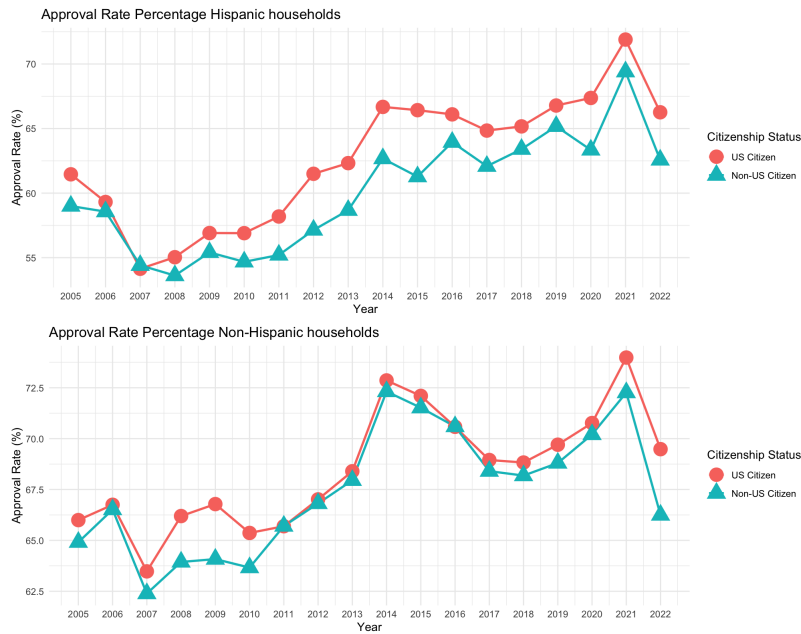


Figure 2.7: Mortgage approval rate by predicted citizenship status

Figure 2.8 shows mean income in the HMDA dataset by predicted citizenship status, with Hispanic households shown in the first panel and non-Hispanic households in the second panel. Among Hispanic households, U.S. citizens who apply for a mortgage consistently have a higher average income than non-U.S. citizens. Notably, non-U.S. citizens experience a significant increase in their mean income during 2021 and 2022, reaching a level comparable to their U.S. citizen counterparts.⁴ This trend suggests that, during the COVID-19 crisis, Hispanic non-U.S. citizens applying for mortgages were predominantly from higher-income households. On the other hand, among non-Hispanic households, non-U.S. citizens who apply for a mortgage generally have a higher average income than U.S. citizens, with the exception of 2009 and 2010, when both demographics had similar incomes.

⁴This could be a selection effect, as lower income Hispanic non-citizens were disproportionately affected by the COVID pandemic. Krannich and Massey (2024) [40] show that during COVID, immigration in the US significantly decreased and immigrants disproportionately experienced higher rates of unemployment and greater losses of income, and thus were probably not in a good position to get a mortgage

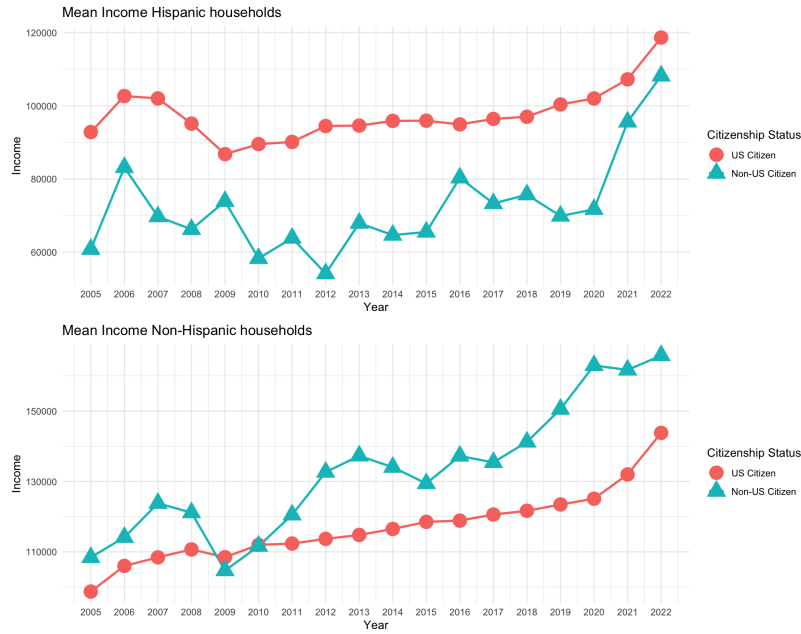


Figure 2.8: Mean Applicant income by predicted citizenship status

2.5 Micro-Level Regression Results

We employ a linear probability model on HMDA data pooled across years to estimate the effect of being a non-U.S. citizen on the likelihood of a person being approved for a mortgage application. We fit the following equation:

$$\begin{aligned}
 y_{i,t} = & \beta_0 + \beta_1 \cdot \text{non_us_citizen}_{i,t} + \beta_2 \cdot \text{female}_{i,t} + \beta_3 \cdot \text{loan_amount}_{i,t} \\
 & + \beta_4 \cdot \text{income}_{i,t} + \beta_5 \cdot \text{race}_{i,t} + \beta_6 \text{race}_{i,t} \cdot \text{non_us_citizen}_{i,t} + X + \epsilon
 \end{aligned}
 \tag{2.1}$$

Where $y_{i,t}$ represents our variable of interest, namely mortgage approval. The variable *non_us_citizen* is a binary indicator derived from our machine learning model's prediction, coded as 1 if the applicant is classified as a not a U.S. citizen and 0 otherwise.

The variable *female* is a dummy variable representing the applicant's sex and *loan_amount* represents the requested loan amount in thousands of dollars.

We also control for the applicant's income and indicators for Asian, Hispanic and Black households, even though these variables are also used in our machine learning model. Since our machine learning models are highly non-linear, multicollinearity is not a significant concern, and we test for it by measuring the Variance Inflation Factor (VIF). Finally, X denotes fixed effects. We run two different regressions; one includes census tract and year fixed effects, and the other adds bank fixed effects. We also implement two-way clustered standard errors by bank and census tract.

In the HMDA dataset, we omit loans coded as 'action taken: purchased by the institution', as these are loans bought by one financial institution from another. Then we divide our remaining dataset into two categories: mortgages completed (those that were either denied or approved) and applications that were withdrawn or closed due to being incomplete.

In our first regression, the dependent variable, is a dummy variable that takes the value 1 if the loan application was approved and 0 otherwise. We limit our sample to completed applications. The regression results are presented in Table 2.4, with the first column includes census tract, year and bank fixed effects and the second column excludes bank fixed effects.

We find a negative and statistically significant effect of being a non-U.S. citizen on the probability of mortgage approval, When we include census tract, year, and bank fixed effects, the estimate indicates that being a non-U.S. citizen decreases the probability of approval by 0.15 percentage points. As a reference, the average approval rate in our

Table 2.4: Estimates of the Effect of Non-U.S. Citizenship on Approval Rates

Dependent Variable: Model:	Indicator for mortgage approval	
	(1)	(2)
<i>Variables</i>		
Female	-0.8717*** (0.0948)	-1.578*** (0.0158)
Income (000s)	0.0159*** (0.0010)	0.0171*** (0.0001)
Loan amount (000s)	-0.0014** (0.0006)	3.97×10^{-5} (6.19×10^{-5})
Asian	-2.383*** (0.0920)	-2.520*** (0.0335)
Black or African American	-8.557*** (0.3162)	-12.63*** (0.0592)
Hispanic	-4.003*** (0.2609)	-6.020*** (0.0359)
Non-US Citizen	-0.1551** (0.0768)	-0.1901*** (0.0680)
Asian \times Non-US Citizen	0.6794*** (0.1141)	0.5334*** (0.0840)
Black or African American \times Non-US Citizen	-0.5736** (0.2316)	0.0766 (0.2187)
Hispanic \times Non-US Citizen	-2.087*** (0.2277)	-2.489*** (0.0998)
<i>Fixed-effects</i>		
Census Tract	Yes	Yes
Bank	Yes	
Year	Yes	Yes
<i>Fit statistics</i>		
Observations	48,601,527	48,601,527
R ²	0.17952	0.07883
Within R ²	0.00593	0.01071

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

data set is 85.57%.

We also find a statistically significant negative effect of being female, which decreases the likelihood of approval by 0.87 percentage points. Income, on the other hand, has a positive effect on the probability of approval.

Additionally, non-whites experience a negative effect on the probability of loan approval compared to White applicants. Hispanic households face a decrease in the probability of approval of 4 percentage points compared to their White counterparts, while being a Hispanic non US citizen decreases the probability of an extra 2.087 percentage points. The largest negative effect is observed for Black or African American applicants, whose probability of approval decreases by 8.55 percentage points relative to the control group of White applicants, with an additional decrease of 0.57 percentage points if they are non-U.S. citizens. For Asian applicants, the probability of approval decreases by 2.38 percentage points; however, being a non-U.S. citizen has a positive effect of 0.67 percentage points.

These results suggest that being a non-U.S. citizen significantly decreases the likelihood of mortgage approval for Hispanic households but has a much smaller negative effect on the approval rate for non-Hispanic households, potentially reflecting systemic biases or institutional barriers that disproportionately affect non-citizen Hispanic applicants in the mortgage market.

Since we used race and income as predictors of citizenship status, there is a possibility of collinearity issues. While the XGBoost model is highly nonlinear, we calculate the Variance Inflation Factor (VIF) to rule out this concern.⁵ The results, shown in Table 2.5,

⁵The Variance Inflation Factor (VIF) is a measure of multicollinearity in a regression model. The square

Term	VIF	VIF 95% CI
Non US-citizen	1.10	[1.10, 1.10]
Female	1.04	[1.04, 1.04]
Income (000s)	1.42	[1.42, 1.42]
Loan Amount (000s)	1.41	[1.41, 1.42]
Race	1.04	[1.04, 1.04]

Table 2.5: Collinearity Diagnostics in the Approval Probability Model for Non-U.S. Citizenship

indicate that the VIF values for all variables are relatively low, with all values significantly below 5, thus ruling out concerns of high collinearity.

