

ScholarWorks@GSU

Essays on Housing and Locational Choices

Authors	Pan, Siyu
Citation	Pan, Siyu. Essays on Housing and Locational Choices. 11 Aug. 2020, Georgia State University. https://doi.org/10.57709/18661478 .
DOI	https://doi.org/10.57709/18661478
Download date	2026-03-13 18:31:48
Link to Item	https://hdl.handle.net/20.500.14694/1832

ABSTRACT

ESSAYS ON HOUSING AND LOCATIONAL CHOICES

By

SIYU PAN

August 2020

Committee Chair: Dr. Henry Spencer Banzhaf

Major Department: Department of Economics

This dissertation explores people's decision-making process of residential location and housing. The first chapter examines how environmental amenities and health affect an individual's decision about where to live. Chapters 2 and 3 investigate a historical housing market in Southern China using unique datasets. These two chapters provide insights into house price determinants and the behavior of the housing market in the short and long run, as well as the economic function of lineages.

Chapter 1 provides an important implication for epidemiology, as it implies a naïve estimation of the adverse effect of air pollution on health will be biased, as people sort based on air quality differences. This paper provides direct evidence that air-pollution-related health shocks change how a household evaluates clean air and, as a result, incentivize relocation towards better air quality. I employ a spatial equilibrium model, in which a household chooses a county to live in based on the county-level characteristics including air pollution. Using NLSY79 data, I create a panel tracking respondents respiratory health shocks and county-level location for over three decades. The estimates from a multinomial mixed logit model support the hypothesis that households move toward cleaner air after a female adult is diagnosed with asthma or becomes pregnant. I find that households react more strongly to a new asthma diagnosis for an adult than to a child's diagnosis. For illustration, the probability that a household with an adult diagnosed with asthma chooses to live at a place such as Los Angeles county would decrease from 0.0464 to 0.0459 (a 0.97% decrease) as a result of a 10-point increase in local Air Quality Index (worsened

air quality). If a family has a pregnant member, the respective decrease in probability would be 0.39%. The estimated increase in expected marginal willingness to pay for a one-unit reduction in AQI contributed by adult asthma and pregnancy are \$14.08 and \$4.97, respectively (in 1982-1984 dollars).

Chapter 2 uses a unique housing transaction dataset that I collect from the historic Huizhou prefecture in China's Anhui Province. The richness of these records provide a distinct opportunity to study economic life in late imperial China. In this paper, I focus on house and residential land transaction records to develop price indices for the Ming Dynasty, the Qing Dynasty and the Republic of China era using a hedonic regression model. The current sample includes more than 1000 records and is continuing to expand. The price indices show that house and land prices are correlated with the layout of the property, familial ties between sellers and buyers, location of the house, and whether the seller is in an urgent financial situation.

Chapter 3 aims at examining the effect of historical shocks, such as natural disasters and political turmoils, on the short and long run performance of lineages in the Huizhou area, and how lineages with different attributes respond differently to the shocks. The effect of a lineage's characteristics on the socioeconomic outcomes of its descendants has attracted attention in the economic history literature. However, identification of the long term effect is limited by the availability of historical individual-level data. This chapter obtains detailed individual-level labor and education information from genealogy records collected in China. The outcome variables that describe lineage performance include wealth, imperial civil examination results, and the labor market achievements of lineage members. The current paper does not have the complete lineage-level information; it uses the housing information and discusses the effect of historical shocks on housing prices. The current results find that in the first ten years following the shock, floods decrease house prices; national and local political turbulences have heterogeneous effects on house prices; famines drastically decrease house prices.

ESSAYS ON HOUSING AND LOCATIONAL CHOICES

BY

SIYU PAN

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree
of
Doctor of Philosophy
in the
Andrew Young School of Policy Studies
of
Georgia State University

GEORGIA STATE UNIVERSITY
2020

Copyright by
Siyu Pan
2020

ACCEPTANCE

This dissertation was prepared under the direction of the candidates Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics in the Andrew Young School of Policy Studies of Georgia State University.

Dissertation Chair: Dr. H. Spencer Banzhaf

Committee: Dr. Garth Heutel
Dr. Thomas A. Mroz
Dr. Vincent Yao

Electronic Version Approved:

Sally Wallace, Dean
Andrew Young School of Policy Studies
Georgia State University
August, 2020

Acknowledgment

Chapter 1 of this dissertation was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here do not necessarily reflect the views of the BLS.

I thank my advisor H. Spencer Banzhaf for his guidance throughout the whole process, as well as my committee members, Garth Heutel, Tom Mroz, and Vincent Yao for their generous help and insightful suggestions. I am thankful for the comments from Dan Kreisman and the audience in the PhD seminar at the Economic Department, Georgia State University. I thank the participants of the CU Environmental Economics Workshop for their comments on Chapter 1; I am also thankful for the insightful feedback from Ruixue Jia and Li Gan on Chapter 2 and 3 of this dissertation. All errors are my own.

The data collection for Chapter 2 and 3 was made possible with the generous help from the staff at the Archive Office of Yi County, the Archive Office of Huangshan City, and the Huangshan City Library; I thank them for letting me access their collections of the Huizhou Documents. I also thank Wu Bingkun at the Huangshan University Library and the faculty at the Anhui Normal University Library for granting me access to their restricted collections. Lastly, I want to thank Liu Boshan from Anhui University for his extensive effort in collecting and photographing the Huizhou Documents since the 1980s.

Contents

1	Health Shock, Air Quality and Location Choice	1
1	Introduction	1
2	Literature Review	4
2.1	Air Pollution and Health Outcomes	4
2.2	Air Pollution and Averting Behavior	5
3	Model and Empirical Strategy	7
3.1	Basic Framework	7
3.2	A Residential Sorting Model with Mixed Logit Specification	8
3.3	Marginal Willingness to Pay for Air Quality	11
4	Estimation Strategy	12
4.1	Income Imputation	12
4.2	Estimation	15
4.3	Contraction Mapping	16
5	Data and Descriptive Statistics	17
5.1	Health Conditions	19
5.2	Air Quality Measure	22
6	Results	25
6.1	Reduced Form Analysis	25
6.2	Conditional and Mixed Logit	29
6.3	Prediction of Changes in Population Composition	33

6.4	The Effect on Marginal Willingness to Pay for Air Quality	35
7	Conclusion and Discussion	37
2	The Huizhou Index	38
1	Introduction	38
2	Literature Review	40
2.1	Historical Housing Markets in China	40
2.2	Historical House Price Indices	41
3	The Data	43
3.1	House Layout	46
3.2	Currency	47
3.3	Other Variables	49
4	Summary Statistics	51
5	Results	55
5.1	Empirical Strategy	55
5.2	Price Indices	57
5.3	The Price Determinants	60
6	Conclusion and Future Work	63
7	Appendix	65
3	Historic Shocks and Lineage Performance	69
1	Introduction	69
2	Literature Review	71
2.1	Historical Data on Families and Lineages	71
2.2	The Importance of Lineages in Historical Studies	72
3	Data Description	74
3.1	Historical Shocks	74
3.2	Outcome Variables	79

3.3	Lineage Characteristics	80
4	Methodology	80
5	The Effect of Shocks on Housing Price	83
6	Future Work	88
7	Conclusion and Discussion	89
8	Appendix	91

List of Figures

4.1	Distributions of Imputed log(Income) and Actual log(Income), All Years	14
5.1	Share of Households with Any First-Time Respiratory Diagnosis (left); Share of Households with Any Respiratory Disease (right)	20
5.2	Share of Households with Children with First-Time Child Asthma Diagnosis (left); Share of Households with Children with Child Asthma (right)	20
5.3	Share of Households with First-Time Adult Asthma Diagnosis (left); Share of Households with Adult Asthma (right)	20
5.4	Share of Households with Pregnancy	21
5.5	AQI by Region	23
5.6	AQI by County, 1980	23
5.7	AQI by County, 1990	24
5.8	AQI by County, 2000	24
5.9	AQI by County, 2010	25
6.1	Relative AQI at Residing County for Households with Respiratory Shocks	26
6.2	Relative AQI at Residing County for Households with Child Asthma	26
6.3	Relative AQI at Residing County for Households with Adult Asthma	27
6.4	Relative AQI at Residing County for Households with Pregnant Members	27
6.5	Increase in MWTP for Air Quality due to Adult Asthma (1982-1984 Dollars)	36
6.6	The Increase in MWTP for Air Quality due to Pregnancy (1982-1984 Dollars)	36
3.1	Locations of Anhui and Huizhou	44

3.2	A Sample House Transaction Document	45
3.3	First Floor Floor Plan of a “Sanjian”	46
3.4	First Floor Floor Plan of a “Sihe”	48
4.1	Number of Transactions by Year	51
4.2	Number of House Transactions by Year	52
4.3	Number of Land Transactions by Year	52
4.4	The Average Room Price (log) of Residential Houses (Silver)	54
5.1	The Decade FE of (log) House Price	57
5.2	The Decade FE of (log) Residential Land Price	58
5.3	The Year FE of (log) Houses Price with Local Polynomial Smoothing	59
5.4	The Year FE of (log) Land Price with Local Polynomial Smoothing	59
7.1	The Decade FE of House Price (log) by County	65
3.1	Number of Shocks by Decade	76
3.2	Number of Natural Disasters by Decade	76
3.3	Number of Political Turmoils by Decade	77
3.4	Number of Shocks by Detailed Categories	77
3.5	Number of Natural Disasters by Detailed Categories	78
3.6	Number of Political Turmoils by Detailed Categories	78
4.1	Illustration of a D-in-D Analysis	81
5.1	Impact on House Price: Natural Disasters	85
5.2	Impact on House Price: Political Turmoils	86
5.3	Impact on House Price: Public Constructions	86
5.4	Impact on House Price: Famine	87
5.5	Impact on House Price: Industrial Development	87

List of Tables

5.1	Mean and Standard Deviation of Variables in 1980, 1990, 2000 and 2010	18
6.1	Results - Household Fixed Effects OLS Regression	28
6.2	Results - Conditional and Mixed Multinomial Logit Models	32
6.3	Results - Conditional and Mixed Multinomial Logit Models (Excluding Hospital Beds)	33
6.4	Changes in Population Associated with an Increase of 10 in AQI	34
6.5	Changes in MWTP for a 1-unit drop in AQI, Induced by Health Conditions	35
3.1	Currencies Adopted in Huizhou During the Republic of China Era	49
3.2	Different Versions of Silver Coins During the Republic of China Era	49
4.1	Number of Transactions by Item	53
4.2	Number of House Transactions by Layout	53
4.3	Number of Residential House Transaction by County	54
4.4	Summary Statistics of Variables	55
5.1	Hedonic Regression of House and Land Transaction	62
7.1	The Historic Huizhou House and Land Price Indices (log)	66
7.2	The Historic Huizhou House and Land Price Indices	67
7.3	Estimated Exchange Rate between Silver and Copper Cash (Qian)	68
3.1	The Definitions of General and Detailed Categories of the Historical Shocks	75
5.1	Results of the OLS Regressions	83

8.1 Regression Coef. of the Hedonic Var. in Eq. 4.3 91

Chapter 1

Health Shock, Air Quality and Location Choice

1 Introduction

The adverse effects of air pollution on health is an important concern of public policy, with health damages playing a dominant role in benefit-cost analyses of air pollution regulations. Much of the epidemiology driving the regulations is based on cross-sectional variation in health and pollution. These estimates have the advantage of estimating long-run effects. However, if the population vulnerable to polluted air tends to avert health risk by choosing to live in places with better air quality, cross-sectional estimates that do not consider selection based on health and air quality could be biased downward. For example, the more sensitive population might avoid a city with worse air quality. By comparing the average health of population in the city to a cleaner city, we are comparing two different populations. The adverse effect would have been larger if the selection effect could be controlled.

Recent work provides evidence showing that air quality-related health conditions could lead to migration through different pathways. Interestingly, Decker and Schmitz (2016) show that experiencing a health shock increases individual risk aversion for at least four years following the

shock; it is reasonable to assume that the increased risk aversion might nudge people into taking risk-averting actions such as moving away from dirty air. Empirical tests of Tiebout sorting suggest that people do respond to air pollution (Banzhaf and Walsh, 2008) . Given these results, we can expect populations with air-quality-related diseases to have a strong incentive to “vote with their feet.”

To test whether people actually sort themselves into different areas based on health conditions and air quality, I take advantage of health shocks and examine the effect of air quality on households’ locational choice. The panel nature of my data allows me to leverage the effect of changes in health status conditional on unobserved heterogeneity in tastes. These health shocks include first time diagnosis of asthma and other respiratory diseases for adults and children and pregnancy. I employ a spatial equilibrium model, in which a household decides which US county to live in based on the air quality, housing expenditure and per capita medical resource in each county, and an unobserved location-specific mean utility.

Empirically, I use a discrete choice logit panel similar to Bayer et al. (2009), but with micro data on locational choices matched to health status to family members. These panel data allow me to control for unobserved tastes in locational preferences using a mixed multinomial logit model. The panel data set tracks respondents’ health conditions and county-level location for over three decades using the National Longitudinal Survey of Youth (specifically, NLSY79, NLSY79 Children and Young Adults, and the NLSY Geocode data). I merge the individual-level data with county information on an Air Quality Index (AQI) from the Environmental Protection Agency (EPA), housing price from the Federal Reserve Economic Data (FRED)’s House Price Index, and per capita hospital bed count from the NLSY Geocode data.

My results support the hypothesis that households with certain health conditions are more likely to move toward cleaner air, specifically, when an adult family member is diagnosed with asthma, or when the the family is expecting a newborn. I find that households react more strongly to a new asthma diagnosis for an adult household member than to a child’s diagnosis. This result could reflect the fact that adult onset asthma is usually more severe than childhood asthma. For

illustration, the probability that a household with at least one adult diagnosed with asthma chooses to live at a place such as Los Angeles county would decrease from 0.046377 to 0.045936 (a 0.95% decrease) as a result of a one-point increase in AQI in Los Angeles county. These findings suggest that analysis of air quality's effect on health needs to account for locational sorting triggered by health conditions.

This paper contributes to the literature in the following ways. First, it contributes to the migration literature by adding an answer to the question of "why people move." Although several previous studies have provided evidence that residents tend to sort based on exogenous changes in air quality using natural experiments, such as Banzhaf and Walsh (2008), or using instrumental variables, such as Freeman et al. (2019), the question of how health affects such locational decisions in a more general setting is not clearly answered. Second, insofar as this migration is a particular kind of averting behavior intended to minimize the health impact of pollution, the paper contributes to the broader literature on averting behavior. Some work, such as Moretti and Neidell (2009), documents that people are willing to invest in defensive measures against air pollution to avoid health risk. Third, it relates to the literature on health perception and behavioral adjustment on changes in health. Smith et al. (2001) was among the first to show that health shocks provide new information and change people's perception on health by comparing data of smokers and nonsmokers. Carbone et al. (2006) and Smith et al. (2006) explore Florida residents' migration pattern as a behavioral response to environmental risks, such as hurricane damage. This paper contributes to this literature by providing new knowledge on the behavioral response in terms of location choice to air quality-related health conditions.

This paper is structured as follows. In section 2, I introduce related literature that motivates this paper. In section 3, I introduce the theoretical model and my empirical strategy. Section 4 presents summary statistics. In section 5, I show and discuss the results. Section 6 concludes and discusses the plan for future work.

2 Literature Review

2.1 Air Pollution and Health Outcomes

To construct my health shock variable, I need to identify what types of health events could be considered as both common and strongly perceived to be associated with air quality. There is a large body of research exploring the adverse effect of air quality on health. Ambient air pollutants, including particulate matter (PM), ground-level ozone, sulphur dioxide (SO₂), carbon monoxide (CO), and oxides of nitrogen (NO_x), have various negative health effects and cause economic damages. A report from the World Health Organization (2016) summarizes studies on the health impact of ambient air pollution; the diseases that could be attributed to air pollution include acute lower respiratory infection (ALRI), chronic obstructive pulmonary disease (COPD), lung cancer, ischemic heart disease (IHD), stroke, adverse birth outcomes, childhood respiratory disease, diabetes, atherosclerosis, and neurodevelopment and cognitive function. Among all outdoor air pollution-caused deaths, 40% are attributed to ischemic heart disease, 40% to stroke, 11% to chronic obstructive pulmonary disease, 6% to lung cancer and 3% to acute lower respiratory infections in children.

Respiratory health is one of the prominent effects of ambient air pollution. For example, a large evidence base in the medical literature that has been growing for several decades confirms exacerbations of pre-existing asthma could be caused by air pollution, in both epidemiological and experimental studies, from either a statistical or mechanistic perspective (Guarnieri and Balmes, 2014). Pope et al. (1995) finds that particulate matter, especially sulfate and fine particulates, was associated with cardiopulmonary and lung cancer mortality. In a more recent study, Bishop et al. (2018) find that long-term exposure of PM_{2.5} increases the probability of receiving a dementia diagnosis, adding evidence of additional damage caused by air pollution to the literature.

Air pollution has a more immediate effect on infant and children's health. Chay and Greenston (2005) find that a 1-percent reduction in total suspended particulates in the air results in a 0.35 percent decline in the infant mortality rate at the county level, with most of the effect

occurring within one month of birth. Currie and Neidell (2005) find a significant effect of CO on infant mortality. Air quality is geographically varying and the literature show that the location of residence is correlated with the distribution of respiratory diseases. Location also matters in this context. Alexander and Currie (2017) discuss the important role that residential segregation and neighborhoods play in the persistent discrepancies in asthma among children. Epidemiological literature also documents the association between prenatal pollution exposure on low birth weight, short gestation, and fetal death.

2.2 Air Pollution and Averting Behavior

Facing the compelling evidence of the detrimental effect of air pollution on health, various studies confirm that people invest in defensive measures to alleviate pollution. Gerking and Stanley (1986) consider defensive medical expenditure and were among the first to estimate marginal willingness to pay for air quality based on averting behavior. Mansfield et al. (2006) document reduction in outdoor time as an averting behavior for air pollution exposure; Zivin and Neidell (2009) confirm that the cost of intertemporally substituting activities to avoid air pollution increases with the number of consecutive days with ozone alerts. Using daily boat traffic at the port of Los Angeles as an instrumental variable for ozone, Moretti and Neidell (2009) estimate that the local cost of avoidance behavior is at least \$11 million per year. Other averting behaviors, including purchase of medication (Deschenes et al., 2017) and air purifiers (Ito & Zhang, 2016), also have been found to be used to alleviate the damage of air pollution, or used to calculate the willingness to pay for air quality. Similarly, Barreca et al. (2016) find that air conditioners have been used to reduce the health effects of extreme heat events.

The literature also documents migration in response to changes in air quality. Empirical tests of Tiebout sorting suggest that people do respond to spatial heterogeneity in environment quality. Banzhaf and Walsh (2008) use a spatial equilibrium model and predict that neighborhoods experiencing exogenous improvements in public goods also experience increased population. They test this hypothesis in the context of air quality improvements and show that

improved air quality leads to a growth in local population density. Consistent with the results, a large hedonic literature finds that, in a cross-section, houses in areas with cleaner air are more expensive. A meta-analysis done by Smith and Huang (1995) summarizes estimates of the marginal willingness to pay for reduced PM concentration from earlier studies; the mean MWTP is \$109.9 while the median MWTP is \$22.4. Chattopadhyay (1999) uses a two-step hedonic model to estimate willingness to pay for non-marginal reduction in the concentration of pollution. Bishop and Timmins (2018) estimate demand functions for clean air in the Bay Area of California and find significant heterogeneity in both MWTP and the elasticities of MWTP for ozone exposure. Currie et al. (2015) links the house price impacts of the opening and closing of polluting facilities with the health effects in a unified framework.

Bayer et al. (2009) show that the conventional hedonic model tends to underestimate the willingness to pay for local amenities when households are tied to a specific location as their “home.” Freeman et al. (2017) similarly apply their model to willingness to pay for air quality in China. My model likewise builds upon Bayer et al. (2009), controlling for distance from home, but introducing micro-level panels. It serves as a detailed investigation into determining how, and to what extent, health conditions affect the willingness to pay for air quality while taking both sorting and moving cost into account.

If the population that is more vulnerable to air pollution, such as the population with air pollution-related diseases, tends to sort away from areas with dirty air, the cross-sectional estimation of air pollution’s true effects on health would be downward biased. Coffey (2003) examines the effect of air quality on infant health after taking locational choice into account using data from Los Angeles metropolitan area in 2000. Using a two-step Heckman model to difference out the sorting bias, he finds that households with weaker infant health live closer to better air quality, proves households do sort themselves into different locations based on their health situation and air quality. This sorting behavior might happen due to different mechanisms. For example, we might expect an increase in risk aversion following a health shock to nudge people into taking risk-averting actions (Decker and Schmitz, 2016). Coffey (2003) and this paper

overlap in addressing sorting behavior, but differ in the emphases, methodologies and data. This paper emphasizes the effect of health on the valuation of air quality, while Coffey (2003) focuses on recovering a better estimate of the adverse health effect of air quality. This paper also uses panel data that covers three decades and represents the entire U.S.

Overall, the literature provides me with strong motivation to investigate the relationship between air-quality-related health conditions and locational choice from both empirical and the theoretical perspectives. The negative effect of air pollution on health is well-documented; studies show people do sort themselves into different locations based on local amenities and personal characteristics. Health shocks incentivize people to adopt defensive behaviors. Hedonic models that address moving costs generate more accurate estimation of MWTP for air quality. Based on the literature, I expect households to react to air pollution differently after experiencing air quality related health conditions.

3 Model and Empirical Strategy

In this section, I introduce a simple static model that connects air quality with a household's decision on where to live. The model is developed based on Bayer et al. (2009) with an inclusion of a health production function. The model assumes each household makes independent migration decisions each period; they choose among many alternative locations by comparing the locational attributes attached, conditioning on their individual characteristics in that period.

3.1 Basic Framework

Assume the utility of a household i in period t at location j is a function of a composite good that household i consumes (C_{ijt}), housing (F_{ijt}), health (H_{it}), the cost of moving away from the birth location and the location they lived in the last period (vector M_{ijt}), and other amenities (θ_{jt}). Health is assumed to be a function of A_{jt} , air quality (or the inverse of air pollution) that varies

among locations, and R_{jt} , medical resources that each location provides (e.g. hospitals). I also assume that health is an increasing function of air quality and medical resources with diminishing returns: $H'(A_{jt}) \geq 0$, $H''(A_{jt}) \leq 0$, and $H'(R_{jt}) \geq 0$, $H''(R_{jt}) \leq 0$. In each period, a household faces the following maximization problem:

$$\begin{aligned} \max_{j,C,F} U &= U_{ijt}(C_{ijt}, F_{ijt}; H_{ijt}(A_{jt}, R_{jt}), M_{ijt}, \theta_{jt}, \varepsilon_{ijt}) \\ \text{s.t.} \quad C_{ijt} + r_{jt}F_{ijt} &= I_{ijt} \end{aligned} \tag{3.1}$$

where r_{jt} is the housing rental price that the household faces at each location, I_{ijt} is income that they could receive at each location and ε_{ijt} represent the idiosyncratic tastes. In the equilibrium, household i maximizes utility by choosing the amount of housing consumption F_{ijt} and residential location j , which is attached to the location-varying A_{jt} , R_{jt} , I_{ijt} , r_{jt} , M_{jt} and θ_{jt} .

3.2 A Residential Sorting Model with Mixed Logit Specification

To link air quality, health and migration decisions together, as well as to generate empirical predictions, I now specify additional structure. For the health production function, assume $H_{it}(A_{jt}, R_{jt}) = A_{jt}^{\gamma_{it}^A} R_{jt}^{\gamma_{it}^R} e^{\gamma^D D_{it}}$ where $\gamma_{it}^A = \gamma^A D_{it} + \gamma_i^A$ and $\gamma_{it}^R = \gamma^R D_{it} + \gamma_i^R$. D_{it} is a vector of normalized individual demographic variables that do not change with location. γ^A and γ^R are vectors that account for observable heterogeneity in tastes. γ_i^A and γ_i^R are random coefficients that capture the unobserved individual tastes. We can write $D_{it} = S_{it} + D_{it}^{-S}$, where S_{it} is a dummy variable indicating health condition and D_{it}^{-S} contains other variables in D_{it} . In this model, health and other individual-level variables change how much a household's health responds to air quality and medical resources at each location. A limitation of the basic conditional logit model without the random coefficients is that the unobserved idiosyncratic taste could cause an endogeneity problem. For example, if it has an unobserved taste factor that makes it prefer an urban lifestyle, a family would be more likely to live in urban areas and its members would be more likely to be

diagnosed with respiratory diseases due to the worse air quality in cities. Thus, it may appear as though families with asthma prefer areas with more pollution. If we cannot tease out the unobserved variation in taste, the coefficients on the interaction term of air quality and health status would bias the estimated influence of air quality and health on locational choice. As a result, I use mixed logit to allow for variation in the coefficients of the location-varying variables, including air quality; which is captured in γ_i^A and γ_i^R . A mixed logit model takes full advantage of the panel structure of the data.

Applying a Cobb-Douglas utility function, the maximization problem can thus be expressed as

$$\begin{aligned}
\max \quad U_{ijt} &= C_{ijt}^{\beta^C} F_{ijt}^{\beta^F} H_{ijt}(A_{jt}, R_{jt}) e^{M_{ijt} + \theta_{jt} + \varepsilon_{ijt}} \\
&= C_{ijt}^{\beta^C} F_{ijt}^{\beta^F} A_{jt}^{\gamma_{it}^A} R_{jt}^{\gamma_{it}^R} e^{\gamma^D D_{it}} e^{M_{ijt} + \theta_{jt} + \varepsilon_{ijt}} \\
s.t. \quad C_{ijt} + r_{jt} F_{ijt} &= I_{ijt}.
\end{aligned} \tag{3.2}$$

ε_{ij} is assumed to have an independently and identically distributed (i.i.d.) extreme value distribution. Other variables are the same as specified earlier. Substituting the income constraint into the utility function and solving the first order condition for optimal housing demand (F^*) yields:

$$F_{ijt}^* = \frac{\beta^F}{\beta^F + \beta^C} \frac{I_{ijt}}{r_{jt}} = \frac{\beta^F}{\beta^I} \frac{I_{ijt}}{r_{jt}}. \tag{3.3}$$

Thus,

$$C_{ijt}^* = I_{ijt} - \frac{\beta^F}{\beta^F + \beta^C} \frac{I_{ijt}}{r_{jt}} r_{jt} = \frac{\beta^C}{\beta^I} I_{ijt}. \tag{3.4}$$

Substituting equation (3.3) and (3.4) into the maximization problem (3.2), the household's indirect utility can be expressed as

$$\begin{aligned}
V_{ijt} &= \left(\frac{\beta^C}{\beta^C + \beta^F} \right)^{\beta^C} \left(\frac{\beta^F}{\beta^C + \beta^F} \right)^{\beta^F} I_{ijt}^{\beta^C + \beta^F} r_{jt}^{-\beta^F} H_{ijt}(A_{jt}, R_{jt}) e^{M_{ijt} + \theta_{jt} + \varepsilon_{ijt}} \\
&= B I_{ijt}^{\beta^I} r_{jt}^{-\beta^F} A_{jt}^{\gamma_{it}^A} R_{jt}^{\gamma_{it}^R} e^{\gamma^D D_{it}} e^{M_{ijt} + \theta_{jt} + \varepsilon_{ijt}}.
\end{aligned} \tag{3.5}$$

Finally, take the natural log of each side of equation (1.5):

$$\ln V_{ijt} = \ln B + \beta^I \ln I_{ijt} + \gamma_{it}^A \ln A_{jt} + \gamma_{it}^R \ln R_{jt} + \gamma^D D_{it} + M_{ijt} + \theta_{jt} + \varepsilon_{ijt} \quad (3.6)$$

where

$$\begin{aligned} \theta_{jt} &\equiv -\beta^F \ln r_{jt} + \beta^A \ln A_{jt} + \beta^R \ln R_{jt} + \xi_{jt} \\ \beta^I &= \beta^C + \beta^F. \end{aligned} \quad (3.7)$$

θ_{jt} captures all of the location-varying attributes that do not change among individual households. ξ_{jt} is a constant that is unique to each location in each year; one can think of it as the average utility that all potential residents would gain from living in each location in a year. It controls nonparametrically for the mean utility of an area. I use the ‘‘contraction mapping’’ method to calculate these mean utilities (Berry, 1994). Assuming the idiosyncratic error term ε_{ijt} is i.i.d. logit, the probability of a utility-maximizing household i choosing location j conditional on the vector of random coefficient γ_i is:

$$L_{ij}(\gamma_i) = \prod_{t=1}^T \frac{e^{\ln B + \beta^I \ln I_{ijt} + \gamma_{it}^A \ln A_{jt} + \gamma_{it}^R \ln R_{jt} + \gamma^D D_{it} + M_{ijt} + \theta_{jt} + \varepsilon_{ijt}}}{\sum_{q=1}^J e^{\ln B + \beta^I \ln I_{iqt} + \gamma_{it}^A \ln A_{qt} + \gamma_{it}^R \ln R_{qt} + \gamma^D D_{it} + M_{iqt} + \theta_{qt} + \varepsilon_{iqt}}} \quad (3.8)$$

$\gamma^D D_{it}$ drops out of equation (3.8) as it does not change among locations. The unconditional probability is the integral of L_{it} over all values of γ_i :

$$P(\ln V_{ij} \geq \ln V_{ik}, \forall k \neq j) = \int L_{ij}(\gamma_i) f(\gamma_i | \Omega) d\gamma_i = \int L_{ij}(\gamma_i) dF(\gamma_i) \quad (3.9)$$

where $f(\gamma_i | \Omega)$ is the probability distribution function of the random coefficients and Ω are the parameters of the distribution.

3.3 Marginal Willingness to Pay for Air Quality

By calculating the marginal rate of substitution between air quality and income, the expected marginal willingness to pay for better air quality is

$$\begin{aligned} E(MWTP^A) &= \int \frac{V_A}{V_I} d\Phi(D_{it}, I_{ijt}, \gamma_i) = \int \frac{1}{\beta^I} \frac{I}{H_{it}(A_{jt})} H'_{it}(A_{jt}) d\Phi(D_{it}, I_{ijt}, A_{jt}, \gamma_i^A) \\ &= \int \frac{\gamma^A D_{it} + \gamma_i^A}{\beta^I} \frac{I_{ijt}}{A_{jt}} d\Phi(D_{it}, I_{ijt}, A_{jt}, \gamma_i^A). \end{aligned} \quad (3.10)$$

$\Phi(D_{it}, I_{ijt}, A_{jt}, \gamma_i^A)$ is the joint distribution of all individual level characteristics and the random coefficients that appear in equation (3.10). As mentioned above, D_{it} is a vector of individual demographic variables that does not vary among locations; air quality-related health status is one of them. I_{ijt} enters the expectation because it varies by location, with location determined by the realization of the other random variables.

Take the first order derivative of an individual's marginal willingness to pay with respect to air quality, we can see diminishing MWTP for air quality of a household at a given location at a certain period:

$$\frac{\delta MWTP^A_{ijt}}{\delta A_{jt}} = -[\gamma_{it}^A / (\beta^C + \beta^F)] [I_{ijt} / A_{jt}^2] < 0. \quad (3.11)$$

Better air quality increases the health of a household, which leads to a decrease in the marginal willingness to pay for clean air; as a result, we might observe households in areas with better air quality, or less air pollution, to show lower $MWTP^A$. Health conditions that are related to air quality increase households' willingness to pay for clean air, thereby nudging them to move to areas with cleaner air, at the cost of potentially lower income, change in housing expenditure and other consumption, as well as moving costs.

To examine further what effect a health shock, i.e. a change in health condition, has on

$MWTP^A$, we can rewrite equation (3.10) by splitting D_{it} into S_{it} and D_{it}^{-S} . Thus we have

$$\begin{aligned} E(MWTP^A) &= \int \frac{\gamma^A S_{it} + \gamma^A D_{it}^{-S} + \gamma_i^A I_{ijt}}{\beta^I A_{jt}} d\Phi(D_{it}, I_{ijt}, A_{jt}, \gamma_i) \\ &= \int \frac{\gamma^A S_{it}}{\beta^I A_{jt}} I_{ijt} d\Phi + \int \frac{\gamma^A D_{it}^{-S} + \gamma_i^A}{\beta^I A_{jt}} I_{ijt} d\Phi \end{aligned} \quad (3.12)$$

S_{it} is a dummy variable indicating if the household has experienced a health shock. The effect of experiencing a health shock on the expected marginal willingness to pay for air quality can be expressed as

$$E(MWTP^A | S_{it} = 1) - E(MWTP^A | S_{it} = 0) = \int \frac{\gamma^A I_{ijt}}{\beta^I A_{jt}} d\Phi \quad (3.13)$$

Note that if $\gamma^A > 0$, meaning when air quality is positively associated with a household's utility (or when air pollution is negatively associated with a household's utility), we have $E(MWTP^A | S_{it} = 1) > E(MWTP^A | S_{it} = 0)$; a household that has experienced negative health shocks would have higher willingness to pay for better air than if it have not. Equation (3.13) is one of the main estimation goals of this paper, it tells us how much an air-quality-related change in health condition could influence the marginal willingness to pay for air quality.