We next run the same linear probability model, but this time using *withdrawal*, a dummy variable equal to 1 if the application was withdrawn or incomplete and 0 if it was completed as the dependent variable. Mortgage applicants may withdraw their applications for various reasons—they may have decided not to proceed with the purchase, experienced changes in their financial situation, or sought better deals by approaching multiple lenders to secure the best interest rate while not proceeding with other applications. Applications can also be incomplete due to a lack of support from the financial institution in guiding the applicant, which can be particularly significant for immigrants with limited knowledge of the U.S. financial system or limited English proficiency.

The results of this regression are presented in Table B.1 in the Appendix. For Hispanic households, the probability of withdrawal increases by 0.43 percentage points but being a Hispanic non U.S. citizen does not have a statistically significant effect.

root of the VIF indicates how much larger the standard error of a coefficient is compared to a scenario in which the variable has no collinearity with other predictors. The smallest possible VIF value is 1, which indicates no collinearity. As a general rule of thumb, a VIF greater than 5 or 10 suggests a problematic level of multicollinearity (Gareth et. al. 2017 [41])

We also find that Asian households have a 1.6 percentage point higher probability of withdrawing compared to their white counterparts, and this probability increases by an additional 0.37 percentage points for non-U.S. citizens.

Finally, we examine the impact of citizenship status on interest rates for approved loans. Table B.3 in the appendix presents the results of this regression. We do not find a significant effect of non-U.S. citizenship on interest rates. However, we do observe a small but statistically significant effect for Asian households, whose interest rates are 0.089 percentage points lower on average, with an additional decrease of 0.03 percentage points for non-U.S. citizen Asians. For Hispanic households, interest rates are on average 0.03 percentage points higher, with an additional 0.02 percentage points for non-U.S. citizens; both effects are statistically significant. These results suggest that Asian households may spend more time searching for loans than White households, while Hispanic households might withdraw for adverse reasons, as they end up paying a slightly higher interest rate.

2.5.1 Alternative Specification using Propensity Score Matching

Since non-U.S. citizens represent the minority group in our dataset, as a robustness check we use propensity score matching to compare non-citizens with their most similar U.S. citizen counterparts based on observable characteristics — specifically, income, loan amount, year of origination and gender. We implement nearest neighbor matching, which pairs each non-U.S. citizen with the closest U.S. citizen in terms of estimated propensity score (i.e., the probability of being a non-U.S. citizen given their characteristics).

This process creates a balanced dataset where both groups are comparable. Using this matched sample, we then estimate a logistic regression model to analyze whether non-U.S. citizenship is associated with lower approval probabilities, and compute the average marginal effect (AME) to interpret the impact in terms of changes in approval probability. The results are shown in Table 2.6 for Hispanic and non-Hispanic households. We find that, for Hispanic households, being a non-U.S. citizen decreases the probability of approval by 1.64 percentage points. For non-Hispanic households, this probability declines by 0.8 percentage points. Both effects are statistically significant.

Table 2.6: Average Marginal Effect of Non-US Citizenship on Mortgage Approval

Group	AME	SE	z	p-value	95% CI Lower	95% CI Upper
Hispanic Households	-0.0164	0.0007	-23.84	0.0000	-0.0178	-0.0151
Non-Hispanic Households	-0.0086	0.0004	-20.27	0.0000	-0.0095	-0.0078

2.5.2 Mortgages for unauthorized immigrants

Next we apply our previously trained unauthorized categorizer to the HMDA dataset, specifically to the observations classified by our previous model as non-U.S. citizens, to predict whether a mortgage application originates from an unauthorized immigrant. The percentages of predicted unauthorized immigrants for Hispanic and non-Hispanic households are shown in figure 2.9.

Finally, we re-estimate the previous linear probability model for mortgage approval in the HMDA dataset, but now we include an extra dummy indicating if the person is a likely unauthorized non-US citizen in the regression. The results are presented in Table 2.7.

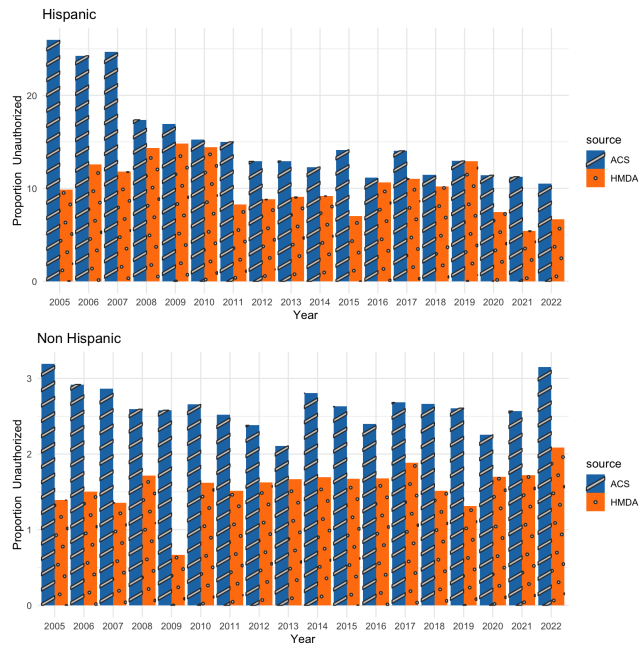


Figure 2.9: Predicted percentage of unauthorized immigrants among all observations by year in HMDA, compared to the percentage of unauthorized immigrants in our ACS training data.

For authorized non-citizens, there is not a statistically significant effect on approval. However, being an unauthorized non-U.S. citizen has a negative and statistically significant effect, reducing the probability of approval by 0.46 percentage points when census tract, bank, and year fixed effects are included.

As before, we find that being any race other than White is associated with a lower probability of approval. For Hispanic households, the probability of approval decreases by 4 percentage points, being a non us citizen decreases the probability by an extra 1.73 percentage points and if they are unauthorized by an extra 0.43 percentage points, although this effect is only significant at the 10% confidence level.

The largest negative effect is observed for Black or African American applicants, whose approval probability decreases by 8.5 percentage points. For Asian applicants, the probability of approval decreases by 2.38 percentage points, but non-U.S. citizenship

is associated with a 0.80 percentage point increase, with both effects being statistically significant.

Since these regressions include a new generated variable that uses income and race as a predictor, we calculate the VIF for both Hispanic and non-Hispanic samples. The results are shown in Table 2.8. Every variable has a reasonably low VIF, with the non-US citizen indicator having the highest value at 2.76. This suggests that multicollinearity is not a major problem.

Next, we run a regression where the dependent variable is a withdrawal indicator. The results are presented in Table B.2 in the appendix. For Hispanic households, we do not find a statistically significant effect of being a non-U.S. citizen but being an unauthorized immigrant has a statistically significant effect decreasing the probability of withdrawal by 0.38 percentage points. For Asian households, being non-US citizen increases the probability of withdrawal by 0.71 percentage points while being unauthorized decreases the probability by 0.54 percentage points. Finally, the results for the interest rate regression are presented in Table B.4 in the appendix. As before, Asian non-citizens have a statistically significant negative effect decreasing the interest rates by 0.03 percentage points and being unauthorized does not have a significant extra effect. Hispanic non-US citizens have on average higher interest rates by 0.04 percentage points, while being unauthorized decreases it by 0.03 percentage points.

Table 2.7: Estimates of the Effect of being Non Us Citizen and unauthorized on Mortgage Approval

Dependent Variable:	Indicator for mortgage approval	
<i>Variables</i>		
Female	-0.8723*** (0.0949)	-1.579*** (0.0158)
Income (000s)	0.0159*** (0.0010)	0.0171*** (0.0001)
Loan amount (000s)	-0.0014** (0.0006)	3.89×10^{-5} (6.16×10^{-5})
Asian	-2.384*** (0.0920)	-2.520*** (0.0335)
Black or African American	-8.557*** (0.3162)	-12.63*** (0.0592)
Hispanic	-4.004*** (0.2610)	-6.021*** (0.0359)
Non-US Citizen	0.1682 (0.1221)	0.3065** (0.1233)
Unauthorized	-0.4605*** (0.1440)	-0.7071*** (0.1462)
Asian \times Non-US Citizen	0.8017*** (0.1598)	0.4366*** (0.1392)
Black or African American \times Non-US Citizen	0.1056 (0.3499)	0.8107** (0.3533)
Hispanic \times Non-US Citizen	-1.734*** (0.2462)	-2.276*** (0.1708)
Asian \times Unauthorized	-0.3788** (0.1710)	-0.0465 (0.1661)
Black or African American \times Unauthorized	-1.071** (0.4528)	-1.173*** (0.4375)
Hispanic \times Unauthorized	-0.4712* (0.2439)	-0.2710 (0.1987)
<i>Fixed-effects</i>		
Census tract	Yes	Yes
Bank	Yes	
Year	Yes	Yes
<i>Fit statistics</i>		
Observations	48,601,527	48,601,527
R ²	0.17953	0.07884
Within R ²	0.00593	0.01072

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table 2.8: Collinearity Diagnostics

Term	VIF	VIF 95% CI
Non-Us Citizen	2.76	[2.76, 2.77]
Unauthorized	2.63	[2.63, 2.63]
female	1.04	[1.04, 1.04]
income (000s)	1.37	[1.37, 1.37]
Loan amount (000s)	1.36	[1.36, 1.36]
race	1.22	[1.22, 1.22]

2.6 Default Risk

To assess whether individuals categorized as likely non-U.S. citizens face higher denial rates due to higher true credit risk, we analyze their credit score and the subsequent mortgage performance. We use performance data from Fannie Mae and Freddie Mac, which provide monthly performance records throughout the life of the mortgage for loans acquired from 2005 to 2022. We merge this data with our HMDA dataset and analyze the effect of being a likely non-U.S. citizen on credit score and measures of delinquency and default.⁶

Frame et al. (2023) [42] merge HMDA data from 2018 and 2019 with performance data from Black Knight McDash, using origination date, loan type, loan purpose, loan amount, ZIP code, lien type, and occupancy type as matching variables. Similarly, we use these variables to merge our datasets, additionally incorporating bank name as a matching criterion. Rosen (2011) [43] also does a similar merge between HMDA and the Lender Processing Services (LPS) Applied Analytics (formerly known as McDash Analytics), matching 18.4% of his HMDA dataset.