4 Estimation Strategy

4.1 Income Imputation

Equations (3.8) and (3.9) require data on I_{ijt} , the household's conditional income at each location. However, I do not observe I_{ijt} in locations other than where the family actually resides in each year.

Accordingly, I use income using US census data (IPUMS-USA). I estimate the unobserved potential income each household could earn in all locations in their choice set in each year, which is all of the counties that show up in my sample, by running location-and-year-specific regressions

of incomes on a vector of individual characteristics. Since the basic unit of observation in this paper is a household, I thus use information of the household head for the income imputation.

However, if we simply regress annual household income on demographics, a potential selection problem could arise - households that choose to live in j might be different from households that do not live in j in terms of unobserved factors. Dahl (2002) and Bayer et al. (2009) include a control function in the location-specific regression as a solution. Following their approach, my wage regression function is:

$$\begin{aligned} \ln I_{i,j,t} = & \alpha_{i,j,t} + \beta_{j,t}^1 \text{race}_{i,t} + \beta_{j,t}^2 \text{agegroup}_{i,t} + \beta_{j,t}^3 \text{edu}_{i,t} \\ & + \delta_{j,t}^1 \text{Pr}(C_{kt}, C_{jt} | \text{edu}) + \delta_{j,t}^2 \text{Pr}(C_{kt}, C_{jt} | \text{edu})^2 + \varepsilon_{i,j,t}^I \end{aligned} \quad (4.1)$$

$\text{race}_{i,t}$ is a vector of variables indicating the household head's race, $\text{agegroup}_{i,t}$ is a vector of dummy variables indicating if the household head belongs in a certain age interval, $\text{edu}_{i,t}$ is a vector of dummy variables indicating the household head's education level. This equation is estimated separately for each location and time period. The control function, $\text{Prob}(C_{kt}, C_{jt} | \text{edu})$ is the probability the household head was born in region k but lives in region j in period t . Since the IPUMS data do not provide birth location on the county level, I run the income imputation regressions for each state by urban status, for each year. For example, I run equation (4.1) for all of the residents in the urban areas of Georgia in year 1990 using the IPUMS data, then use the coefficients to estimate the counterfactual income for all NLSY observations as if they live in a county that is in urban Georgia in 1990. I do the same for all states. I use the Urban-Rural Classification Scheme for Counties from the National Center for Health Statistics (NCHS) to define the urban/rural status of the counties in my sample for each year. The probability is calculated as the average share of household heads with each certain education level that lives in region j but was born in region k :

$$\begin{aligned}
Pr(C_{kt}, C_{jt} | edu) = & somems * Pr(C_{kt}, C_{jt} | somems) + msgrd * Pr(C_{kt}, C_{jt} | msgrd) \\
& + somehs * Pr(C_{kt}, C_{jt} | somehs) + hsgrd * Pr(C_{kt}, C_{jt} | hsgrd) \\
& + somecol * Pr(C_{kt}, C_{jt} | somecol) + colgrd * Pr(C_{kt}, C_{jt} | colgrd) \\
& + grd * Pr(C_{kt}, C_{jt} | grd).
\end{aligned} \tag{4.2}$$

In the probability function, *somems*, *msgrd*, *somehs*, *hsgrd*, *somecol*, *colgrd* and *grd* are a set of dummy variables representing some middle school, middle school graduates, some high school, high school graduates, some college, college graduates and graduate school respectively. Each variable equals 1 if a household head has the corresponding education level, and equals 0 otherwise. For example, equation (4.2) would become $Pr(C_{kt}, C_{jt} | hsgrd) = Pr(C_{kt}, C_{jt} | hsgrd)$ for a household if the household head is a college graduate. Figure 4.1 shows the histogram of the log of actually household income decumented in the NLSY data and the imputed income. The Census measure of income takes debt into consideration while the NLSY measure does not, as a result, the distribution of the imputed income has a lower mean comparing to the NLSY income distribution.

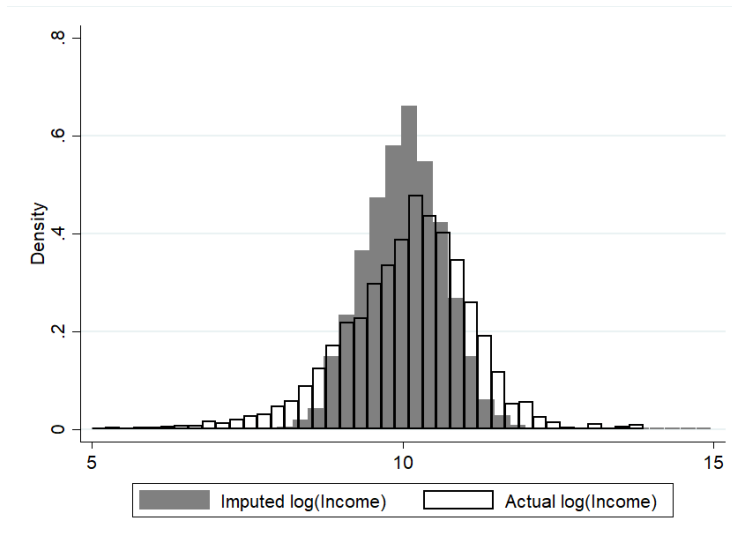


Figure 4.1: Distributions of Imputed log(Income) and Actual log(Income), All Years

4.2 Estimation

Before implementing the mixed logit model, I first estimate a conditional multinomial logit model, which is almost identical to the mixed logit model except that it does not include the random coefficients. This simpler model provides a point of comparison and a gauge of how important it is to control for unobserved heterogeneity.

A household chooses to live in region j if and only if they can achieve the highest indirect utility in region j . Based on equation (1.10) and (1.11), the log likelihood function conditioning on the random coefficients γ_i^A and γ_i^R is:

$$\begin{aligned} L_{ij}(\gamma_i) &= \prod_{t=1}^T \frac{e^{\ln B + \beta^l \ln I_{ijt} + \gamma_{it}^A \ln A_{jt} + \gamma_{it}^R \ln R_{jt} + M_{ijt} + \theta_{jt} + \varepsilon_{ijt}}}{\sum_{q=1}^J e^{\ln B + \beta^l \ln I_{iqt} + \gamma_{it}^A \ln A_{qt} + \gamma_{it}^R \ln R_{qt} + M_{iqt} + \theta_{jt} + \varepsilon_{iqt}}} \\ &= \prod_{t=1}^T \frac{e^{\ln B + \beta^l \ln I_{ijt} + \gamma^A D_{it} \ln A_{jt} + \gamma_i^A \ln A_{jt} + \gamma^R D_{it} \ln R_{jt} + \gamma_i^R \ln R_{jt} + M_{ijt} + \theta_{jt} + \varepsilon_{ijt}}}{\sum_{q=1}^J e^{\ln B + \beta^l \ln I_{iqt} + \gamma^A D_{it} \ln A_{qt} + \gamma_i^A \ln A_{qt} + \gamma^R D_{it} \ln R_{qt} + \gamma_i^R \ln R_{qt} + M_{iqt} + \theta_{jt} + \varepsilon_{iqt}}} \end{aligned} \quad (4.3)$$

I specify the random coefficients follow a normal distribution, thus the unconditional log-likelihood function is

$$P(\ln V_{ij} \geq \ln V_{ik}, \forall k \neq j) = \int L_{ij}(\gamma_i) f(\gamma_i | \bar{\gamma}, \sigma_\gamma) d\gamma_i \quad (4.4)$$

A_{tj} , R_{tj} are as previously defined. M_{itj} is a vector of variables that are proxies for moving cost, including $M1_{itj}$ and $M2_{itj}$, as defined in section 5.

The main coefficient of interest in this paper is γ^A , the coefficients of the interaction term of health status and the air quality index. It shows the significance of how experiencing a health shock changes a household's attitude towards air quality, revealed through choices on residential location. To explain it further, note that when $S_{it} = 0$, the whole term $\gamma^A S_{it} \ln A_{jt}$ disappears; when $S_{it} = 1$, $\gamma^A S_{it} \ln A_{jt} = \gamma^A \ln A_{jt}$. γ^A is capturing the part of the effect that A_{jt} has on migration decision through a health shock, it is how the population that has a certain health condition reacts to air pollution differently than the population does not have any health condition.

4.3 Contraction Mapping

The location-varying characteristic θ_{jt} consists of housing price, air quality, medical resource, and ξ_{jt} , the mean utility of unobserved county-level attributes in each year t . I borrow the method that Berry (1994) developed to calculate the values of ξ_{jt} . To do this, I run iterations of mixed logit regression to find the value of ξ_{jt} so that the predicted population shares of each location in each year is the same as the actual population shares. Which implies

$$\begin{aligned} pop_share_j &= \widehat{pop_share}_j, \text{ where} \\ \widehat{pop_share}_j &= \frac{1}{N} \sum_i P(V_{ij} > V_{ik} \forall l \neq k \mid I_i, D_i, M_i \gamma_i, A, R, r) \end{aligned} \quad (4.5)$$

would be true for all years and all locations. In each iteration, the algorithm updates the value of ξ_{jt} by adding the difference between the log of actually population share of each county j in year t and the log of estimated population share of each county j in year t to ξ_{jt} :

$$\xi_{jt}^{n+1} = \xi_{jt}^n + \ln(pop_share_j) - \ln(\widehat{pop_share}_j)^n \quad (4.6)$$

In practice, the most ideal way of doing the contraction mapping in this context is to use the complete panel data and include all U.S. counties in each observation's choice set. However, this would create a dataset that is too large to estimate with mixed logit in a timely manner. To reconcile the conflict between the ideal contraction mapping algorithm and limited computational power, I attempted to restrict the choice set to be ten randomly drawn counties in addition to the county each observation actually live in to reduce the sample size, as discussed in McFadden (1978) and Guevara and Ben-Akiva (2013). However, doing so implies that the population shares are estimated on a random subset of the full sample, thus the predicted population shares for all U.S. counties in a given year does not necessarily always add up to one, which prevents the contraction mapping algorithm from converging. I thus take another route and estimate ξ_{jt}^n for each year separately using the full data set for that year. I then insert those ξ_{jt}^n into the model for all years, treating them as data and restricting the coefficient to be one and subsampling the

choice set for the final mixed logit model regression. Since the amount of time that Stata needs for estimating the model increases exponentially with the sample size, dividing the sample by year would significantly reduce the time of contraction mapping.

5 Data and Descriptive Statistics

In this section I discuss the sources of my data and provide some descriptive statistics. Before I show the more specific summary statistics of main variables in the subsection below, Table 5.1 summarizes the means and standard deviations of all variables that I use in my estimations in 1980, 1990, 2000 and 2010 respectively.

Table 5.1: Mean and Standard Deviation of Variables in 1980, 1990, 2000 and 2010

Variable	(1)		(2)		(3)		(4)	
	1980		1990		2000		2010	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
<i>Health Conditions (S_{it})</i>								
Resp ¹	0.004	0.064	0.016	0.124	0.048	0.214	0.069	0.253
ChildAsthma ^{1a}	0	0	0.003	0.051	0.060	0.238	0.101	0.301
MotherAsthma ^{1b}	0.006	0.078	0.018	0.134	0.032	0.176	0.038	0.190
Pregnant ²	0.088	0.284	0.045	0.207	0.006	0.079	0	0
<i>Other HH Head Information (D_{it})</i>								
Female	0.510	0.500	0.504	0.500	0.507	0.500	0.507	0.500
Years of Education	11.36	3.020	11.53	4.021	9.274	4.861	10.78	4.073
Hispanic	0.201	0.400	0.199	0.400	0.198	0.398	0.197	0.397
Black	0.276	0.447	0.286	0.452	0.294	0.456	0.293	0.455
Age > 60	0.022	0.146	0.045	0.207	0.083	0.276	0.107	0.309
Imputed log(Income) ³ (I_{ijt})	9.344	0.660	9.986	0.587	10.23	0.648	10.49	0.821
<i>Moving Costs (M_{ijt})</i>								
Not in Birth County	0.167	0.373	0.316	0.465	0.362	0.481	0.377	0.485
Miles to Birth County	67.07	277.8	137.6	394.0	157.6	418.8	164.0	428.7
Moved Last year	0.075	0.263	0.068	0.251	0.052	0.222	0.030	0.170
<i>County Level Characteristics (θ_{jt})</i>								
Population (m)	1064	1683	1085	1685	1217	1929	1102	2062
Population in Poverty	118768	182712	146416	236102	137977	228847	134481	225849
Median Age	280.7	30.81	298.9	23.39	301.7	72.64	310.3	215.3
Median Family Income	14995	2613	20315	3498	36727	7860	35512	16654
Square Mile	1325	2230	1428	2533	1442	2571	1361	2632
EPA Air Quality Index ⁴	61.95	32.17	54.49	25.80	53.91	20.98	46.78	14.15
Hospital Beds	6957	4801	6492	3867	3997	2096	3925	2114
FRED Housing Price Index ⁵	52.51	14.29	74.94	12.02	100.28	11.09	133.4	25.48
<i>Imputed County Attributes</i>								
Mean Utility ⁶ (ξ_{jt})	0.02	0.723	0.010	0.569	0.012	0.708	0.008	0.837

Note: $N=7,869$ for each year. Unless otherwise specified, all variables are from NLSY79 and NLSY79 Children & Young Adults.

¹ Equals one if any member in a household has ever been diagnosed with a respiratory disease.

^{1a} Equals one if a child in a household has ever been diagnosed with asthma.

^{1b} Equals one if the mother in a household has ever been diagnosed with asthma.

^{1b,1a} The sample used for the summary statistics of those two variables is the linked mother and children pairs from NLSY79 and NLSY79 Children & Young Adults. The sample size is 3,151 for each year.

² Equals one if any member in a household is pregnant or within a year after giving birth.

³ Imputed with Census data as in Dahl (2002).

⁴ Obtained from the Environmental Protection Agency (EPA)'s pre-generated Annual Summary Data.

⁵ Obtained from the Federal Reserve Economic Data (FRED)'s All-Transactions House Price Index for the United States .

⁶ Estimated with contraction mapping.

5.1 Health Conditions

For health information and other geographic variables, I use the National Longitudinal Survey (NLSY79) datasets. The NLSY79 Original Cohort has surveyed a representative sample of American youth born between 1957 and 1964 since 1979. The sample originally included 12,686 respondents aged 17-22 in 1979; 9,964 respondents remain in the eligible samples as of the most recent wave. NLSY79 Children and Young Adults is an addition to all their mother's information in the original NLSY79; this cohort can be linked to NLSY79 through mothers' ID. It includes assessments of each child and additional demographic and development information. This dataset started in 1986. The NLSY79 original cohort is annual on and before 1994 and biannual on and after 1996; NLSY79 Children and Young Adults is biannual throughout the entire survey period. For this paper, I show results separately from two samples: one is the NLSY79 original cohort; I link NLSY79 Children and Young Adults to their mothers in the NLSY79 original cohort to create a sample that consists mother-children pairs as its units of observations, which I refer to as "households" for simplification.

I obtain county-level location information from the restricted geocoded version of NLSY79. The publicly available version of the NLSY provides geographic information only at the region level. In NLSY79 Children and Young Adults, diagnosis information of asthma is provided as a retrospective question. As a result, even though the dataset itself is biannual after 1994, we can still infer in which year the child was first diagnosed with asthma. Table 5.2 shows the number of children being diagnosed with asthma for the first time each year, listed by region.

To present the information above in a more intuitive way with a household as the basic unit, I show the following graphs. The left panel of Figure 5.1 shows the percentage of households with at least one person diagnosed with any respiratory disease for the first time in the sample each year. The right panel of Figure 5.1 shows the the percentage of households with at least one person that were ever diagnosed with any respiratory disease in the sample each year. It is the cumulative measure of the variable shown on the left panel.