⁶The public HMDA dataset does not include credit score information

We match our HMDA dataset with our default data using the property’s ZIP code, mortgage type (in this case, owner-occupied), whether the mortgage was sold to Fannie Mae or Freddie Mac, the bank name, origination year, metropolitan statistical area (MSA), and loan amount, retaining only unique matches.⁷ Our matching rate for Fannie Mae mortgages is 27.14% and for Freddie mac 40.46%. Our resulting dataset includes 1,131,586 mortgages originated from 2005 to 2022.

Table 2.9: Summary statistics by ethnicity and predicted citizenship status

Ethnicity	Predicted Citizenship Status	<i>Mean</i>				
		Income	Credit Score	Delinquent 30	Delinquent 90	Defaulted
Hispanic	US Citizen	84.4	735	23.2	11.1	3.79
Hispanic	Non US Citizen	59.4	730	26.6	13.3	4.81
Non Hispanic	US Citizen	91.0	749	14.1	5.6	2.01
Non Hispanic	Non US Citizen	104.0	748	13.9	6.01	2.04

Table 2.9 presents summary statistics of income, credit score, delinquency rates, and default rates. Among Hispanic households, Non–US Citizens have, on average, lower income, a slightly lower credit score, and slightly higher delinquency and default rates. Among Non-Hispanic households, Non–US Citizens show a higher average income, similar credit scores, and similar delinquency and default rates.

Figure 2.10 presents different measures of mortgage delinquency and default over time by citizenship status, while Figure 2.11 shows the same measures by Race. The first row displays the percentage of mortgages that have been delinquent for at least 30 days, the second row for at least 90 days, and the third row shows the percentage of mortgages that ended in foreclosure or short sale. Overall, non–U.S. citizens and citizens exhibit similar delinquency and default behavior. Additionally, Hispanic and African American borrowers have higher delinquency rates than Asian and white households,

⁷More details about the matching procedure are explained in the appendix.

with the gap being more pronounced for mortgages originated between 2005 and 2008, prior to the financial crisis.

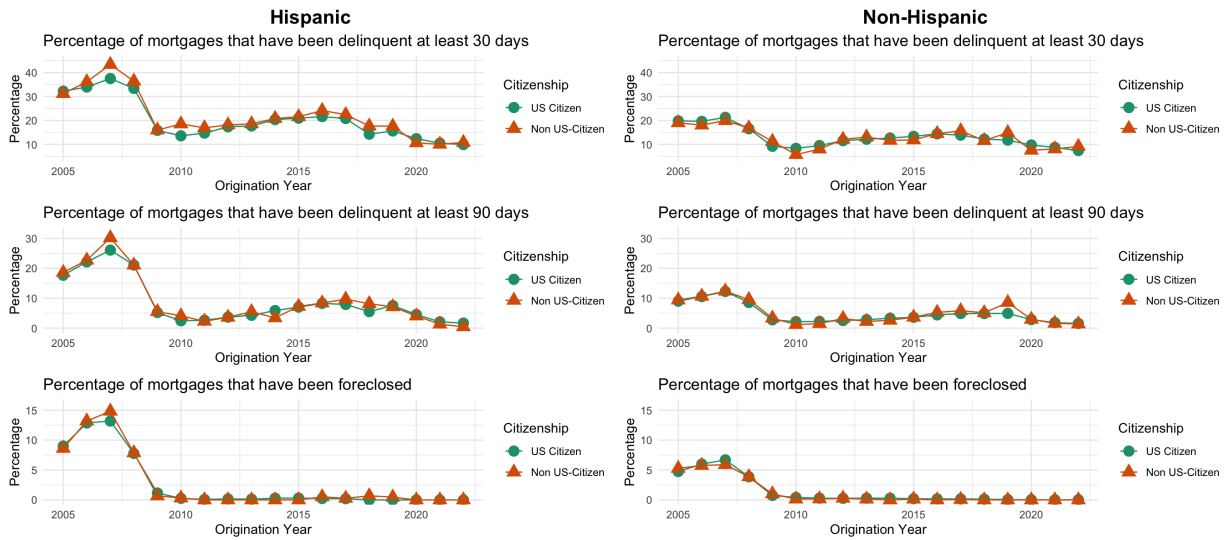


Figure 2.10: Delinquency and default rates by citizenship status.

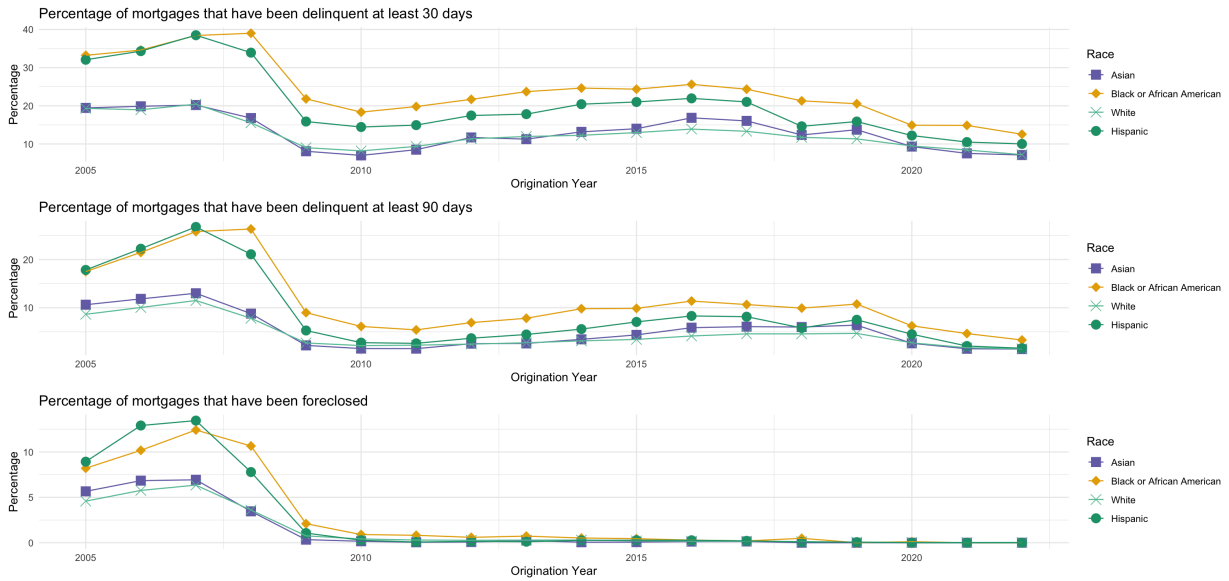


Figure 2.11: Delinquency and default rates by race.

To explore if non-U.S. citizens are riskier, we check if non US-Citizens have a significant lower credit score (ex ante risk) and if they are more likely to become delinquent or default (ex post risk), to do so we run the following linear model, estimated using the

Table 2.10: Estimates of the Effect of Non-U.S. Citizenship on Credit Score

Dependent Variable:	Credit Score
<i>Variables</i>	
Female	-0.3398** (0.1666)
Income (000s)	0.0017 (0.0029)
Loan Amount (000s)	0.0025 (0.0020)
Asian	-3.029*** (0.3572)
Black or African American	-17.89*** (0.6283)
Hispanic	-11.49*** (0.4768)
Non US-Citizen	-0.3462 (0.6481)
Asian × Non US-Citizen	0.3400 (0.7734)
Black or African American × Non US-Citizen	-2.078 (2.315)
Hispanic × Non US-Citizen	-0.2644 (0.9539)
<i>Fixed-effects</i>	
Census Tract	Yes
Bank	Yes
Year	Yes
<i>Fit statistics</i>	
Observations	1,131,586
R ²	0.14440
Within R ²	0.00506

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 2.11: Estimates of the Effect of Non-U.S. Citizenship on default

Dependent Variables:	has_been_delinquent_30	has_been_delinquent_90	defaulted
<i>Variables</i>			
Female	1.391*** (0.1233)	0.7009*** (0.0898)	0.3074*** (0.0443)
Income (000s)	-0.0224*** (0.0011)	-0.0198*** (0.0013)	-0.0080*** (0.0011)
Loan Amount (000s)	0.0103*** (0.0014)	0.0160*** (0.0015)	0.0092*** (0.0012)
Credit Score	-0.1801*** (0.0040)	-0.0948*** (0.0042)	-0.0314*** (0.0034)
Asian	-0.1884 (0.3115)	-0.2048 (0.1402)	-0.2806*** (0.0742)
Black or African American	6.401*** (0.3180)	4.472*** (0.3169)	0.6432*** (0.1561)
Hispanic	3.725*** (0.3217)	2.339*** (0.3758)	0.6750*** (0.1725)
Non US-Citizen	0.0898 (0.4734)	-0.1824 (0.3111)	-0.0454 (0.2388)
Asian × Non US-Citizen	-0.8722 (0.5330)	0.1128 (0.3789)	0.0730 (0.2100)
Black or African American × Non US-Citizen	-0.1972 (2.560)	-0.9030 (1.679)	-2.255** (0.9244)
Hispanic × Non US-Citizen	0.3031 (0.7845)	0.2132 (0.5174)	-0.0582 (0.3508)
<i>Fixed-effects</i>			
Census Tract	Yes	Yes	Yes
Bank	Yes	Yes	Yes
Year	Yes	Yes	Yes
<i>Fit statistics</i>			
Observations	1,131,586	1,131,586	1,131,586
R ²	0.17509	0.16694	0.14082
Within R ²	0.06435	0.04240	0.01278

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

merged HMDA, Fannie Mae and Freddie Mac data, pooled across years.⁸

$$\begin{aligned} y_{it} = & \beta_0 + \beta_1 \cdot non_us_citizen_{it} + \beta_2 \cdot female_{it} + \beta_3 \cdot loan_amount_{it} + \\ & \beta_4 \cdot income_{it} + \beta_5 \cdot race_{it} + \beta_6 credit_score_{it} + \\ & \beta_7 \cdot race_{it} \cdot non_us_citizen_{it} + X + \epsilon_{it} \end{aligned} \quad (2.2)$$

Where y is our variable that indicates, credit score, delinquency or default, $non_us_citizen$ is the indicator variable for being a non-U.S. citizen, and $female$ indicates whether the borrower is a woman. $loan_amount$ represents the loan amount in thousands of dollars, while $income$ corresponds to the borrower's income, also in thousands of dollars. $race$ are indicator variables for Hispanic, Asian, African American and White. $credit_score$ represents the FICO Classic credit score, and we also include the interaction term between $race$ and $non_us_citizen$. Finally, we include time, bank, and Census tract fixed effects.

For ex ante risk, our dependent variable is the credit score. The results from the regression are shown in Table 2.10, where the coefficient for the Non-US Citizen dummy is not statistically significant. Hispanic households have on average 11.4 points lower credit scores, while being a non-U.S. citizen does not have an additional statistically significant effect. For Asian and African American households, there is no statistically significant effect of non-U.S. citizenship. This suggests that, all else equal, Non-US Citizens do not have significantly lower credit scores, and therefore are not ex ante riskier in our dataset.

For ex post risk, we run a linear probability model on dummies for delinquency

⁸Fannie Mae and Freddie Mac only buy mortgages of authorized non-U.S. citizens

and risk, the results from this regression are shown in Table 2.11. In the first column, the dependent variable is an indicator of whether the mortgage has ever been delinquent for 30 days. The second column, examines whether the mortgage has been delinquent for 90 days and the third column examines whether the mortgage ended up in foreclosure, a short sale, or was classified as a non-performing loan.

Our results suggest that being likely a non-U.S. citizen does not increase the probability of loan delinquency or default as we do not find a significant effect for either Hispanic or Asian households for any of the delinquency or default measures. This implies that their likelihood of defaulting is similar to that of their U.S. citizen counterparts. A higher credit score has a statistically significant negative effect on delinquency; each additional point reduces the probability of being 30 days delinquent by 0.18 percentage points and the probability of default by 0.03 percentage points. Similarly, higher income is associated with a lower likelihood of delinquency and default, while a higher loan amount increases these risks.

Additionally, we do find that Black or African American households have an increased probability of being 30 days delinquent by 6.4 percentage points, while their probability of being 90 days delinquent increases by 4.4 percentage points, and the probability of defaulting rises by 0.64 percentage points. However, being a non-U.S. citizen within the Black or African American group is associated with a 2.25 percentage point decrease in the probability of default.

Overall these results suggest higher mortgage denial rates for non-citizen households are not a response to lower credit scores or to higher default or delinquency rates among non-citizen borrowers.

2.6.1 Alternative Specification using propensity score matching

Similarly to the previous section, we use propensity score matching to compare non-U.S. citizens to their U.S. citizen counterparts as a robustness check. This time, we use the same covariates as before—income, loan amount, year, and gender—and additionally include credit score. This matching process creates a balanced dataset in which both groups are comparable. We then estimate the average marginal effect of being a non-U.S. citizen on three binary outcomes: having been delinquent for at least 30 days, being delinquent for at least 90 days, or having defaulted. The results are shown in Table 2.12. None of the effects of being a non-U.S. citizen are statistically significant for either Hispanic or non-Hispanic households. This reinforces our previous findings, indicating that non-U.S. citizens are not at a higher risk of becoming delinquent or defaulting.

Group	Dependant Variable	AME	SE	z	p	lower	upper
Hispanic Households	Has been delinquent 30	0.0010	0.0064	0.1630	0.8705	-0.0114	0.0135
Hispanic Households	Has been delinquent 90	0.0030	0.0023	1.3159	0.1882	-0.0015	0.0075
Hispanic Households	Defaulted	0.0045	0.0030	1.4782	0.1394	-0.0015	0.0104
Non-Hispanic Households	Has been delinquent 30	-0.0018	0.0034	-0.5374	0.5910	-0.0085	0.0048
Non-Hispanic Households	Has been delinquent 90	0.0011	0.0023	0.4561	0.6483	-0.0035	0.0056
Non-Hispanic Households	Defaulted	-0.0017	0.0014	-1.2227	0.2214	-0.0045	0.0010

Table 2.12: Average Marginal Effects (AME) on delinquency and default by citizenship status

2.6.2 Citizenship and loan risk during the financial crisis

In this section, we focus on mortgages originated prior to the financial crisis by restricting our dataset to loans issued between 2005 and 2007. Figure 2.12 shows state-level data on the percentage of mortgages that became delinquent for at least 30 days,

the share of mortgages held by likely non-U.S. citizens and Hispanic borrowers, and the change in the housing price index.⁹ Florida has the highest delinquency rate, with 23.3% of mortgages originated during this period falling into 30-day delinquency, followed by Louisiana, Nevada and Arizona. Notably, states with a high proportion of Hispanic and likely non-U.S. citizen homebuyers experienced some of the steepest price declines. For example, in California, where 18% of the buyers were Hispanic and 5% were likely non-U.S. citizens, home prices dropped by 39.24%. In Florida, which had 13% Hispanic buyers and 4% likely non-U.S. citizen buyers, prices declined by 36.09%.¹⁰

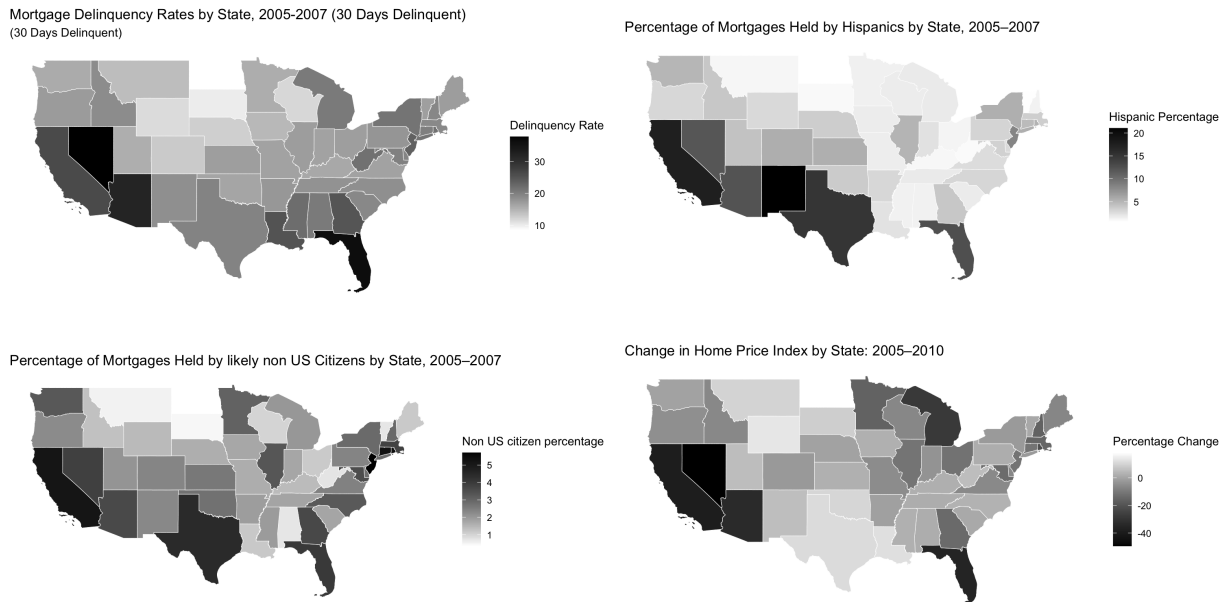


Figure 2.12: Delinquency rate, percentage of mortgages held by Hispanics and percentage of mortgages held by likely non-U.S. citizens by State for mortgages originated before the financial crisis (2005-2007)

In Table 2.13, we run the same default and delinquency regressions as before, but this time we focus only on mortgages originated between 2005 and 2007, to analyze

⁹Home prices are measured using the Federal Housing Finance Agency House Price Index.

¹⁰Percentages for all states are provided in the appendix, Table B.6.

Table 2.13: Estimates of the Effect of Non-U.S. Citizenship on default for mortgages acquired from 2005-2007

Dependent Variables:	has.been.delinquent_30	has.been.delinquent_90	defaulted
<i>Variables</i>			
Female	1.750*** (0.2380)	1.007*** (0.2002)	0.4313*** (0.1072)
Income (000s)	-0.0309*** (0.0021)	-0.0320*** (0.0020)	-0.0188*** (0.0013)
Loan Amount (000s)	0.0134*** (0.0032)	0.0220*** (0.0026)	0.0158*** (0.0017)
Credit Score	-0.2123*** (0.0049)	-0.1284*** (0.0039)	-0.0585*** (0.0016)
Asian	-1.752*** (0.3987)	-0.4047 (0.2733)	-0.4956** (0.2274)
Black or African American	5.451*** (0.4900)	4.602*** (0.5786)	0.5236 (0.3400)
Hispanic	4.108*** (0.4973)	3.505*** (0.5005)	1.306*** (0.2630)
Non US-Citizen	0.5119 (0.7476)	0.4930 (0.6572)	0.3896 (0.5242)
Asian × Non US-Citizen	-0.9503 (0.9531)	-0.2880 (0.9339)	-0.5955 (0.6987)
Black or African American × Non US-Citizen	-5.009* (2.752)	-2.447 (2.194)	-4.120** (1.860)
Hispanic × Non US-Citizen	0.4229 (1.160)	0.0556 (1.036)	-0.9373 (0.7436)
<i>Fixed-effects</i>			
Census Tract	Yes	Yes	Yes
Bank	Yes	Yes	Yes
Year	Yes	Yes	Yes
<i>Fit statistics</i>			
Observations	317,512	317,512	317,512
R ²	0.28107	0.27557	0.23463
Within R ²	0.09422	0.06232	0.02286

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

the effects of race and non-U.S. citizenship on delinquency and foreclosure during the financial crisis.