Similarly, the left panel of Figure 5.2 shows the percentage of households with children

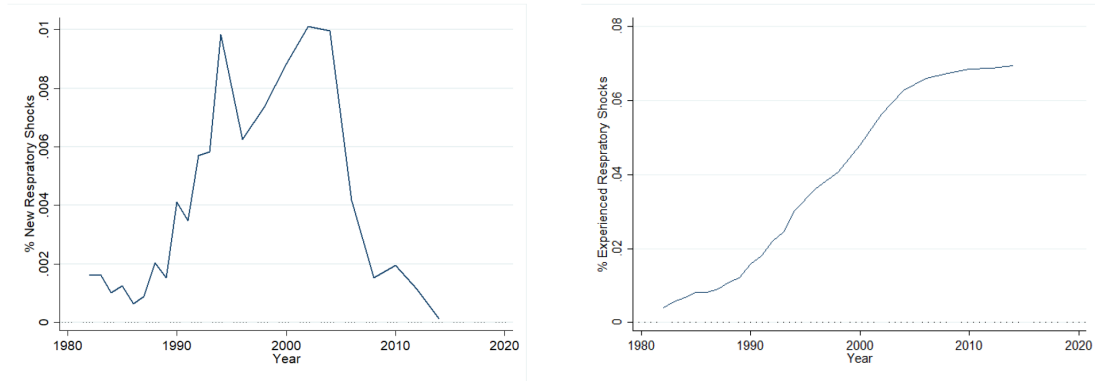


Figure 5.1: Share of Households with Any First-Time Respiratory Diagnosis (left); Share of Households with Any Respiratory Disease (right)

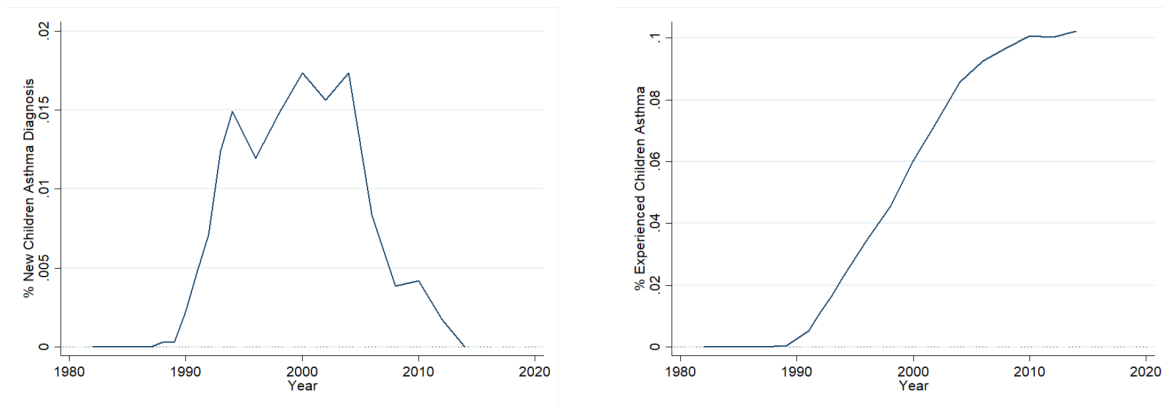


Figure 5.2: Share of Households with Children with First-Time Child Asthma Diagnosis (left); Share of Households with Children with Child Asthma (right)

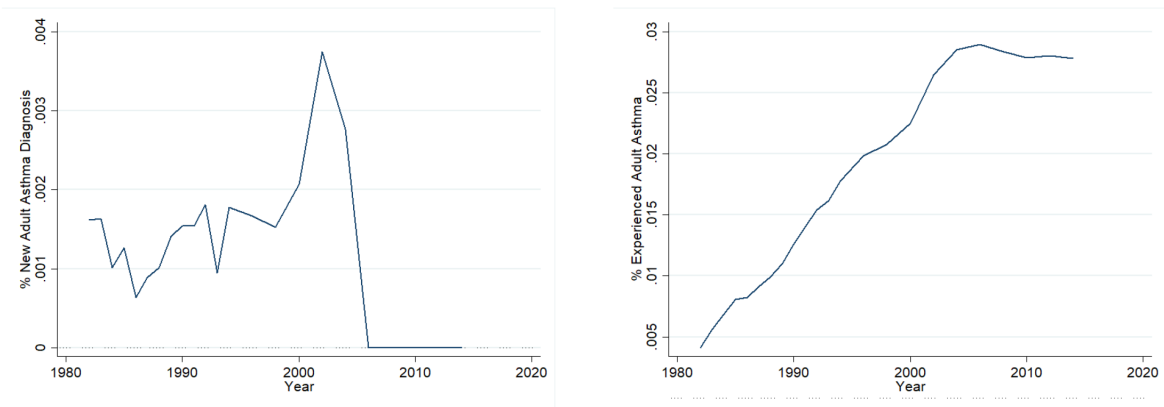


Figure 5.3: Share of Households with First-Time Adult Asthma Diagnosis (left); Share of Households with Adult Asthma (right)

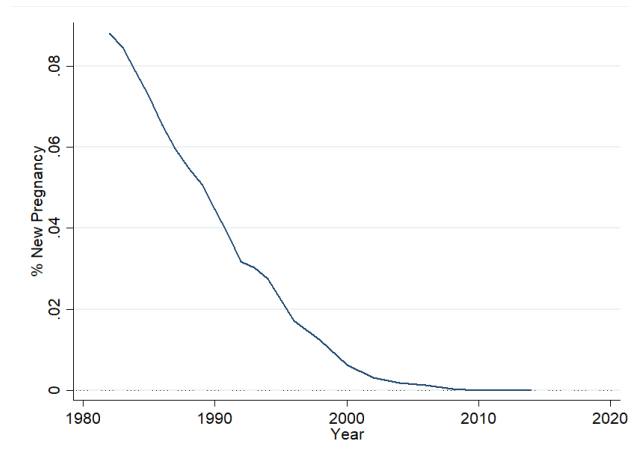


Figure 5.4: Share of Households with Pregnancy

that have at least one child diagnosed with asthma for the first time in the sample each year; the right panel is the cumulative measure. The left panel of Figure 5.3 shows the percentage of households that with at least one adult diagnosed with asthma for the first time in the sample each year; the right panel is the cumulative measure. Figure 5.4 shows the share of families with at least one pregnant woman in each year.

The reason that we see a concurrent increasing-decreasing trend with children’s asthma is that the respondents of NLSY79 are all around the same age (17-22 years old in 1979). As a result, their children also tend to be born around the same time. Asthma is a disease that starts in childhood but after 5 years old. As a result, we observe the highest share of children being diagnosed with asthma for the first time in the sample during 1993-2005. To construct the health shock variable, I also take the health status of other members in a household into account. NLSY79 records when adults are first diagnosed with asthma. The same is true for NLSY79 Children and Young Adults, for which the original cohorts data also provides this information with a retrospective question. As a result, I can identify the adults that were diagnosed with asthma in each year.

5.2 Air Quality Measure

American air quality information is available from the U.S. Environmental Protection Agency (EPA). I use an annual air pollutant concentration summary, known as the Air Quality Index (AQI), collected by outdoor monitors across the US to construct my air quality measurement variable for each county. The AQI is a piecewise linear function of the pollutant concentration, including particulate matters (PM), ground-level ozone, sulphur dioxide (SO₂), carbon monoxide (CO), and oxides of nitrogen (NO_x). The higher the AQI value is, the greater the level of air pollution is. AQI in the range of 0 to 50 is considered satisfactory. AQI between 51 and 100 is considered to be “moderate”. There are moderate health concerns for a very small number of people who are sensitive to certain pollutants. “Unhealthy for sensitive Groups” AQI is between 101 and 150. People with heart and lung disease, older adults and children are at greater risk from exposure to ozone and particulate matters. AQI between 151 and 200 is considered “unhealthy”. Most populations could experience some adverse health effects. “Very Unhealthy” AQI is 201 to 300, everyone may experience more serious health effects. “Hazardous” AQI is greater than 300. This would trigger a health warnings of emergency conditions for the entire population.

Figure 5.5 shows the population-weighted Air Quality Index for each region. Figure 5.6, 5.7, 5.8 and 5.9 shows the annual mean AQI measure for each county in 1980, 1990, 2000 and 2010 respectively. We do observe a trend of reducing air pollution as similarly shown in Figure 5.5, but comparing to the regional data, we have more variation in air pollution with county level information.

This paper also uses hospital information in each county, as another important reason for people with poor health condition to move is living in an area with good medical resource. For this purpose, I use the number of hospital beds per 100,000 residents provided in the NLSY geocode data. Other geographical level data including the Housing Price Index (HPI) from the Federal Reserve Economic Data (FRED), county population, county area and county medium household income from the National Historical Geographic Information System (NHGIS).

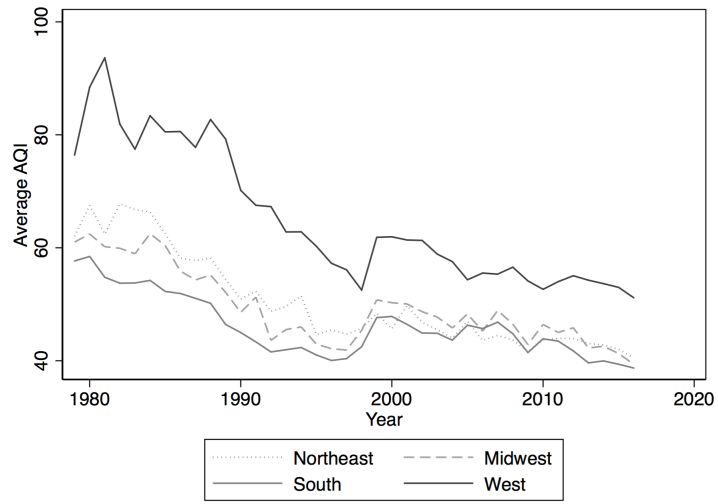


Figure 5.5: AQI by Region

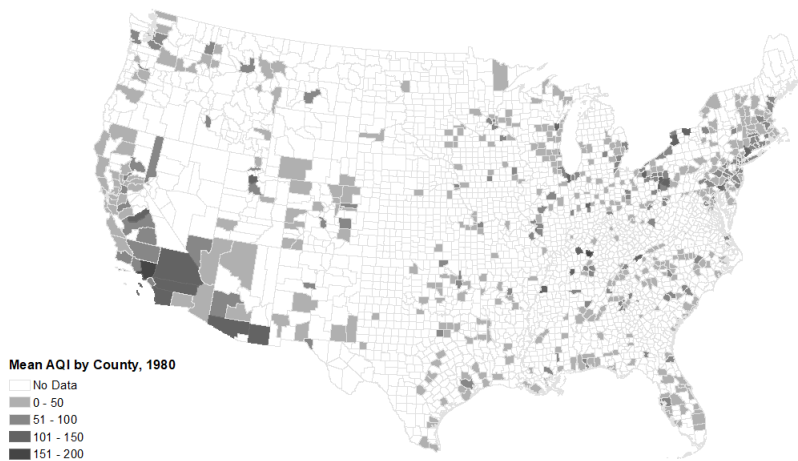


Figure 5.6: AQI by County, 1980

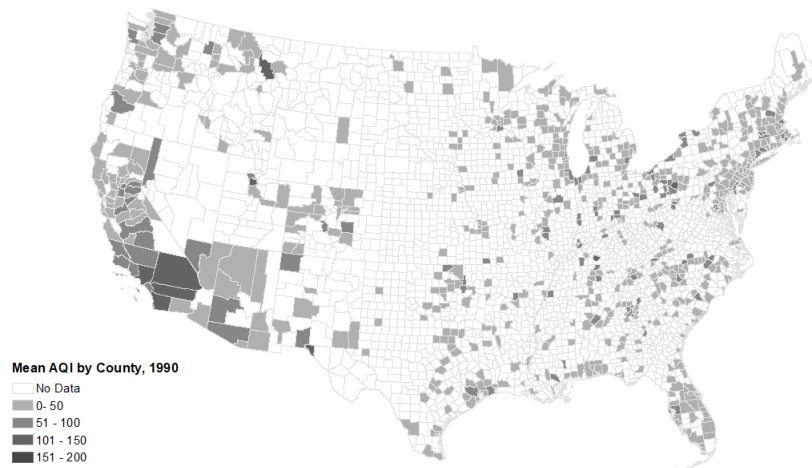


Figure 5.7: AQI by County, 1990

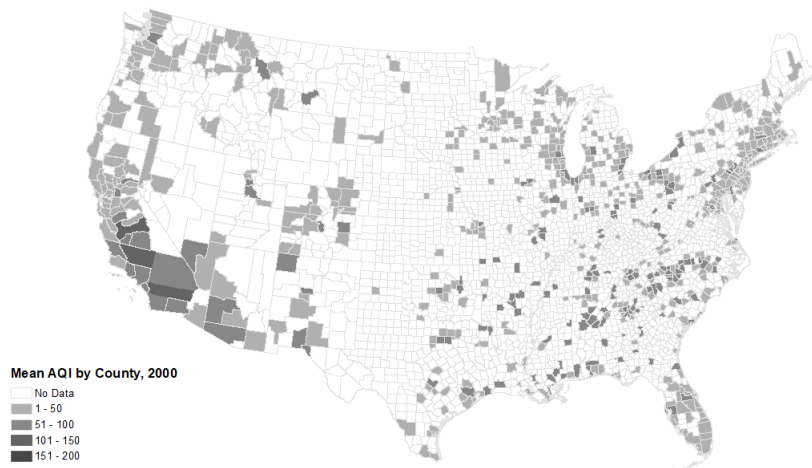


Figure 5.8: AQI by County, 2000

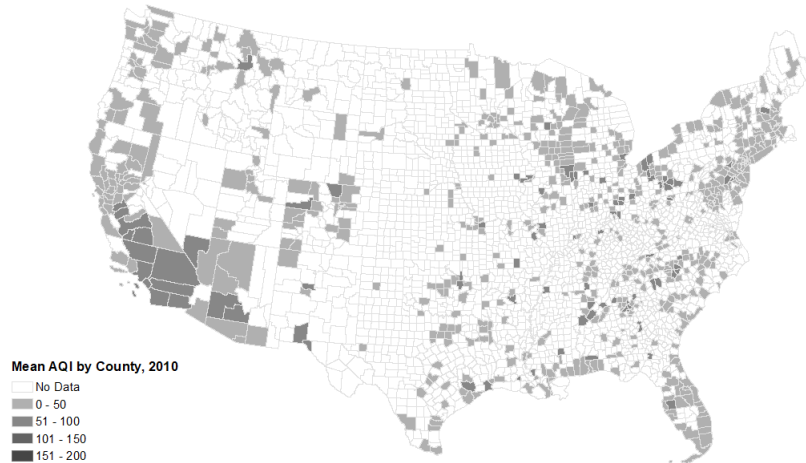


Figure 5.9: AQI by County, 2010

6 Results

6.1 Reduced Form Analysis

Before discussing the estimates generated by the conditional and mixed logit models, I consider patterns in the raw data. Figure 6.1 shows the result from a simple event study using a subsample with households that experienced changes in health conditions, including the first time diagnosis of a respiratory disease and pregnancy. The x axis shows years centered around the year each household had the first health shock. The y axis is the average difference between AQI in counties where households with the health conditions live and the national average AQI. We do not observe any obvious change around the shocks. Figure 6.2 similarly shows the difference in average AQI at counties where households with a asthmatic child live and the national average AQI. There is also no noticeable change in the relative AQI around the shocks. Figure 6.3 shows the difference in average AQI at counties where households with an asthmatic adult live and the national average AQI. We can see a pattern of relative county AQI dropping starting about three years before the health shocks. Figure 6.4 shows the relative AQI for households with a pregnant member. The relative AQI starts lowering about one year before the pregnancy. A common pattern that we see here is that the households with any type of respiratory health conditions tend to live

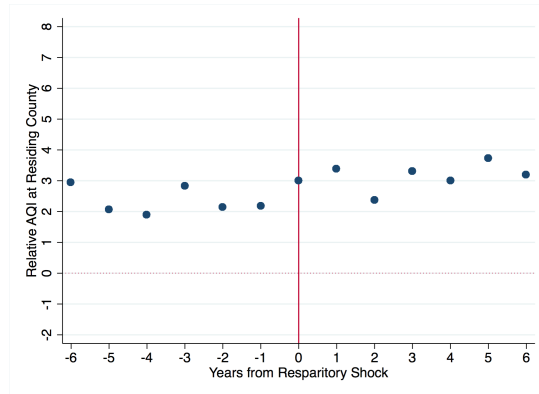


Figure 6.1: Relative AQI at Residing County for Households with Respiratory Shocks

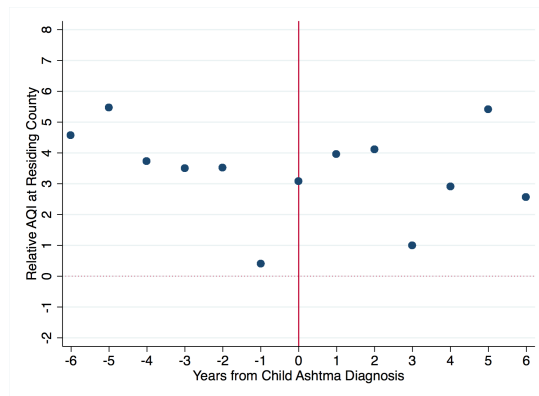


Figure 6.2: Relative AQI at Residing County for Households with Child Asthma

at locations that consistently has dirtier air than the national average. However, more detailed investigation is needed to determine whether that pattern is driven by the changes in health.