We find that being a non-U.S. citizen does not have a statistically significant effect on mortgage delinquency or default. Hispanic households have a 4.1 percentage points higher probability of being 30 days delinquent but being a Hispanic non-citizen does not have a statistically significant effect. Among non-Hispanic households, being Asian is associated with a 1.7 percentage point decrease in the probability of being 30 days delinquent, while being Black increases this probability by 5.4 percentage points with both effects being statistically significant. However, non-U.S. citizenship within these racial groups does not have a significant impact.

Once again, our findings indicate no significant effect of non-U.S. citizenship on mortgage delinquency or default, even during the financial crisis. This suggests that non-U.S. citizens were not riskier borrowers and did not face higher default rates therefore, they should not face a lower likelihood of mortgage approval.

2.7 Conclusion

In this paper, we first train a machine learning classifier using ACS data to predict whether a person is a U.S. citizen. We then use this classifier to categorize observations in the HMDA dataset to estimate whether non-U.S. citizens face greater challenges in the mortgage market. Our findings show that, for Hispanic households, being a non-U.S. citizen has a negative effect on the probability of approval; for African-American noncitizens, this effect is negative but smaller, while for Asian households, this effect is

slightly positive.

We apply the method developed by Borjas [38] [39] to classify non-citizen observations in our ACS dataset as either authorized or unauthorized immigrants. Using this classification, we train a model to distinguish between authorized and unauthorized non-U.S. citizens and apply it to our HMDA dataset. Our findings reveal that, among Hispanic households, non-U.S. citizens face a lower probability of loan approval, with an even larger reduction for those classified as unauthorized. In contrast, in Asian households, non-U.S. citizens have a higher probability of approval, although this effect turns negative when the individual is classified as unauthorized.

We then show that non-U.S. citizens are not riskier borrowers than their U.S. citizen counterparts, using Fannie Mae and Freddie Mac performance data matched to our HMDA dataset. Non-U.S. citizens do not exhibit higher probabilities of delinquency or default. This finding holds even for mortgages originated prior to the financial crisis, despite the fact that non-U.S. citizens were more active in the mortgage market in states that experienced sharp price declines, such as Florida and California.

Overall, these results suggest that Hispanic non-citizens face difficulties in the mortgage market, making it harder for them to get approved, even if they are authorized.

Appendix A: Proofs and Robustness test in Chapter 1

A.0.1 Proof of Bubbles existence

Let's first define $x_t = \frac{b_t}{sk_t^\alpha}$, $x_t^I = \frac{b_t^I}{sk_t^\alpha}$ and $x_t^U = \frac{b_t^U}{sk_t^\alpha}$ and analyze the case where $\frac{b_t + b_t^I}{(1-\epsilon)sk_t^\alpha} < 1$. from equation 1.9 we have

$$E \left[\frac{b_{t+1}}{b_t + b_t^I + b_t^U} \right] = q\alpha k_{t+1}^{\alpha-1}$$

$$E \left[\frac{b_{t+1}}{sk_{t+1}^\alpha} \right] = \frac{q\alpha k_{t+1}^{\alpha-1} (b_t + b_t^I + b_t^U)}{sk_{t+1}^\alpha}$$

$$E[x_{t+1}] = \frac{\alpha q (b_t + b_t^I + b_t^U)}{sk_{t+1}}$$

Using equation 1.13 we then get:

$$E[x_{t+1}] = \frac{\alpha q (b_t + b_t^I + b_t^U)}{s \{ [Q\epsilon + q(1-\epsilon)]sk_t^\alpha + [Q-q]b_t^I - qb_t \}}$$

If we divide the numerator and denominator by sk_t^α we get:

$$E[x_{t+1}] = \frac{\alpha q (x_t + x_t^I + x_t^U)}{s \{ [Q\epsilon + q(1-\epsilon)] + [Q-q]x_t^I - qx_t \}}$$

This equation describes how the bubble behaves when $\frac{x_t + x_t^I}{(1-\epsilon)} < 1$

Now let's analyze the other case, i.e. when $\frac{b_t + b_t^I}{(1-\epsilon)sk_t^\alpha} > 1$ from equation 1.9 we have:

$$E \left[\frac{b_{t+1}}{b_t + b_t^I + b_t^U} \right] = n_t^* Q \alpha k_{t+1}^{\alpha-1}$$

$$E \left[\frac{b_{t+1}}{sk_{t+1}^\alpha} \right] = \frac{n_t^* Q \alpha k_{t+1}^{\alpha-1} (b_t + b_t^I + b_t^U)}{sk_{t+1}^\alpha}$$

$$E[x_{t+1}] = \frac{n_t^* Q \alpha (b_t + b_t^I + b_t^U)}{sk_{t+1}}$$

Using equation 1.13 we get that when $\frac{x_t + x_t^I}{(1-\epsilon)} > 1$ the bubble evolves according to the following equation:

$$E[x_{t+1}] = \frac{n_t^* Q \alpha (b_t + b_t^I + b_t^U)}{s \{ [Q\epsilon + (1-\epsilon)] sk_t^\alpha + [Q-1] b_t^I - b_t \}}$$

$$E[x_{t+1}] = \frac{n_t^* Q \alpha (x_t + x_t^I + x_t^U)}{s \{ [Q\epsilon + 1 - \epsilon] + x_t^I [Q-1] - x_t \}}$$

A.0.2 Derivation Under No Bubble Creation

When $x < (1-\epsilon)$ and calculating where $x_{t+1} = 1$

$$x = \frac{\alpha q x}{s [Q\epsilon + q(1-\epsilon) - qx]}$$

$$s [Q\epsilon + q(1-\epsilon) - qx] = \alpha q$$

$$s[Q\epsilon + q(1 - \epsilon)] - sqx = \alpha q$$

$$x = \frac{Q\epsilon + q(1 - \epsilon)}{q} - \frac{\alpha}{s}$$

In a bubbly equilibrium $x > 0$ therefore:

$$\frac{Q\epsilon + q(1 - \epsilon)}{q} - \frac{\alpha}{s} > 0$$

$$\alpha < \frac{[Q\epsilon + q(1 - \epsilon)]s}{q}$$

A.0.3 Interval in Which Bubble Creation Shifts Up $E[x_{t+1}]$

We show here that when $x_t \in (0, \frac{Q\epsilon + (1-\epsilon)q}{Q}] \cup [1 - \epsilon, 1)$,

We consider the first case of equation 1.15 and take the first derivative with respect to x_t^I , assuming $X_t^U = 0$ for simplicity. We get

$$\frac{\partial}{\partial x_t^I} E[x_{t+1}] = -\frac{\alpha q[(\epsilon - 1)q + Q(x_t - \epsilon)]}{s[Q(\epsilon + x_t^I) - q(\epsilon + x_t^I + x_t - 1)]^2} \Big|_{\frac{x_t + x_t^I}{1 - \epsilon}} < 1 \quad (\text{A.1})$$

The denominator is clearly positive as well as αq so this type of creation shifts $E[x_{t+1}]$ up as long as

$$x_t < \frac{Q\epsilon + (1 - \epsilon)q}{Q} \quad (\text{A.2})$$

Now taking the second part of equation 1.15, we get

$$\frac{\partial}{\partial X_t^I} E[X_{t+1}] = \frac{\alpha Q n_t [\epsilon(Q - 1) - Qx_t + 1]}{s(\epsilon Q - \epsilon + Qx_t^I - x_t^I - x_t + 1)^2} \quad (\text{A.3})$$

The denominator and $\alpha Q n_t$ are both positive so $E[x_{t+1}]$ is increasing as long as

$$x_t \leq \frac{\epsilon(Q-1) + 1}{Q} \quad (\text{A.4})$$

Now from equation 1.16 that $x_t < 1$ so we have that this is true as long as

$$\frac{\epsilon(Q-1) + 1}{Q} \leq 1 \quad (\text{A.5})$$

which implies that $Q \geq 1$ which was one of our initial assumptions, therefore when $x_t > (1 - \epsilon)$, $E[x_{t+1}]$ it is increasing.

A.0.4 Robustness Test: Probit with Fixed Effects

Table A.1: Probit estimates of bubbles on Breakthrough patents

VARIABLES	(1) Dummy: Breakthrough Patents > 0
Bubble (0)	-0.0115 (0.1460)
Bubble (-1)	0.0529*** (0.0134)
Bubble (-2)	0.0367** (0.0156)
Bubble (-3)	-0.0276 (0.0172)
Bubble (-4)	0.0370** (0.0148)
Bubble (-5)	-0.0013 (0.0165)
Brkt. Pats. (-1)	0.0566*** (0.0069)
Brkt. Pats. (-2)	0.0234*** (0.0050)
Brkt. Pats. (-3)	0.0113** (0.0044)
Brkt. Pats. (-4)	0.0126*** (0.0044)
Brkt. Pats. (-5)	0.0021 (0.0041)
Observations	7,350

Note: *Jackknife standard errors in parentheses*
 *** p<0.01, ** p<0.05, * p<0.1

Here we apply the same specification as in Equation 1.1, but with a new y-variable: we simply define a dummy which equals 1 if there were any breakthrough patents in a given sector-year and zero otherwise. We display the average partial effects of bubbles and the lagged dummy on the probability of having non-zero breakthrough patents per-capita in a given sector-year. We see that the first, second, and fourth lags of the

bubble dummy are still significant and of reasonable magnitude. For example, having a bubble in the previous year increases the probability of having a breakthrough patent by approximately 5%.

Appendix B: Additional Tables and Robustness Tests for Chapter 2

B.0.1 Shapley values

SHAP (SHapley Additive exPlanations) values introduced by Lundberg and Lee (2007) [44] are a method to explain how much each feature contributes to a model's prediction for a given observation. They represent the marginal contribution of a feature to the prediction, considering all possible combinations of features. By summing the SHAP values for all features and adding the model's bias term, you can recover the model's actual prediction for that instance.

In the figure B.1 we summarize the Shapley values for every observation and the 4 variables with the highest Shapley values. The dots represent individual data points. The x-axis shows the SHAP value, the color indicates the feature's actual value and the Y axis is the mean absolute SHAP value for each feature.

For example, if purple dots (high feature values) for a variable appear mostly on the right (positive SHAP values), it means that higher values of that feature tend to increase the model's predicted outcome. This allows you to see not only which features are most important overall but also how they influence predictions in different directions depending on their values.

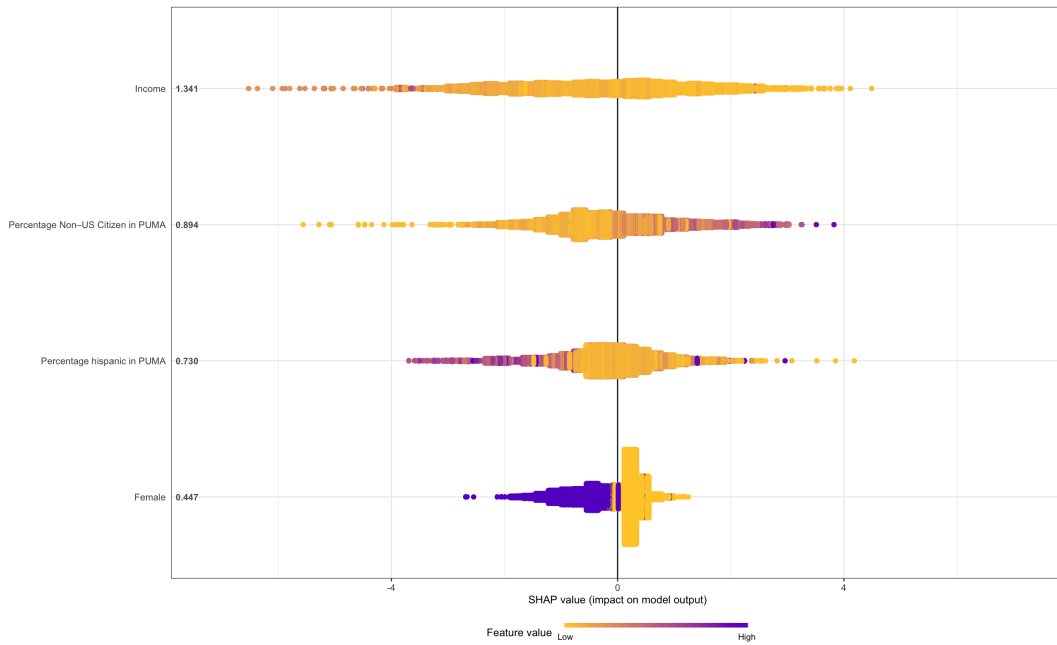


Figure B.1: Shapley Values

B.0.2 Withdrawal Rates

For non-Hispanic households, non-U.S. citizens consistently exhibit a higher application withdrawal rate, as shown in Figure B.2. This may indicate that they tend to search more and apply for multiple mortgages—in other words, they “shop around” for more favorable terms. On the other hand, Hispanic non citizens withdraw less frequently, suggesting that they may not shop around as much.



Figure B.2: Withdrawal Rate by predicted citizenship status

Table B.1: Estimates of the Effect of Non-U.S. Citizenship on Withdrawal Rates

Dependent Variable: Model:	Indicator for withdrawal of mortgage application	
	(1)	(2)
<i>Variables</i>		
Female	0.1434*** (0.0210)	0.1719*** (0.0097)
Income (000s)	0.0060*** (0.0004)	0.0073*** (7.48×10^{-5})
Loan amount (000s)	4.4×10^{-6} *** (1.15×10^{-6})	7.07×10^{-6} ** (3.52×10^{-6})
Asian	1.602*** (0.1008)	1.369*** (0.0259)
Black or African American	1.231*** (0.1215)	1.311*** (0.0249)
Hispanic	0.4319*** (0.0842)	0.6125*** (0.0211)
Non-US Citizen	0.1045* (0.0581)	0.0736 (0.0583)
Asian \times Non-US Citizen	0.3743*** (0.0919)	0.6343*** (0.0750)
Black or African American \times Non-US Citizen	0.0078 (0.1506)	0.2647* (0.1437)
Hispanic \times Non-US Citizen	-0.1939 (0.1222)	-0.4711*** (0.0720)
<i>Fixed-effects</i>		
Census Tract	Yes	Yes
Bank	Yes	
Year	Yes	Yes
<i>Fit statistics</i>		
Observations	53,463,517	53,463,517
R ²	0.07132	0.01545
Within R ²	0.00047	0.00056

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table B.2: Estimates of the Effect of being Non Us Citizen and unauthorized on Withdrawal Rates

Dependent Variable:	Indicator for withdrawal of mortgage application	
<i>Variables</i>		
Female	0.1435*** (0.0210)	0.1719*** (0.0097)
Income (000s)	0.0060*** (0.0004)	0.0073*** (7.48×10^{-5})
Loan amount (000s)	4.4×10^{-6} *** (1.15×10^{-6})	7.07×10^{-6} ** (3.52×10^{-6})
Asian	1.602*** (0.1008)	1.369*** (0.0259)
Black or African American	1.230*** (0.1215)	1.311*** (0.0249)
Hispanic	0.4319*** (0.0842)	0.6123*** (0.0211)
Non-US Citizen	-0.1023 (0.1113)	-0.0222 (0.1074)
Unauthorized	0.2935** (0.1288)	0.1358 (0.1268)
Asian \times Non-US Citizen	0.7144*** (0.1455)	0.9145*** (0.1247)
Black or African American \times Non-US Citizen	0.2836 (0.2705)	0.4662* (0.2443)
Hispanic \times Non-US Citizen	0.0758 (0.1619)	-0.2000 (0.1305)
Asian \times Unauthorized	-0.5447*** (0.1578)	-0.4836*** (0.1486)
Black or African American \times Unauthorized	-0.3988 (0.3254)	-0.2975 (0.2997)
Hispanic \times Unauthorized	-0.3802** (0.1574)	-0.3774** (0.1513)
<i>Fixed-effects</i>		
full_census_tract	Yes	Yes
rid	Yes	
year	Yes	Yes
<i>Fit statistics</i>		
Observations	53,463,517	53,463,517
R ²	0.07132	0.01545
Within R ²	0.00047	0.00056

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

B.0.3 Interest Rate

Table B.3: Estimates of the Effect of Non-U.S. Citizenship on Interest Rate

Dependent Variable: Model:	Interest rate	
	(1)	(2)
<i>Variables</i>		
Female	0.0200*** (0.0020)	0.0231*** (0.0007)
Income (000s)	0.0001 (7.38×10^{-5})	0.0004*** (7.82×10^{-6})
Loan Amount (000s)	-6.93×10^{-5} (8.23×10^{-5})	-0.0005*** (6.83×10^{-6})
Asian	-0.0896*** (0.0037)	-0.0815*** (0.0016)
Black or African American	0.0370* (0.0222)	0.0978*** (0.0030)
Hispanic	0.0352*** (0.0097)	0.1034*** (0.0021)
Non-US Citizen	0.0043 (0.0040)	0.0035 (0.0040)
Asian \times Non-US Citizen	-0.0330*** (0.0060)	-0.0418*** (0.0047)
Black or African American \times Non-US Citizen	-0.0225 (0.0143)	-0.0644*** (0.0131)
Hispanic \times Non-US Citizen	0.0270*** (0.0104)	0.0772*** (0.0062)
<i>Fixed-effects</i>		
Census Tract	Yes	Yes
Bank	Yes	
Year	Yes	Yes
<i>Fit statistics</i>		
Observations	10,950,283	10,950,283
R ²	0.57275	0.41460
Within R ²	0.00144	0.00737

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table B.4: Estimates of the Effect of being authorized and unauthorized on Interest Rates

Dependent Variable: Model:	Interest rate	
	(1)	(2)
<i>Variables</i>		
Female	0.0200*** (0.0020)	0.0231*** (0.0007)
Income (000s)	0.0001 (7.38×10^{-5})	0.0004*** (7.82×10^{-6})
Loan Amount (000s)	-6.92×10^{-5} (8.23×10^{-5})	-0.0005*** (6.83×10^{-6})
Asian	-0.0896*** (0.0037)	-0.0816*** (0.0016)
Black or African American	0.0370* (0.0222)	0.0978*** (0.0030)
Hispanic	0.0352*** (0.0097)	0.1034*** (0.0021)
Non-US Citizen	0.0030 (0.0059)	-0.0031 (0.0068)
Unauthorized	0.0019 (0.0069)	0.0095 (0.0083)
Asian \times Non-US Citizen	-0.0391*** (0.0091)	-0.0464*** (0.0075)
Black or African American \times Non-US Citizen	-0.0440* (0.0243)	-0.0947*** (0.0223)
Hispanic \times Non-US Citizen	0.0484*** (0.0125)	0.0984*** (0.0104)
Asian \times Unauthorized	0.0131 (0.0089)	0.0130 (0.0091)
Black or African American \times Unauthorized	0.0340 (0.0263)	0.0484* (0.0269)
Hispanic \times Unauthorized	-0.0296** (0.0124)	-0.0296** (0.0123)
<i>Fixed-effects</i>		
Census Tract	Yes	Yes
Bank	Yes	
Year	Yes	Yes
<i>Fit statistics</i>		
Observations	10,950,283	10,950,283
R ²	0.57275	0.41461
Within R ²	0.00144	0.00738

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

B.0.4 Risk Regression including Co-applicant

As a robustness test, we run the same delinquency and default regressions as before, but this time we include an additional dummy variable indicating whether the mortgage application had a co-applicant. Since the mortgage data does not contain co-applicant information for the years 2005 to 2007, these observations are omitted. Once again, the effect of being a non-U.S. citizen is not statistically significant.