I also consider a fixed effect OLS estimation with the following equation:

$$AQI_{jt} = shocks_{it} + \log(income_{it}) + D_{it} + \theta_{jt} + county_j + year_t + \varepsilon_{it} \quad (6.1)$$

Note that all of the observations that I use to run this regression are actual observations, without the counterfactual ones that the conditional and mixed logit models require. The coefficients of the shock variables in this regression show the effect of various health conditions on the mean level of air pollution level in the counties of residence in the sample. Table 6.1 provides results from three separate regressions. Each column include a respiratory health condition and a pregnancy.

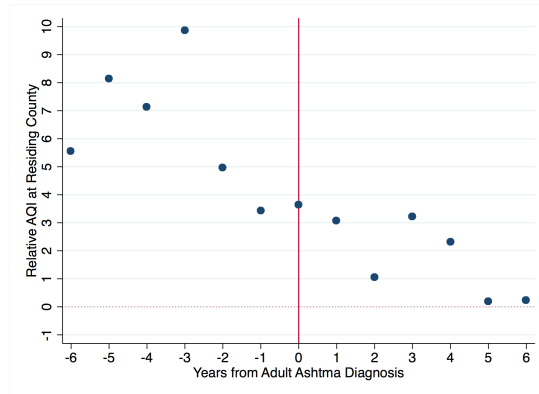


Figure 6.3: Relative AQI at Residing County for Households with Adult Asthma

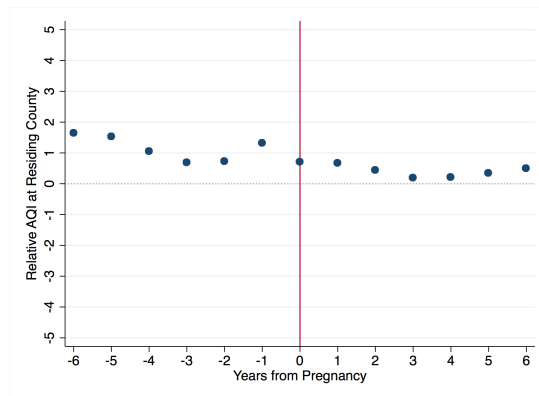


Figure 6.4: Relative AQI at Residing County for Households with Pregnant Members

Column (1) defines the respiratory health shock as whether a household has experienced any diagnosis of a respiratory disease. Column (2) defines the respiratory health shock as whether a household has at least one child experienced a diagnosis of asthma. Similarly, column (3) defines the respiratory health shock as whether a household has at least one adult experienced a diagnosis asthma. Pregnancy as a health shock is included in all three regressions. Column (1) and column (3) use the full sample with all observations, while column (2) uses a subsample with only families that have children.

Table 6.1: Results - Household Fixed Effects OLS Regression

Outcome Variable: AQI at Residing Counties			
	(1)	(2)	(3)
	All Respiratory	Children Asthma	Mothers Asthma
Resp	-0.473** (-2.64)	-0.550* (-2.54)	-1.051** (-2.65)
Pregnant	0.017 (0.14)	-0.017 (-0.14)	-0.012 (-0.10)
log(Income)	-0.382*** (-13.76)	-0.402*** (-10.37)	-0.411*** (-10.26)
HH heads Edu	-0.016* (-2.42)	0.021* (2.23)	0.022* (2.34)
County Level Characteristics (θ_{it})			
Area (sqmi)	0.001*** (7.05)	0.002*** (6.02)	0.002*** (6.00)
Population	-0.000*** (-143.65)	-0.000*** (-89.49)	-0.000*** (-89.51)
Median Family Income	0.001*** (75.52)	0.001*** (44.64)	0.001*** (44.61)
Median Age	0.014*** (32.11)	0.015*** (19.78)	0.014*** (19.76)
Poverty Population	-0.000*** (-28.75)	-0.000*** (-17.26)	-0.000*** (-17.28)
Hospital Beds	0.000*** (5.43)	0.000*** (3.39)	0.000*** (3.35)
Housing Price Index	-0.051*** (-18.48)	-0.053*** (-11.77)	-0.051*** (-11.77)
County FE	✓	✓	✓
Year FE	✓	✓	✓
State × Year FE	✓	✓	✓
R^2	0.862	0.851	0.887
N	173118	69322	69322

t statistics in parentheses

* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the results from a reduced form three household fixed effect OLS regressions which examine the effect of health shocks on the AQI at residing locations. Column (1), (2) and (3) show results from the same models, but they differ in the definition of the respiratory conditions. The respiratory condition in column (1) is defined as any diagnosis of a respiratory disease of a household member. The respiratory condition in column (2) is defined as asthma diagnosis for a child. The respiratory condition in column (3) is defined as asthma diagnosis for the mother. For panel (2) and (3) I only include households with children in the estimation sample. All three column contain pregnancy as a health shock. I also include county fixed effects, year fixed effects, and state by year fixed effects in all three columns.

The results of the fixed effects OLS regression, shown in the first column of Table 6.1, suggest that any respiratory health condition of a family member would lead to a somewhat

significant reduction in air pollution (AQI) of the counties they choose to reside in. However, by comparing column (2) and column (3), we can see that the main driving force for the relocation concerning air pollution shown in column (1) is the diagnosis of asthma for an adult in the family. The coefficient of asthma diagnosis for a child in the family is not significant, although still negative. This result aligns with the estimates from my mixed logit regression, which I will discuss more in detail in the next subsection.

Being pregnant does not have a significant effect on the AQI of residing counties in all three regressions. This particular result, however, does not align with what the mixed logit model suggests. We will learn more from the results of the conditional and mixed logit regressions below.

6.2 Conditional and Mixed Logit

Estimating the structural model of locational choice provides a more complete picture, as described in equation (4.4) and (4.5). Table 6.2 displays the results. Each column differs in terms of the definition of the respiratory health shock, which are similarly defined as in the three columns of Table 6.1. In each column, there are two subcolumns showing the different results from the conditional logit and the mixed logit model. The conditional logit regression does not include random coefficients on AQI and per capita hospital beds. The results estimated using the mixed logit model is the main result, with the conditional logit model is serves as a point of comparison model.

Column 1 and 3 of Table 6.2 shows the results from the sample with all households, regardless of whether there are children in the family. Column 2 of Table 6.2 shows the results from the subsample of households with a mother from the original NLSY79 cohort who has at least one child in the NLSY79 Children and Young Adults cohort. In this column, the respiratory health shock is defined as whether the family has had a child diagnosed with asthma. Pregnancy as a different type of health shock is included in all of the regressions with different definitions of respiratory health shocks. Overall, I examine the coefficients as presented on the first row of

Table 6.2, which are the coefficients of the interaction term between respiratory health shocks and AQI, and the third row of Table 6.2, the coefficients of the interaction term between pregnancy and AQI. Overall, I found no significant effect of a general respiratory health condition for any household member on the change in preference for air pollution. I also found no significant effect of a asthma diagnosis for a child in the household. I do find that households who experience a diagnosis of adult-onset asthma or a pregnancy tend to significantly increase their avoidance of air pollution after the shock. These estimates condition on time-invariant heterogeneous tastes.

The coefficients of Resp^*AQI and $\text{Pregnant}^*\text{AQI}$ could be interpreted as how much the health shock has made a household to dislike air pollution more than if they did not experience such a shock. A negative coefficient suggests that experiencing a health shock increases the dislike for air pollution. Closer examination of Table 6.2 reveals that mothers' asthma diagnosis is the major contributor for the changes in the preference for AQI. The coefficient in columns 3 are larger than those of children's in column 2, which implies that children's health shock has limited effect on the household's locational choice comparing to mothers' health situation. Being diagnosed with asthma for the first time as an adult on average decreases one's possibility in living in a county with higher AQI (worse air quality). A possible reason is that adult-onset asthma is usually associated with more severe health consequences comparing to being diagnosed as a child (CDC, 2013). Childhood asthma often has occasional symptoms in response to allergic triggers or respiratory infections that could be easily controlled, while asthma developed in adulthood tend to have more persistent and more severe symptoms that requires daily medication. As a result, people react more strongly to adult asthma. This result is consistent with the fixed-effect OLS estimation.

The coefficients on the all of the interaction terms of health shocks S_{it} with hospital beds tend to be extremely small. Those variables might be correlated with AQI while not providing additional explanation to the story. Accordingly, I consider another version of the model, dropping the interaction terms involving hospital beds ($beds_{jt}$). As shown in Table 6.3, excluding hospital beds does not change the results by much.

Due to the strong correlation between air pollution and health of infants and fetuses documented in the literature, the significant effect of being pregnant and expecting a newborn on the preference on AQI is shown in my results. This finding does not align with the results from the fixed-effect OLS estimation, which finds no significant effect of being pregnant on the choice of air quality at residing county. This could be caused by the absence of ξ_{jt} in the OLS regression as some of the unobserved county-level attributes are ignored. A household expecting a new born might move due to reasons that are more complex compared to a household experiencing a respiratory health shock. For example, following pregnancy, a mother might move to area with better public welfare system, or move closer to family and friends. They could also move into bigger apartments or houses. Some might even take the quality of school district into consideration. I argue that the contraction mapping term ξ_{jt} captures some of the unobserved characteristics which is contributing to the different results in table 6.1 and 6.2. However, it would be interesting to analyze the details of migration behavior of a household expecting newborns in the future work.

The coefficients on income is always negatively significant, simply suggest that households get higher utility from places there they earn higher income. Lastly, it is also no surprise that all of the migration cost variables have significantly negative effects on the location preference, meaning the cost of moving is tying people to where they were born, where they lived, or where that are closer to their birth counties.

Table 6.2: Results - Conditional and Mixed Multinomial Logit Models

	(1)		(2)		(3)	
	All Respiratory		Children's Asthma		Adults' Asthma	
	Clogit	Mlogit	Clogit	Mlogit	Clogit	Mlogit
Resp*AQI	-0.003 (-0.61)	-0.002 (-0.33)	-0.000 (-0.35)	0.000 (0.01)	-0.017* (-2.19)	-0.017* (-1.77)
Resp*Beds	0.000 (0.01)	-0.000 (-0.05)	0.000 (0.42)	0.000 (0.28)	0.000 (0.35)	0.000 (0.24)
Pregnant*AQI	-0.008** (-2.91)	-0.008* (-2.52)	-0.006** (-2.11)	-0.006* (-1.96)	-0.006** (-2.16)	-0.006* (-2.01)
Pregnant*Beds	0.000 (0.34)	0.000 (0.21)	0.000 (0.21)	0.000 (0.05)	0.000 (0.21)	0.000 (0.04)
log(Income)	0.070*** (4.86)	0.068*** (5.07)	0.067** (2.46)	0.056* (2.60)	0.067** (2.47)	0.057* (2.60)
AQI (SD)		-0.019*** (-18.79)		-0.019*** (-11.25)		-0.018*** (-11.14)
Beds (SD)		0.000 (0.79)		0.000* (2.52)		0.000* (2.56)
ξ_{jt}	✓	✓	✓	✓	✓	✓
AQI* D_{it}^{-S}	✓	✓	✓	✓	✓	✓
Beds* D_{it}^{-S}	✓	✓	✓	✓	✓	✓
Moving Costs	✓***	✓***	✓***	✓***	✓***	✓***
<i>N</i>	1,885,950		753,863		753,863	

t statistics in parentheses

* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$

The coefficient of ξ_{jt} , the contraction mapping term, is constrained to 1.

Note: This table shows the results from the conditional logit model and the mixed logit model. The mixed logit model differs from the conditional logit model in that it assigns random coefficients to AQI and per capita hospital beds, and that it acknowledges the panel structure of the data. As a result, this table shows the coefficients on the mean and standard deviation of AQI only for the mixed logit model. Panel (1), (2) and (3) show estimates from the same models, but they differ in the definition of the respiratory shocks. The respiratory shock in panel (1) is defined as any diagnosis of a respiratory disease of a household member, the respiratory shock in panel (2) is defined as asthma diagnosis for a child. The respiratory shock in panel (3) is defined as asthma diagnosis for the mother in the household. I only include families with children in the estimation sample for panel (2) and (3). All three panels contain pregnancy as a health shock.

Table 6.3: Results - Conditional and Mixed Multinomial Logit Models (Excluding Hospital Beds)

	(1)		(2)		(3)	
	All Respiratory		Children's Asthma		Mothers' Asthma	
	Clogit	Mlogit	Clogit	Mlogit	Clogit	Mlogit
Resp*AQI	-0.004 (-0.61)	-0.002 (-0.33)	-0.000 (-0.31)	0.000 (0.08)	-0.018* (-2.15)	-0.016* (-1.73)
Pregnant*AQI	-0.008*** (-2.93)	-0.007* (-2.53)	-0.006** (-2.13)	-0.006* (-1.96)	-0.006** (-2.18)	-0.006* (-2.02)
log(Income)	0.074*** (5.88)	0.076*** (6.13)	0.071*** (2.77)	0.063*** (2.93)	0.071*** (2.77)	0.063** (2.93)
AQI (SD)		-0.019*** (-18.72)		-0.018*** (-11.12)		-0.018*** (-11.01)
ξ_{jt}	✓	✓	✓	✓	✓	✓
AQI* D_{it}^{-S}	✓	✓	✓	✓	✓	✓
Moving Costs	✓***	✓***	✓***	✓***	✓***	✓***
<i>N</i>	1,885,950		753,863		753,863	

t statistics in parentheses

* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$

The coefficient of ξ_{jt} , the contraction mapping term, is constrained to 1.

Note: This table shows the results from the conditional logit model and the mixed logit model. The mixed logit model differs from the conditional logit model in that it assigns random coefficients to AQI and per capita hospital beds, and that it acknowledges the panel structure of the data. As a result, this table shows the coefficients on the mean and standard deviation of AQI only for the mixed logit model. Panel (1), (2) and (3) show estimates from the same models, but they differ in the definition of the respiratory shocks. The respiratory shock in panel (1) is defined as any diagnosis of a respiratory disease of a household member, the respiratory shock in panel (2) is defined as asthma diagnosis for a child. The respiratory shock in panel (3) is defined as asthma diagnosis for the mother in the household. I only include families with children in the estimation sample for panel (2) and (3). All three panels contain pregnancy as a health shock.

6.3 Prediction of Changes in Population Composition

To present the effect of air pollution on location choice in a more intuitive way, Table 6.3 demonstrates the effect of hypothetically worsened air quality in Los Angeles county on the probability that each population group would choose to reside there, using data from 2000 and 2002. Column (1) shows the current prediction of the average probabilities of residing in LA county, column (2) shows the average probabilities of residing in LA county if the AQI increases by 10, while holding everything else constant. There are slight increases in the total population

and among some sub-population, this could be explained by the fact that AQI is tightly related to local economic development. Even with an increase in the total population, the worsened air pollution pushes families that have experienced an adult asthma or families expecting newborns out of the county. This suggests that the associated benefit of local economic development is impaired by the increased air pollution. Such estimates could be important for policy makers to grasp a more comprehensive picture of the effects of air pollution on the local population that are associated with health outcomes.

Table 6.4: Changes in Population Associated with an Increase of 10 in AQI

	(1)	(2)	(3)
	Probabilities of Living in LA Before and After AQI Increases by 10		
	Before	After	Average Change
All Population	.04566765	.04571033	+.00004268 (+0.09%)
<i>Health Conditions:</i>			
Adult Asthma	.04635124	.04590715	-.00044409 (-0.97%)
Pregnant	.05944041	.05921062	-.00022979 (-0.39%)
<i>Populations by HH Head Characteristics:</i>			
Age > 60	.04054004	.04048711	-.00005293 (-0.13%)
College Graduate	.04854411	.04867658	+.00013247 (+0.27%)
Female	.04407196	.04411236	+.0000404 (+0.09%)
Black	.02942283	.02952134	+.00009851 (+0.33%)
Hispanic	.12996762	.12996325	-.00000437 (-0.00%)
Other Race	.02330095	.02332925	+.0000283 (+0.12%)

* Predictions based on observations from 2000 and 2002.

Note: the first column shows the predicted probability of living in the County of Los Angeles. The probabilities in column (1) are calculated using the estimates from the mixed logit model. I first predict the possibilities of residing in LA County in 2000 and 2002 for each observation. I then calculate the mean possibilities for each subsample, identified by the conditions listed in the left panel. These include: households that has experienced adult asthma, household that has a pregnant member, households that has experienced adult asthma and has a pregnant member at the same time; households heads that are older than 60, households heads with college or higher education, household heads are female, household heads are black, household heads are Hispanic, and households heads that are of other races. The probabilities in column (2) are calculated in a similar fashion, after I increase the AQI at LA county by 10. Prediction is based on data from 2000 and 2002. I choose these two years as they contain large enough amount of households that have respiratory health conditions, while they are not too late so they contain reasonable amount of pregnant observations as well.