Table B.5: Estimates of the Effect of Non-U.S. Citizenship on default

Dependent Variables: Model:	has_been_delinquent_30 (1)	has_been_delinquent_90 (2)	defaulted (3)
<i>Variables</i>			
Female	0.2323** (0.0983)	0.0435 (0.0700)	0.0592** (0.0258)
Income (000s)	-0.0127*** (0.0009)	-0.0132*** (0.0009)	-0.0042*** (0.0007)
Loan amount (000s)	0.0100*** (0.0012)	0.0140*** (0.0014)	0.0064*** (0.0010)
Credit score	-0.1712*** (0.0038)	-0.0864*** (0.0042)	-0.0236*** (0.0031)
Asian	-0.2098 (0.3458)	-0.3081* (0.1814)	-0.3026*** (0.0837)
Black or African American	6.242*** (0.3479)	4.426*** (0.3285)	0.6945*** (0.1646)
Hispanic	3.486*** (0.3348)	2.124*** (0.3885)	0.5331*** (0.1644)
Non-US Citizen	0.2493 (0.4996)	-0.2167 (0.3079)	-0.1208 (0.2101)
Co-applicant	-4.390*** (0.1537)	-2.235*** (0.1203)	-0.6982*** (0.0925)
Asian × Non-US Citizen	-1.214** (0.5694)	-0.0717 (0.3809)	0.0881 (0.1832)
Black or African American × Non-US Citizen	0.3843 (2.728)	-0.3254 (1.822)	-1.139 (0.9711)
Hispanic × Non-US Citizen	0.1174 (0.7964)	0.0925 (0.5518)	0.1194 (0.3078)
<i>Fixed-effects</i>			
Census tract	Yes	Yes	Yes
Bank	Yes	Yes	Yes
Year	Yes	Yes	Yes
<i>Fit statistics</i>			
Observations	921,568	921,568	921,568
R ²	0.18203	0.17649	0.15883
Within R ²	0.06149	0.04037	0.01050

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

B.0.5 Mortgage Data by State

Table B.6: Mortgage data by state

State	Percentage Delinquent	Percentage Hispanic	Percentage Non Us Citizen	House Price index change from 2005 until 2010
New Mexico	0.19	0.21	0.02	5.47
California	0.26	0.18	0.05	-39.24
Texas	0.20	0.15	0.04	11.30
Florida	0.37	0.13	0.04	-36.09
Arizona	0.32	0.12	0.04	-33.80
Nevada	0.38	0.12	0.04	-49.43
New Jersey	0.23	0.08	0.06	-11.29
Colorado	0.13	0.06	0.02	-2.53
Kansas	0.17	0.05	0.03	2.64
New York	0.21	0.05	0.03	-2.42
Rhode Island	0.22	0.05	0.06	-21.84
Illinois	0.17	0.05	0.03	-11.45
Washington	0.16	0.05	0.03	-0.73
Connecticut	0.20	0.05	0.05	-11.25
Utah	0.15	0.04	0.02	5.35
District of Columbia	0.16	0.04	0.04	2.67
Georgia	0.25	0.04	0.04	-14.05
Idaho	0.19	0.04	0.01	-6.48
Oklahoma	0.16	0.04	0.03	10.32
Nebraska	0.13	0.04	0.02	-0.51
Massachusetts	0.19	0.04	0.04	-15.07
Maryland	0.20	0.03	0.04	-13.62
Pennsylvania	0.18	0.03	0.02	1.82
North Carolina	0.18	0.03	0.03	2.75
Arkansas	0.18	0.03	0.02	-0.90
Oregon	0.17	0.03	0.02	-4.80
Wyoming	0.12	0.03	0.01	13.47
Delaware	0.20	0.03	0.03	-6.80
Virginia	0.17	0.03	0.02	-6.29
Indiana	0.17	0.02	0.02	-3.86
Louisiana	0.25	0.02	0.01	12.03
Tennessee	0.18	0.02	0.02	1.85
Michigan	0.21	0.02	0.02	-28.52
South Carolina	0.19	0.02	0.02	1.04
Wisconsin	0.12	0.02	0.01	-6.75
Iowa	0.15	0.02	0.02	2.63
Missouri	0.17	0.02	0.02	-5.31
Minnesota	0.16	0.02	0.03	-16.08
Mississippi	0.22	0.01	0.02	2.53
Alabama	0.21	0.01	0.01	1.91
New Hampshire	0.19	0.01	0.03	-16.26
Ohio	0.17	0.01	0.01	-11.81
Montana	0.14	0.01	0.01	9.78
Kentucky	0.16	0.01	0.01	2.37
South Dakota	0.10	0.01	0.00	8.60
West Virginia	0.22	0.01	0.01	5.20
Vermont	0.17	0.01	0.01	1.01
North Dakota	0.09	0.01	0.00	18.19
Maine	0.17	0.01	0.01	-6.95

B.0.6 Default Matching Details

The HMDA data set provides data at the census tract level, while our default data are reported at the three-digit ZIP code level. To align them, we map each census tract to its corresponding three-digit ZIP code. Since the boundaries of the census tract change

every 10 years, we adjust the mapping accordingly to ensure accurate alignment for each year. Because a census tract can correspond to multiple three-digit ZIP codes, we retain only those census tracts that uniquely match a single three-digit ZIP code.

Since our analysis focuses only on owner-occupied mortgages used for property purchases, we filter our default data to include only mortgages that meet these same criteria.

Additionally, HMDA reports whether a mortgage was sold to Fannie Mae or Freddie Mac, so we apply these conditions to ensure proper matching.

For Fannie Mae and Freddie Mac data, certain banks are explicitly reported by name. To account for this, we map the agency and respondent ID in HMDA to these banks and classify all other institutions under "Other". Finally, within each bank, we match records to HMDA data based on the same bank, origination year, ZIP code, metropolitan statistical area (MSA), and loan amount, keeping only unique matches.

B.0.7 Addressing Selection in our training data with Kernel Density Estimation

Since our original training sample included only individuals who were successful in the mortgage market—and because the income distributions from the ACS and HMDA datasets did not match—we perform a robustness test using kernel density estimation (KDE) which is a non-parametric method for estimating a probability density function, combined with conditional reweighting.

First, we stratify both datasets by race and gender to ensure that the overall proportions in the source data (ACS) match those in the target data (HMDA). Then, within each

stratum, we apply KDE to estimate the probability density function of income in the HMDA and the ACS datasets without filtering for individuals with mortgages; that is, we include all adults and heads of households who moved in the last year.

For each observation x in the source dataset, we calculated a weight as follows:

$$w = \frac{f_{target}(x)}{f_{source}(x)} \quad (\text{B.1})$$

where $f_{target}(x)$ is the KDE-estimated density for income in the HMDA data, and $f_{source}(x)$ is the KDE-estimated density for income in the ACS data, evaluated at the same income level.

This weight indicates how much more (or less) an observation should be represented to match the target distribution. If, within a particular group, the target density is higher than the source density at a given income level, the corresponding observation receives a larger weight (and vice versa). We then resample the ACS data using these weights, ensuring that both the categorical proportions and the continuous income distribution align more closely with the HMDA data.

As an example, we compare the resulting distributions in our original training dataset with those obtained after sampling from the ACS dataset using the computed weights. Figure B.3 shows the income distribution for non-Hispanic households in 2007. It is clear that, after applying this method, the income distribution in our new training data closely matches that observed in HMDA. Similarly, Figure B.4 presents the proportion of female observations; after reweighting, this proportion aligns with that in HMDA.

We retrain our models using this new training dataset, categorizing the HMDA observations as Citizens and non-US Citizens before rerunning our regressions. Table B.7 shows the approval regression results. For Hispanic households, being a non-US Citizen decreases the probability of approval by 3.04 percentage points—a significantly larger effect than our original estimate of 2.28 percentage points. For non-Hispanic households, our original estimate indicated that being a non-US Citizen increased the probability of approval by 0.23 percentage points; with the new method, this estimate increases to 0.36 percentage points.

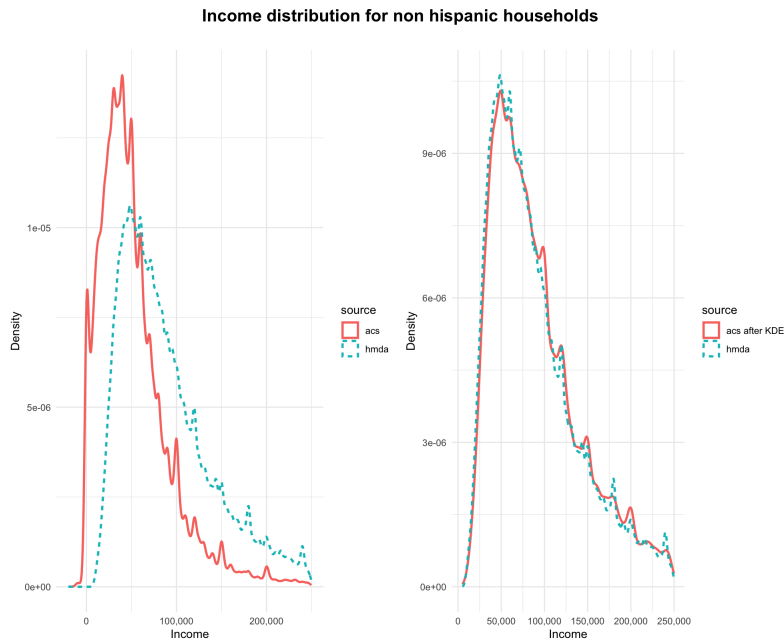


Figure B.3: Income distribution for non-Hispanic households in 2007. The left panel displays the distribution in our original ACS training dataset, and the right panel shows the distribution after sampling using the KDE-estimated densities.