6.4 The Effect on Marginal Willingness to Pay for Air Quality

As suggested in Equation (3.13), I use the results to generate two estimates of the increase in the marginal willingness to pay for air quality induced by diagnoses of adult asthma and pregnancy respectively. The mean increase in the marginal willingness to pay for a one-unit reduction in AQI that is induced by diagnoses of adult-onset asthma is about \$14.08 (1982-1984 dollar), while the mean increase in the MWTP for a one-unit reduction in AQI induced by pregnancy of adult-onset asthma is about \$4.97 (1982-1984 dollar). The distribution of estimated marginal willingness to pay for a unit of reduction in shown in Figure 6.5 and 6.6. We can see that both distributions have a rightward skew.

Table 6.5: Changes in MWTP for a 1-unit drop in AQI, Induced by Health Conditions

Health Conditions	(1) Mean	(2) Mean (2014 dollar)	(3) Median	(3) Median (2014 dollar)
Adult Asthma	\$14.08	\$33.34	\$94.70	\$224.18
Pregnant	\$4.97	\$11.77	\$33.42	\$79.12

Note: All estimates are in constant 1982-1984 dollar if not otherwise noted. Predictions based on true observations; counter-factual observations are not included. The means are calculated as in Equation(3.13).

To compare with the estimate of MWTP for air quality in the existing literature, Bayer et al. (2009) estimated that the median household would pay \$149 - \$185 (in constant 1982–1984 constant dollars) for a one-unit reduction in average ambient concentrations of particulate matter after accounting for migration. This paper suggests that those estimates would differ based on the whether if a household has an air quality-related health condition. In addition, it might be worth mentioning that apart from the changes in MWTP induced by the health condition, this paper does not provide an estimate of the overall marginal willingness to pay for air quality itself. This is because the objective of this paper does not require a exogenous measure of the air quality - the emphasis is on the effect induced by the health condition. As a result, the estimation of an overall MWTP for air quality using the observed AQI could be biased.

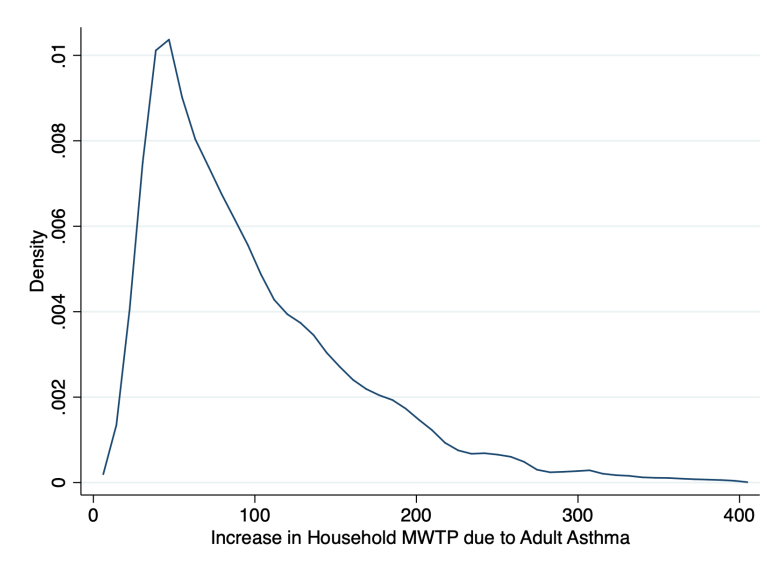


Figure 6.5: Increase in MWTP for Air Quality due to Adult Asthma (1982-1984 Dollars)

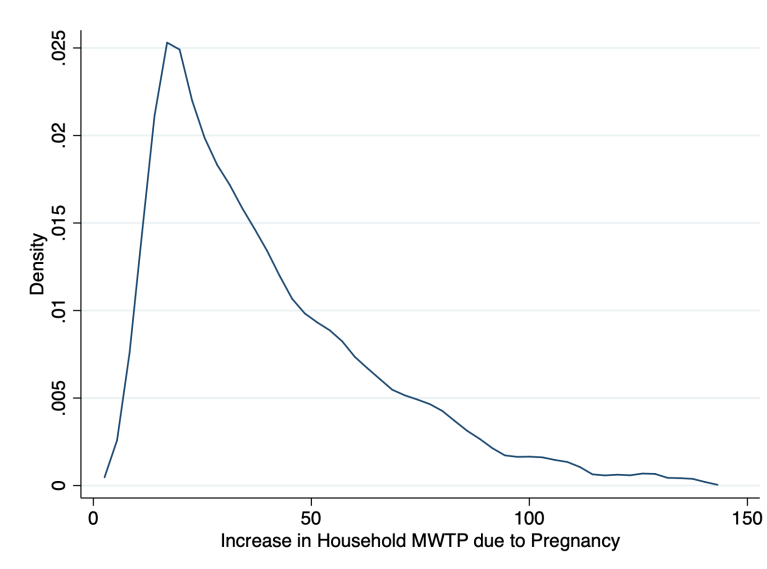


Figure 6.6: The Increase in MWTP for Air Quality due to Pregnancy (1982-1984 Dollars)

7 Conclusion and Discussion

This paper provides an answer for the question of whether people with air-quality related health conditions move to areas with better air quality by adopting a spatial sorting model. The results estimated by a mixed logit model does support the hypothesis that households move toward clean air after an adult member was diagnosed with asthma. In addition, the comparison of estimates using adults only health conditions and children only health conditions reveals that people tend to react more to first time asthma diagnosis of adults. This provide us with directions for further investigation. The results also suggest households tend to move to areas with less air pollution if the family is expecting a newborn, this could be a result of the household's precaution to avoid the damage of the air pollution on the infants. However, pregnancy is usually an important event for a household that might trigger a migration decision for reasons other than concerns over air quality. Further analysis is required to discover the true cause of such relocation.

The estimates from this paper could be used as a tool to gauge the effect of a change in the air quality on the change in the population, as well as the change in the population composition in terms of health, race education and other individual characteristics for a certain county. The next step of this paper is to provide an estimation of households' marginal willingness to pay for air quality as revealed in their locational choice. This estimate will be useful for policy makers in a cost-benefit analysis of air pollution. It will also be a great tool for private industry to decide how much compensation they should offer a potential worker to attract them to move to a city with higher levels of air pollution, based on the individual characteristics of the worker. In the near future, I would also like to explore the distributional effect of air quality on different populations, especially using data from developing counties, such as China or India.

Chapter 2

A Price Index from the Historic Huizhou Housing Market: 1912 - 1949

1 Introduction

The housing market is an essential component of the modern economy. But economists today know relatively very little about the housing market in the very long run relative to other aspects of the economy. Limited data has compromised studies on the housing market in economic history. This paper takes advantage of a unique individual-level transaction data set which covers an entire ancient prefecture Huizhou in southern China for about 400 years. This paper provides very long run hedonic price indices using around a thousand house and land transactions. Meanwhile, I analyze the specific price determinants in this housing market that are different from the modern housing markets.

Huizhou, a historical prefecture located in south Anhui Province in China, is a mountainous region with a distinct local culture. It is home to one of the most prominent groups of merchants in China that were active from the fourteenth to nineteenth centuries. Those merchants left hometown in their early adulthood and were mostly engaged in the salt and tea businesses. They sent most of their profits back home, and their families used a large amount of

that remittance to build houses in their hometown. Those houses were built carefully following a systematic guideline of Huizhou residential houses and have a similar structure and layout.

The data that this paper uses is a unique historic house transaction dataset from the Huizhou Documents that I collected from southern Anhui Province during 2017-2019. The Huizhou Documents are a collection of hand-written files that cover transaction records, contracts, inheritance documents, household ledgers, government decrees, litigation documents, lineage agreements, private letters, etc. Thanks to the mountainous terrain, Huizhou was able to keep a large number of historical records and artifacts safe from its prosperous past during wartime and the cultural revolution. Since the early 20th century, booksellers and collectors had been collecting those documents from Huizhou households over a long period. In the 1980s, researchers and local governments started to realize the uniqueness and potential research value of the Huizhou Documents and initiated a collective effort to purchase the documents from the private collectors and store them in museums and libraries. I collect copies of those documents from archive offices, libraries, and book collections in the Huizhou area. Those documents are mainly dated in the Ming Dynasty (1368 - 1644), Qing Dynasty (1644 - 1912), and the Republic of China (1912 - 1949). The sheer abundance of the records provides a rich mine that allows us to learn about details of rural economic life in late imperial China. Today, more than 500,000 documents had been discovered and the number is still growing.

Housing has always been an large component of household consumption. This paper develops price indices that capture an important component of an overall measure of inflation in a very long period in Chinese history. For this purpose, I focus primarily on the house and land transaction records. Those transaction records are contracts written by a third party that state information including the names of the seller(s) and buyer(s), the location of the house/land, a brief description of the condition of the house/land, room count, a rough or accurate measure of the size, orientation, price, etc. I currently have transcribed 1093 transaction documents covering the Ming Dynasty, the Qing Dynasty, and the Republic of China (RoC) era. I estimate two very long-run price indices using hedonic regressions. Meanwhile, the results show that the house and

land price had a general upward trend in the sample period that covers 1570 - 1949. The results also suggest that the distinct lineage culture in the Huizhou area affect the price-determination in the housing market.

This paper is structured as follows. In section 2, I introduce the previous literature that investigates the historic housing market in China and other parts of the world. In section 3, I introduce the data and how I deal with the information on the written records. I also discuss the main variables. In Section 4, I present summary statistics. In section 5, I introduce the empirical strategy and show the preliminary results. Section 6 concludes and discusses the next step.

2 Literature Review

This paper is one of the very first works that explores the historic Chinese housing market using micro-level data. In this section, I will introduce the previous work that has been done on the housing market in China before 1949. Due to the scarcity of such research, I will also present research that utilizes historical house and land transaction data from other countries.

2.1 Historical Housing Markets in China

The economic history literature on the historic housing market in the Republic of China era is minimal, and the research on the housing market in late imperial China is almost a blank field. The studies available are focused on major urban areas, such as Shanghai, which covers the RoC era. Zhu (2011) documents the land and house prices in Shanghai using information the author collected from newspapers and historical documents. The land price increased substantially during the Second Sino-Japanese War (1937-1945). The house and land prices increased by 133%-500% from 1938 to 1944 in different districts of Shanghai. Zhu (2011) shows that the growth rate of house and land price far surpassed the growth rate of overall commodity prices.

Other historical work discusses such facts anecdotally. For example, Li (2012) also collects historical anecdotes from historical newspapers and other historians' work about the

housing market in Shanghai and other cities. Those anecdotes reflect a turbulent housing market with prices that were closely tied to historical events. However, Huizhou prefecture, the area where my data is from, could have had different circumstances.

The literature suggests that in imperial China, people used the housing transaction records as proof of the property right ownership. There are also two types of transaction records: the "red" contract and the "white" contract. The red contracts are "red" with the government's red stamps on them, meaning the government certified those transactions. The white contracts recorded the sales done privately without being brought to the court. The white contracts cover most of my sample. The third party that signs the white contract was likely to be the most prestigious member of the lineage; the authority of lineages (McDermott, 2013) made the white records equally effective as the official red contract.

Historians have put in the effort to achieve a consensus on what the late imperial Chinese land system was like in detail. Zhang (2020) summarizes the key characteristics of the land system: (i) the majority of the landowners were smallholders; (ii) the security of private land ownership was quite robust; (iii) land use regulations were relatively laissez-faire comparing to modern laws; (iv) property transactions were done mainly over written contracts, and contracts were respected by both communal authorities (such as lineages) and formal authorities. Overall, historians agree that property rights were effectively protected and provided the economic incentives that are usually associated with private ownership.

2.2 Historical House Price Indices

There are several papers that use long term housing price data from other countries to construct price indices and explore the determinants of historical housing prices. Common data with which economists have created price indices include professional appraisals, self-assessments, and transactions; historical data, including what I collected from the Huizhou document, tend to be transaction-based.

One popular method of creating historical housing price indices is the repeat sale method.

Eichholtz (1997) uses the repeat sale method to construct a historical price index that covers more than 200 years based on the house transactions of buildings in Herengracht, the Netherlands. The Case-Shiller Home Price Indices¹ also use repeated sales to construct a long-term house price index from 1890 till today for major U.S. cities (Shiller, 2005). This approach requires observing the same house repeatedly in the sample period. Some economists use hedonic method to investigate the performance of housing market during historic events. White, Snowden and Fishback's (2014) discussion on different indexing methods to study the housing market during the Great Depression is an example. Nicholas and Scherbina (2010) also estimate hedonic housing price index for Manhattan using transaction information from 1920 to 1940. My data does not currently contain enough information on repeat sales; I thus use a standard hedonic model to calculate decade and year fixed effects of the prices to construct hedonic indices.

With hedonic regressions, this paper also aims to find out the historical determinants of house prices in Huizhou. Dröoes and Minne (2017) re-estimate the Herengracht index with additional data of the macro economy and find that population growth, construction costs, and new housing supply are the main contributors to housing prices before 1900, while income and interest rates started to have an impact after 1900. However, based on the literature, the formation of house prices in Huizhou might be a different story. As historian Sucheta Mazumdar (2001) argues, "rather than tools of a market-mediated exchange society, these (Huizhou) land contracts were stitches in the webs of reciprocity and redistribution peculiar to rural society." With the available information from the transaction records, as I will show in later sections, this paper traces out parts of the "webs of reciprocity and redistribution" in the rural Huizhou society. Different types of transactions and properties shown in the records allow me to investigate the preference of homeowners back in time.

¹ Available for download on Robert Shiller's website.

3 The Data

I gathered my house transaction data from historical documents discovered in the Huizhou prefecture. Figure 1 provides a map of the relative position of Huizhou in China. To my knowledge, this paper is the first to digitize and transcribe those documents. I started collecting the transaction records from the Archive office in Yi County, Huangshan City, in the summer of 2017. I subsequently visited the Huangshan City Archive Office, the Huangshan City Library, the Huangshan University Library, and the Anhui Normal University Library in the following summers. The transaction records I read in the archive offices are in their original parchment. The transaction records that I collected from the libraries are from a series of published photo albums of Huizhou documents photographed by Liu Boshan. I took pictures of the original parchments in the archive office and Liu's photo collection of Huizhou Documents as I could not carry them with me. After collecting the images of those documents, I transcribed important information on them into an excel spreadsheet, taking the form of a dataset with variables in text and a unique ID number for each transaction. I then took that information, translated them into English, and finally converted them into variables with continuous or discontinuous numeric values. I will discuss the variables later in this section.



Figure 3.1: Locations of Anhui and Huizhou

The contracts were usually composed by an educated acquaintance of the seller and the buyer, or an elderly in the lineage who often wrote contracts for lineage members. The contracts were usually kept by the buyers, and they serve similar purposes as property ownership certificates. Each document follows a certain template that is more or less constant throughout the Ming Dynasty, the Qing Dynasty, and the RoC era. Figure 2.1 is a sample of a land transaction document. A legal document typically states the name of sellers and buyers, location, the type of house layout, an indicator of the relationship between buyers and sellers (if they belong in the same lineage or household), currency, price, and date. Sometimes a document also includes more detailed information about the house, such as the orientation of the house, area of the house, room count, etc.

To make it possible to apply econometric analysis to the data, I convert the information from the documents into numeric values. Some information is straightforward, such as date and price; some can be converted to an indicator variable, such as the county/village in which the

house was located and the family relationship between buyers and sellers. Other variables, such as house layout types and currencies, need more careful interpretation.

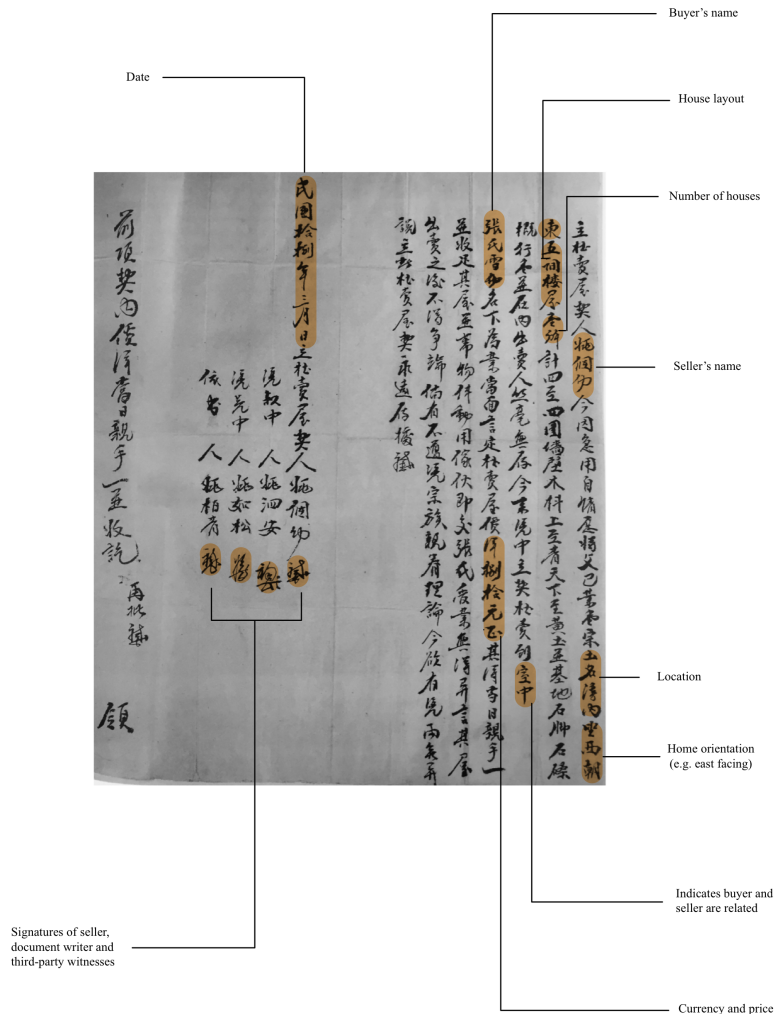


Figure 3.2: A Sample House Transaction Document

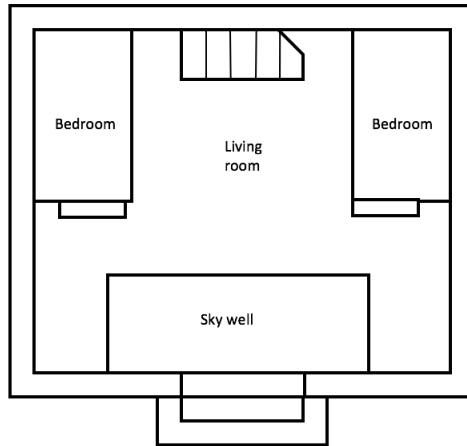


Figure 3.3: First Floor Floor Plan of a “Sanjian”

3.1 House Layout

The document writers usually stated the type of the house layout, the number of rooms being sold, and less frequently, the floor or lot size measurement. In the Huizhou area, craftsmen built residential houses according to a standard design (Liu, 2013); they all had a wooden main frame, brick walls, and black tile roofs. The wooden pillars run through the entire house vertically, dividing the space into several compartments in a symmetric fashion. Residents assign different functions to each compartment, such as a bedroom, living room, or storage room.

Some layout styles were the most popular in the Huizhou area as they suit the local climate and culture the best. Sanjian and sihe are two common ones. Sanjian refers to a symmetric layout that has one living room in the middle and two bedrooms on each side of the living room, it usually has a second floor, and sometimes a third floor. Figure 2.2 shows the typical floor plan for the first floor in a sanjian. Sihe refers to a symmetric layout that has one sky well at the center of the house; two living rooms, one in front of the sky well and one behind; four rooms at the four corners of the house (Figure 2.3). It has two or three floors. Apart from sanjian and sihe, there are other basic layouts, including wujian, big sanjian, etc. A large house that has multiple units (jin) is usually a combination of several units with one of the aforementioned basic layouts or a hybrid of multiple basic layouts.

I first generate an indicator variable for each layout type. I also use the approximated average number of compartments for each layout and construct a single compartment count variable for all house transactions. For creating the price index, I use the layout dummies, compartment count variable, and the interaction of them. I investigate architecture literature that conducts surveys of Huizhou houses. A two-story sanjian house usually has 4-5 bedrooms and 1-2 living rooms, or 6 compartments in total. A three-story sanjian usually has 6-8 bedrooms and 1-3 living rooms, or 9 compartments in total. A two-story sihe house usually has 8-10 bedrooms and 2-4 living rooms, or 12 compartments during total. A two-story wujian house usually has 8-9 bedrooms and 1-2 living rooms, or 10 compartments in total. Due to the flexibility in the function of each compartment, I use the total number of compartments as the room count variable. Another advantage of using the room count as opposed to using an indicator variable is that sometimes only a few rooms of the house were sold, and the room count variable is appropriate for houses sold either wholly or partially.

3.2 Currency

The currency systems in the Ming Dynasty and the Qing Dynasty were reasonably stable. The monetary system consists of two main currencies, silver and cash (Qian). China was not a producer of silver or silver coins; as a result, most of the silver in circulation in the Chinese domestic market were obtained from exporting goods to Europe and Japan. The imported silver was mainly mined in Latin America and Japan. Copper cash, or “Qian”, is a currency that the government minted with copper throughout the Ming and the Qing Dynasty. The Chinese government set the official exchange rate between the coins and silver. However, the de facto exchange rate tended to change with fluctuation in the accessibility of silver. Transactions in the sample include different types of currencies; thus, this paper also sheds light on the actual exchange rate between silver and cash.

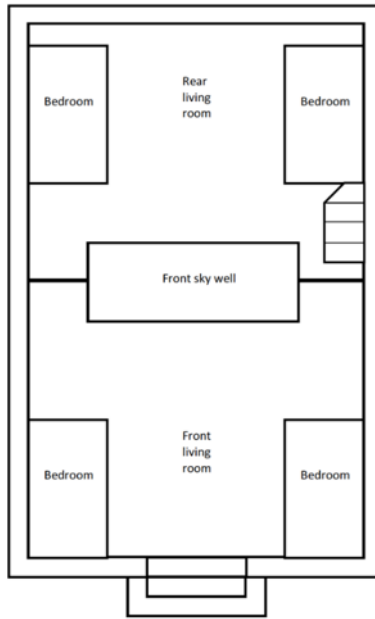


Figure 3.4: First Floor Floor Plan of a “Sihe”

The financial system in the RoC era had experienced turmoil, and the National Government of the Republic of China adopted many different currencies during this period (Peng, 2007). Changes in currency accompanied political turbulence, warfare, hyperinflation, and influences from the world economy. This situation makes it challenging to convert house prices between different currencies. Instead of converting all prices to a base currency, I create an indicator for each currency, then create another variable for the numeric price as recorded. Table 3.1² summarizes all of the currencies that were adopted in China during the RoC era. From 1912 to 1935, the National Government minted several versions of silver coins (Yinyuan). They have a different appearance and a slight difference in silver content; see Table 3.2.

Those different versions of silver coins, including ones minted by non-Chinese government entities such as Yingyang (the Mexican silver coin), had similar silver contents and were approximately interchangeable.

²All official exchange rate information in Table 3.1 is only valid when the currency was first adopted.

Table 3.1: Currencies Adopted in Huizhou During the Republic of China Era

Currency	Yinyuan (Silver coins)	Fabi (Legal tender)	Jinyuanquan (Gold Yuan)	Yinyuanquan (Silver Yuan)	Yingyang (Mexican silver coins)
Year Adopted	1912	Nov 4, 1935	August 19, 1948	July, 1949	1854 - after 1949
Exchange rate b/w currencies	Varies (see Table 2.2)	1 Fabi =1 Yinyuan =1 shilling & 2.5 pence	1 Jinyuanquan =3 million Fabi = US\$0.25	1 Yinyuanquan = 100 million Jinyuanquan	90% silver, 27.58g per yuan
Note	Silver Standard	Banknote gold standard (pegged to £sd)	Gold standard	Silver standard	Silver standard

Note: Data source: Peng, Weixin. The History of Chinese Currency (In Chinese). (1958).

Table 3.2: Different Versions of Silver Coins During the Republic of China Era

Versions	Yuan-big-head	Sun-small-head	Chuanyang
Year minted	1914 - 1928	1912 - 1914, 1928 - 1933	1933
Composition	89% silver, c. 23.9g per yuan	96% silver, c. 24.96g per yuan	88% silver, c. 23.5g per yuan

Note: Data source: Peng, Weixin. The History of Chinese Currency (In Chinese). (1958).

3.3 Other Variables

In this section, I introduce all of the other variables that I use for the construction of price indices, apart from the aforementioned currency.

ln(Price) is the log of the total transaction price written on each contract. It is the numeric value of the written price, regardless of the currency being used.

Type is the type of property being transacted. It could be a residential house or a commercial property. Hypothetically, commercial properties are associated with higher prices.

Layout is the layout of the house, as discussed in the data section. Hypothetically, being a "sihe" or "wujian" is associated with higher prices comparing to "sanjian" since they have larger sizes, and are most likely to be owned by well-off families.

Compartment is the number of compartments each transaction contains if the item is a house. It is the number of compartments as recorded in the documents, or as I infer from the structure of the house.

Face is the orientation of the house, the direction that the house's gate, or the main living room, is facing (the direction from a living room to the sky well is the direction that this living room is facing). Traditional Chinese "Fengshui" suggests that south-facing is the best orientation. However, there is local superstition suggesting that facing north is a safer choice for merchants to do well with their business. We will find out which orientation was preferred in the results section.

Female is a variable indicating if the seller of the house is female. The traditional Chinese society was a typical patriarchy. Huizhou was no exception - when a female family member is involved in the most important household decision-making, such as selling and buying properties, it usually implies that she is a widowed elderly in the family. The impact of the seller's gender on the price could go both ways. It is reasonable to assume that when the seller's household head passed away, the buyer would be willing to pay a higher price to help the seller's family out. It is also possible that the seller sells the house at a lower price if the situation is urgent, or that female sellers have less market power. We will look at the results to understand more about this price determinant.

SameLineage and **SameFamily** indicate whether the seller(s) and buyer(s) are from the same lineage or the same family. Lineages play an important role in organizing communal life in late imperial China; it is worth exploring the effect of such a connection on housing prices.

Urgent is a dummy variable that has the value of 1 if the seller of the property is in a financially urgent situation. The most commonly seen case is that when a parent of the seller passed away, the seller sold the house to collect enough cash to pay for a proper funeral.

Currency indicates the currency this transaction used. This variable is discussed in detail in the section above.

4 Summary Statistics

I obtained records of 1031 property transactions from 1570 to 1949. Figure 4.1 graphs the total number of transactions by year. We see most transactions dated around the late Qing Dynasty and the early RoC period. The distribution of the transactions across time could be due to the local economic activity and the relative difficulty in preserving earlier documents.

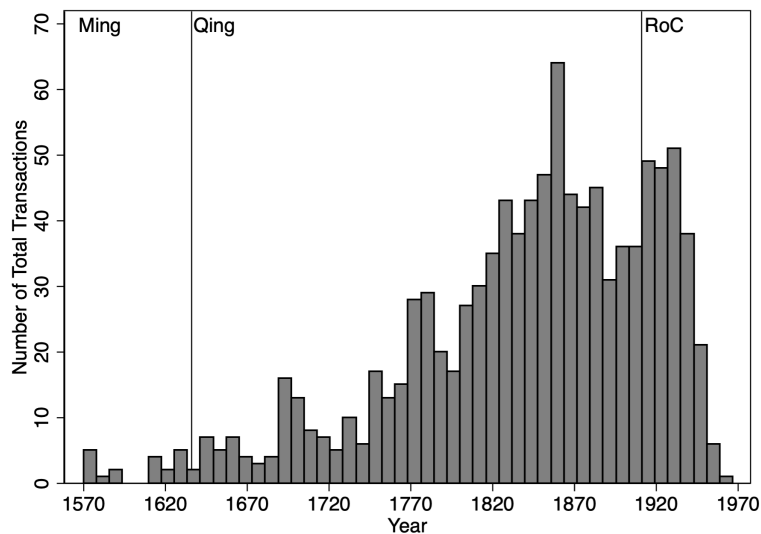


Figure 4.1: Number of Transactions by Year

I also show the distribution of the number of house transactions and land transactions across time in Figure 4.2 and Figure 4.3. They both show a similar peak at around the late Qing Dynasty.

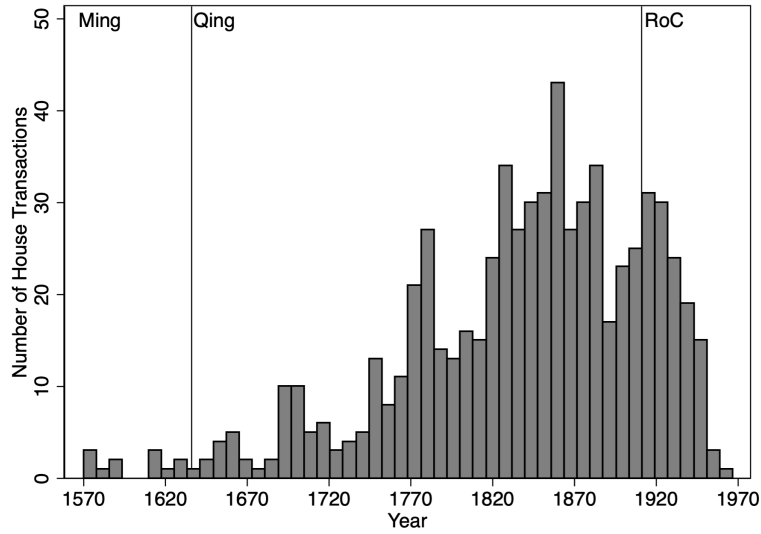


Figure 4.2: Number of House Transactions by Year

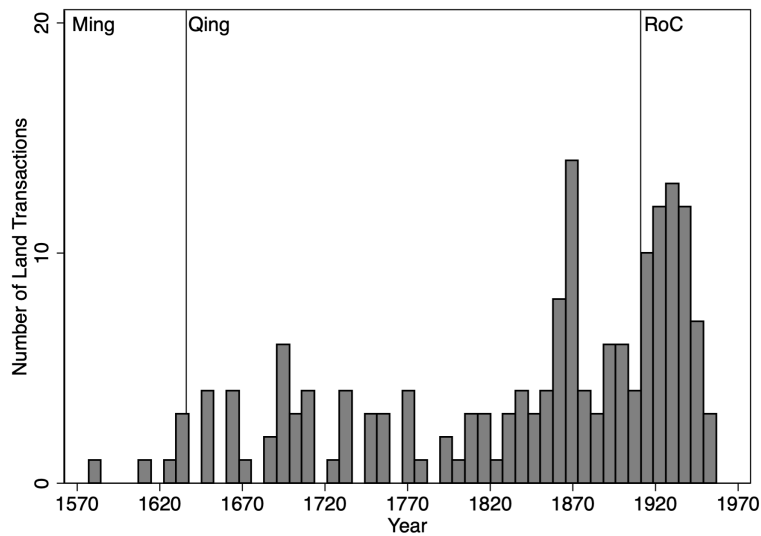


Figure 4.3: Number of Land Transactions by Year

Transactions by item. Transactions in the sample include transactions of residential houses (or part of one), commercial properties (or part of one), and land for residential houses (or part of one). In a typical Huizhou residential house, kitchens and toilets were independent units detached from the main building. As a result, there are transactions of only kitchens/toilets. Other

items include buildings for agricultural production purposes, such as oil mills and pigsties. Table 4.1 shows the numbers of transactions by item.

Table 4.1: Number of Transactions by Item

Item	Frequency	Percentage
Residential Houses	634	61.49%
Residential Land	172	16.68%
Commercial Houses	45	4.36%
Toilets or Kitchens	105	10.18%
Agricultural Production Sites	51	4.95%
Other	24	2.33%
Total	1031	100.00%

Transactions by house layout. Table 4.2 shows the number of transactions for each residential house layout. A detailed description of the layouts is in section 3. As we can see, majority of the houses have the standard “sanjian” or “sihe” layouts.

Table 4.2: Number of House Transactions by Layout

Item	Frequency	Percentage
Sanjian	395	58.26%
Sihe	248	36.58%
Wujian	10	1.47%
Other	25	3.69%
Total	678	100.00%

Transactions by county. There were six counties within the Huizhou region: She (where the prefecture government is located), Yixian, Xiuning, Qimen, Jixi, and Wuyuan. There are also a few transactions that occurred outside of Huizhou; these documents were brought back home by the Huizhou merchants who were doing business in other cities, such as Wuhan. Table 4.3 shows the number of transactions by county. A majority of the current sample is from She county and Qimen county. She is the largest county in Huizhou; Qimen was home to many prominent Huizhou merchants’ families and had a reputation for being affluent in history. In future work, I will visit Jixi and collect more transaction records from there.