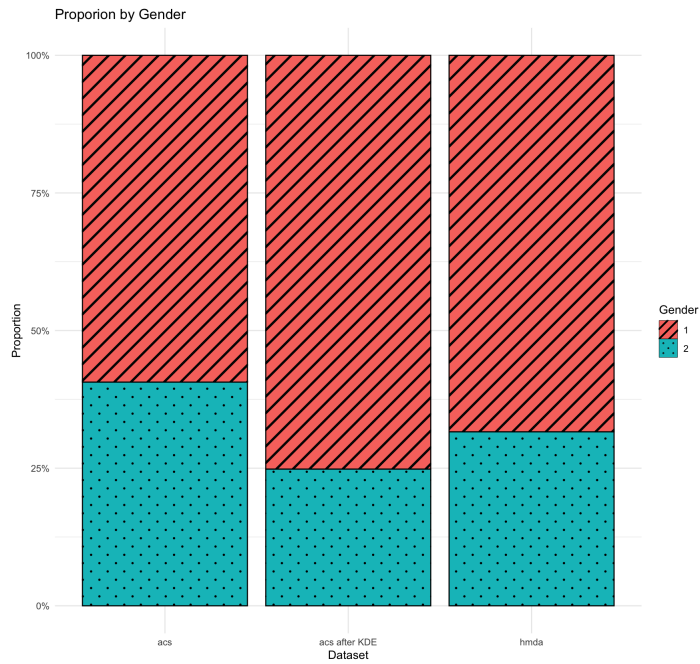


Figure B.4: Gender distribution for non-Hispanic households in 2007 for our original ACS training dataset, ACS after sampling using the KDE-estimated densities and HMDA

Table B.7: Estimates of the Effect of Non-U.S. Citizenship on Approval Rates

Dependent Variable: Model:	Indicator for mortgage approval	
	(1)	(2)
<i>Variables</i>		
Female	-0.9170*** (0.0946)	-1.606*** (0.0162)
Income (000s)	0.0164*** (0.0012)	0.0170*** (0.0002)
Loan amount (000s)	-0.0017** (0.0008)	0.0001 (0.0001)
Asian	-3.028*** (0.1127)	-3.480*** (0.0566)
Black or African American	-8.653*** (0.3108)	-12.77*** (0.0623)
Hispanic	-3.758*** (0.2666)	-5.841*** (0.0365)
Non-US Citizen	-0.0679 (0.0612)	-0.1374*** (0.0407)
Asian × Non-US Citizen	1.052*** (0.1070)	1.422*** (0.0699)
Black or African American × Non-US Citizen	0.5219** (0.2566)	0.9626*** (0.1199)
Hispanic × Non-US Citizen	-2.794*** (0.2583)	-2.915*** (0.0768)
<i>Fixed-effects</i>		
Census tract	Yes	Yes
Bank	Yes	
Year	Yes	Yes
<i>Fit statistics</i>		
Observations	46,723,049	46,723,049
R ²	0.18185	0.08042
Within R ²	0.00591	0.01058

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Bibliography

- [1] Robert J. Gordon. *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*. Princeton University Press, Princeton, NJ, 2016.
- [2] Joel Mokyr. *The Lever of Riches: Technological Creativity and Economic Progress*. 1990.
- [3] Thomas Philippon. Additive growth. Working Paper 29950, National Bureau of Economic Research, 4 2022.
- [4] Aakash Kalyani, Nicholas Bloom, Marcela Carvalho, Tarek Hassan, Josh Lerner, and Ahmed Tahoun. The diffusion of new technologies. (21-144), 6 2021. Revised October 2023.
- [5] N. F. R. Crafts. Macroinventions, economic growth, and ‘industrial revolution’ in britain and france. *The Economic History Review*, 48(3):591–598, 1995.
- [6] Ľuboš Pástor and Pietro Veronesi. Technological revolutions and stock prices. *American Economic Review*, 99(4):1451–83, 9 2009.
- [7] Mary O’Sullivan, Naomi R Lamoreaux, and Kenneth L Sokoloff. Financing innovation in the united states, 1870 to the present. 2007.
- [8] Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–20, 9 2021.
- [9] Alberto Martin and Jaume Ventura. Economic growth with bubbles. *American Economic Review*, 102(6):3033–58, 5 2012.
- [10] Paul A. Samuelson. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy*, 66(6):467–482, 1958.
- [11] Jean Tirole. Asset bubbles and overlapping generations. *Econometrica*, 53(6):1499–1528, 1985.
- [12] Alp Simsek. The macroeconomics of financial speculation. *Annual Review of Economics*, 13:335–369, 2021.
- [13] Valentin Haddad, Paul Ho, and Erik Loualiche. Bubbles and the value of innovation. Working Paper 29917, National Bureau of Economic Research, 4 2022.
- [14] Tri Vi Dang and Zhaoxia Xu. Market sentiment and innovation activities. *Journal of Financial and Quantitative Analysis*, 53(3):1135–1161, 2018.
- [15] Ramana Nanda and Matthew Rhodes-Kropf. Investment cycles and startup innovation. *Journal of Financial Economics*, 110(2):403–418, 2013.
- [16] Ming Dong, David Hirshleifer, and Siew Hong Teoh. *Stock market overvaluation*,

- moon shots, and corporate innovation*. National Bureau of Economic Research, 2017.
- [17] MALCOLM BAKER and JEFFREY WURLER. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006.
- [18] Daron Acemoglu and Fabrizio Zilibotti. Was Prometheus unbound by chance? risk, diversification, and growth. *Journal of Political Economy*, 105(4):709–751, 1997.
- [19] Robin Greenwood, Andrei Shleifer, and Yang You. Bubbles for fama. *Journal of Financial Economics*, 131(1):20–43, 2019.
- [20] Martin Kenney and Donald Patton. Firm database of emerging growth initial public offerings (ipos), 1990-2010. *Inter-university Consortium for Political and Social Research*, 2014.
- [21] Ramana Nanda and Matthew Rhodes-Kropf. Financing risk and innovation. *Management Science*, 63(4):901–918, 2017.
- [22] Azari, Shabnam Shenasi and Jenkins, Virginia and Hahn, Joyce and Medina, Lauren . The Foreign-Born Population in the United States: 2022. *American Community Survey Briefs*, 2022.
- [23] Paulson, Anna and Singer, Audrey and Newberger, Robin, Smith, Jeremy. Financial Access for immigrants: Lessons from different perspectives. 2006.
- [24] Thorsten Beck, Asli Demirguc-Kunt, and Maria Soledad Martinez Peria. Reaching out: Access to and use of banking services across countries. *Journal of Financial Economics*, 85(1):234–266, 2007.
- [25] Una Okonkwo Osili and Anna Paulson. What Can We Learn about Financial Access from U.S. Immigrants? The Role of Country of Origin Institutions and Immigrant Beliefs. *The World Bank Economic Review*, 22(3):431–455, 11 2008.
- [26] Neil Bhutta, Andreas Fuster, and Aurel Hizmo. "paying too much? borrower sophistication and overpayment in the us mortgage market". *FRB of Philadelphia Working Paper No. 24-11*, 2024.
- [27] Alexei Alexandrov and Sergei Koulayev. "no shopping in the u.s. mortgage market: Direct and strategic effects of providing information". *Consumer Financial Protection Bureau Office of Research Working Paper No. 2017-01*, 2018.
- [28] Laurie Goodman, Aniket Mehrotra, and Amalie Zinn. ITIN Mortgages, Barriers and Opportunities to Advance Latino Homeownership. 2024.
- [29] Alicia H. Munnell, Geoffrey M. B. Tootell, Lynn E. Browne, and James McEneaney. Mortgage lending in boston: Interpreting hmda data. *The American Economic Review*, 86(1):25–53, 1996.
- [30] Helen F. Ladd. Evidence on discrimination in mortgage lending. *Journal of Economic Perspectives*, 12(2):41–62, June 1998.
- [31] MICHAEL LaCOUR-LITTLE. Discrimination in mortgage lending: A critical review of the literature. *Journal of Real Estate Literature*, 7(1):15–49, 1999.
- [32] Graziella Bertocchi, Marianna Brunetti, and Anzelika Zaiceva. The Financial Decisions of Immigrant and Native Households: Evidence from Italy. *Italian Economic Journal*, 2023.
- [33] Paolo Emilio Mistrulli, Md Taslim Uddin, and Alberto Zazzaro. Discrimination Of Immigrants In Mortgage Pricing And Approval: Evidence From Italy. Mo.Fi.R. Working Papers 180, Money and Finance Research group (Mo.Fi.R.) - Univ. Politecnica Marche - Dept. Economic and Social Sciences, May 2023.

- [34] Luis Diaz-Serrano and Josep M. Raya. Mortgages, immigrants and discrimination: An analysis of the interest rates in Spain. *Regional Science and Urban Economics*, 45:22–32, 2014.
- [35] Chao Liu. Language frictions in consumer credit. 2023.
- [36] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.
- [37] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.
- [38] George J. Borjas. The labor supply of undocumented immigrants. *Labour Economics*, 46:1–13, 2017.
- [39] George J. Borjas and Hugh Cassidy. The wage penalty to undocumented immigration. *Labour Economics*, 61:101757, 2019.
- [40] S. Krannich and D. S. Massey. The effect of the COVID-19 pandemic on immigration and immigrant wellbeing in the United States. *SSM - Population Health*, 27:101705, August 12 2024.
- [41] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning (8th ed.)*. Springer Science+Business Media New York., 2017. ISBN 978-1-4614-7138-7.
- [42] Scott Frame, Ruidi Huang, Xuwei Erica, Yeonjoon Lee, Will Shuo, Erik Mayer, and Adi Sunderman. The impact of minority representation at mortgage lenders. 2023.
- [43] Richard Rosen. Competition in mortgage markets: The effect of lender type on loan characteristics. *Economic Perspectives*, Vol. 35, No. 1, 2011.
- [44] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.