Table 4.3: Number of Residential House Transaction by County

County	Frequency	Percentage
Yi	134	13.00%
She	499	48.40%
Qimen	232	22.50%
Xiuning	64	6.21%
Wuyuan	70	6.79%
Jixi	5	0.78%
Outside of Huizhou	16	1.55%
Total	1031	100.00%

Average transaction price by year. In figure 4.4, I plot the annual average transaction price of sanjian, the most common layout, for those used silver, the most common currency. We can see a slight increasing trend throughout the period.

Table 4.4 summarizes all of the variables that describe the characteristics of a property. Apart from log Price and the number of rooms, all variables are binary. The upper panel shows the summary statistics for the entire sample, including both house and land transactions. The lower panel shows the summary statistics for the subsample of house transactions only. "Number of rooms" is a feature unique to house transactions; thus, it only shows up in the lower panel.

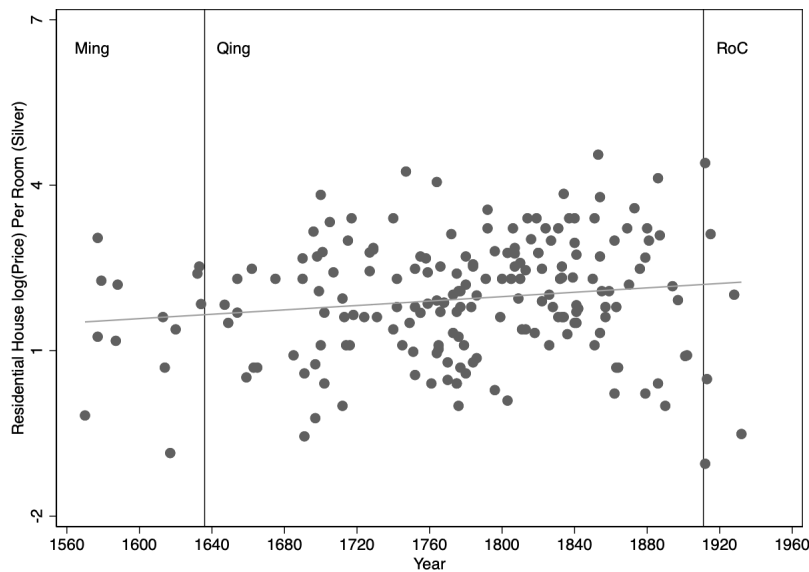


Figure 4.4: The Average Room Price (log) of Residential Houses (Silver)

Table 4.4: Summary Statistics of Variables

Variable	Mean	Std.Dev.	Min	Max
Sample: All Transactions; N=1031				
log(Price)	4.228	3.176	-1.347	16.81
Silver	0.351	0.478	0	1
Female Seller	0.113	0.317	0	1
Same Lineage	0.438	0.496	0	1
Same Family	0.067	0.250	0	1
Seller in Urgent Situation	0.029	0.168	0	1
Next to Ancestral Halls	0.045	0.207	0	1
Subsample: Residential and Commercial Houses; N=679				
log(Price)	4.650	3.191	-1.204	16.81
Silver	0.333	0.472	0	1
Number of Rooms	3.362	4.778	0	72
Female Seller	0.113	0.317	0	1
Same Lineage	0.436	0.496	0	1
Same Family	0.077	0.266	0	1
Seller in Urgent Situation	0.034	0.181	0	1
Next to Ancestral Halls	0.056	0.230	0	1
Abandoned	0.021	0.142	0	1
New	0.038	0.192	0	1

5 Results

5.1 Empirical Strategy

To estimate the house price index, I run the regression shown in equation (5.1) for house transactions. Each i is one property. X_i is a vector of all of the characteristics associated to house transaction i . T_i is a vector of time dummies, and its coefficient γ_1 captures time fixed effects. I will use decade or year for T_i in separate regressions. $County_i$ is a vector of county dummies, and its coefficient γ_2 captures county fixed effects. Note that since multiple currencies showed up in the sample and the relative values among them changed over time, I also include time by currency dummies in X_i for all of the regressions.

$$\ln(\text{Price}_{it}) = \beta X_{it} + \gamma_1 T_i + \gamma_2 \text{county}_i + \varepsilon_{it} \quad (5.1)$$

Note that γ_1 , the time (decade or year) fixed effects, are the basis for constructing the price index. Equation 5.1 is most useful for estimating overall measures of a large component of price inflation that capture aggregate trends. To further explore the effect of each property characteristic X_i on housing prices, as a relationship of independent interest, I also consider conditioning more flexibly on time and space by including an additional county-by-decade interaction term; γ_3 catches the county-by-time fixed effects. I then show the coefficients on β .

$$\ln(\text{Price}_{it}) = \beta X_{it} + \gamma_1 T_i + \gamma_2 \text{county}_i + \gamma_3 T_i * \text{county}_i + \varepsilon_{it} \quad (5.2)$$

I run a similar regression, as shown in equation (5.3) to obtain land price index. Y_j is a vector of residential land characteristics; η_1 and η_2 are respectively the time fixed effects and the county fixed effects.

$$\ln(\text{Price}_{jt}) = \alpha Y_{jt} + \eta_1 T_j + \eta_2 \text{county}_j + \eta_3 T_j * \text{county}_j + \xi_{jt} \quad (5.3)$$

η_1 , η_2 , and η_3 are respectively the time fixed effect, county fixed effects, and county by time fixed effects. The estimation results will be shown and discussed in Section 6.

I also run the regression for house transactions that only used silver and copper cash (qian) with an additional currency by decade interaction term to estimate the exchange rate between the two currencies. More details are discussed in the appendix.

5.2 Price Indices

Figure 5.1 plots the decade fixed effects directly from the estimation of γ_1 in equation 5.1 using only house transactions. The reference decade is the 1570s; the price index shows a stable development before the nineteenth century, and a general increasing trend after 1800. A slight decreasing trend in the housing price index from around the 1840s to the 1860s coincides with the First and the Second Opium War, which severely shattered the foundation of the Qing Dynasty. The surging price during the late RoC reflects the hyperinflation at the time. The estimation results of housing price for each county is shown in Figure 6.1 in the Appendix.

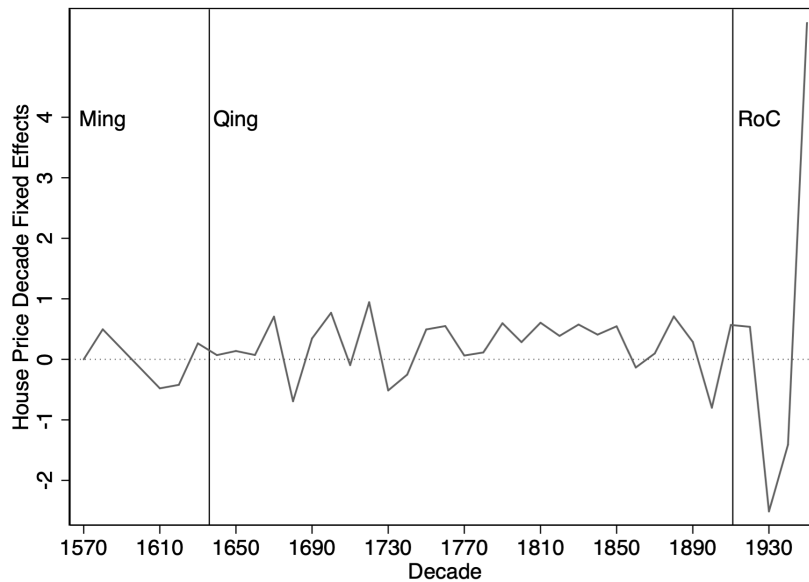


Figure 5.1: The Decade FE of (log) House Price

Figure 5.2 plots decade fixed effects directly from the estimation of η_1 in equation 5.3 using only land transactions. Comparing to the house price index, it shows a more obvious and steady upward trend since the base decade of the 1570s, but a drop in price during the RoC era.

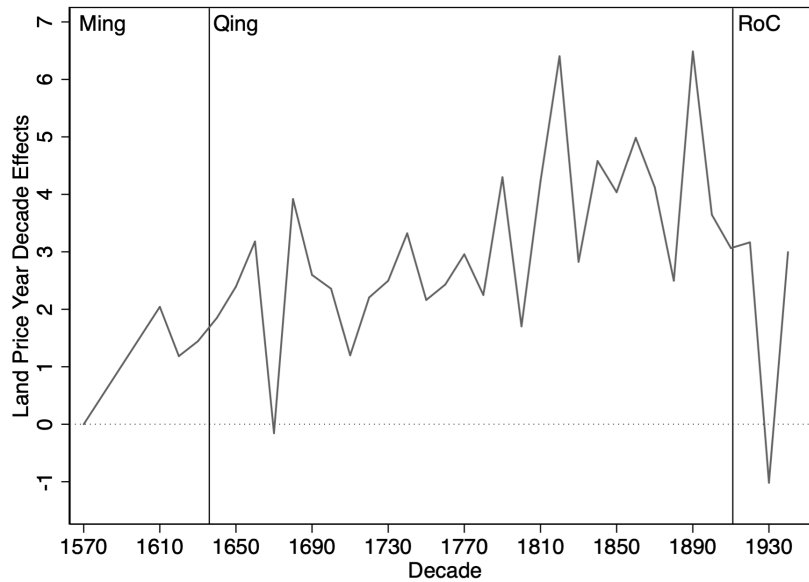


Figure 5.2: The Decade FE of (log) Residential Land Price

Figures 5.3 and 5.4 plot the year fixed effects from the estimations of equations 5.1 and 5.3 for house and land transactions, respectively, using year dummies instead of decade dummies. Each dot represents the year fixed effect of a certain year. The dashed lines in dark gray are the local polynomial smoothing of the year fixed effects at degree 1 with a 10-year bandwidth. The lighter gray area is the 95% confidence interval. Both smoothed indices were weighted by the frequency of observations in each year. We observe a general upward trends in both graphs while the land price has a larger variation. Starting around the 1840s when the Opium War broke out, the house price index started falling until the 1910s. From the 1770s to the 1890s, there are a few outlying years with exceptionally high price index. They are driven by a few very expensive sales during this period. Those transactions might reflect the rich Huizhou merchants purchasing luxurious properties, as this time frame coincides with the years when this merchant group was the most wealthy and powerful.

During the Republic of China era (1911-1949), both house and land prices experienced dramatic increase which again reflect the hyperinflation; this is not captured by the decade fixed effects of land price. The detailed price indices generated by equation 5.1 and 5.3, estimated by

local polynomial smoothing, is shown in Table 6.1 and Table 6.2 in the appendix.

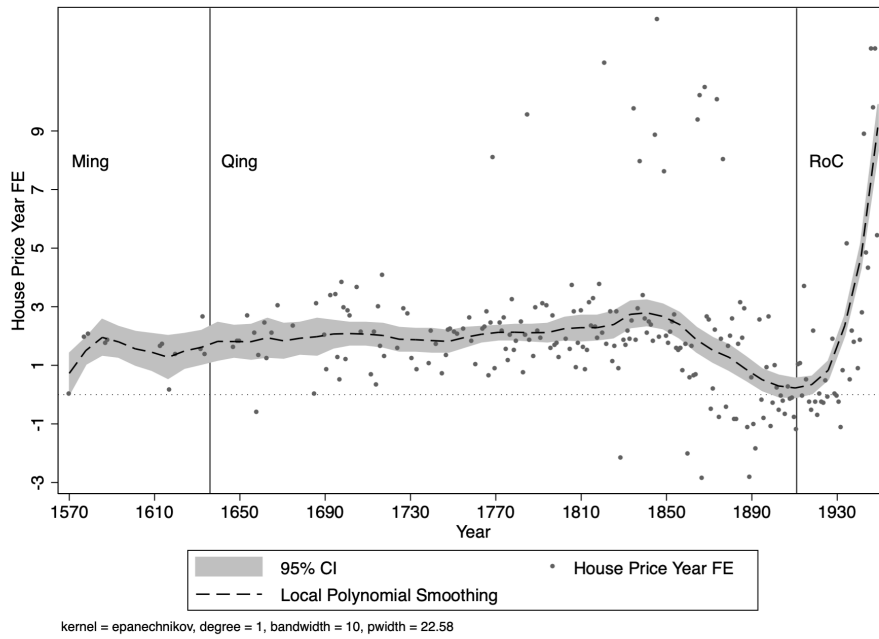


Figure 5.3: The Year FE of (log) Houses Price with Local Polynomial Smoothing

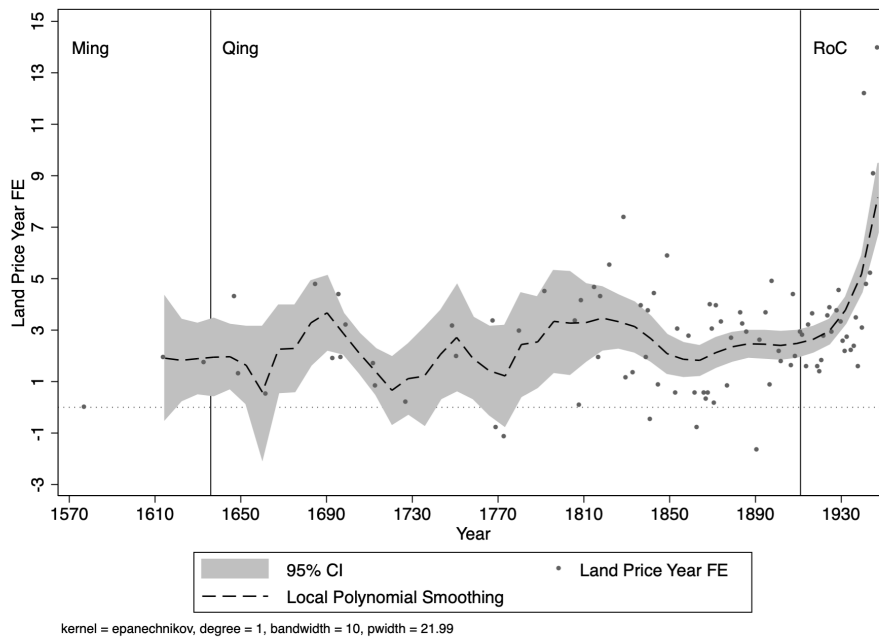


Figure 5.4: The Year FE of (log) Land Price with Local Polynomial Smoothing

5.3 The Price Determinants

In addition to the estimation of the historical price indices, the regressions tell us what was important for house and land buyers back in time. Table 5.1 shows the estimation results of equation 5.2 and 5.4. Panel (1) shows the result from the subsample with residential house transactions; panel (2) shows results from land transactions. The first column in panel (1) and the first column in panel (2) use decade as the basic time variable; the second column in panel (1) and the second column in panel (2) use year as the basic time variable. Currency \times Decade FE is included across all columns as the current sample size cannot afford Currency \times Year FE.

The positive and significant coefficient on the interaction term of “sihe” or “sanjian” with “number of rooms” suggests that these two are more desired layouts; note that the base category of the layout dummy here was wujian or other layouts. A room in a “sihe” or a “sanjian” house would, on average, cost more than a room in houses with other layouts. The “sihe” layout shows larger coefficients than “sanjian”, indicating higher prices. The reason could be that “Sihe” homes tend to be larger, more stable and are associated with better house quality or better interior designs. A female seller leads to lower transaction prices in both house transaction subsample and land transaction subsample, but the coefficients are not significant. The seller and the buyer being from the same lineage or family does not have a significant impact on the house price. However, if we use year fixed effect instead of decade fixed effect, transactions done within the same lineage tend to have lower prices, and transactions done within the same family tend to have higher prices.

Several other factors have significant effects on the house and land prices. If the sold property is located near an ancestral hall, it significantly increases the transaction price. A south-facing house is significantly more valuable than houses that face other directions, reflecting the willingness to pay for more sunlight or other Fengshui benefits that are associated with south-facing of the house. An east-facing house is also significantly more valuable than houses facing other directions except for the south.

The coefficient of variable “south” suggests that a south-facing house is about 2.23 times more expensive than a north facing house. Traditionally, Chinese people have always been

building houses that face south. North-facing houses are the absolute minority. When people build a north-facing house, they usually face other restrictions that prevent them from making the house south-facing, such as having to build the house on a small plot blocked by a tall building to the south. A north-facing property could also be a small north-facing part of a larger south-facing house. As a result, a north-facing property might be less desirable on other aspects other than the orientation, which brings the price further down.

An interesting finding here, in addition to the results discussed above, is that when the sellers were in situations where financial support was urgently needed, they tend to be able to sell their properties at a significantly higher price. Those transaction documents describe background stories such as the death of a parent, in which situation the children needed to sell the properties to pay for the funeral. Other reasons include when the sellers were struggling to maintain a business. The positively significant coefficients on this variable for both house and land transactions suggest a close social network in the Huizhou area that was intensified upon familial ties, and that lineages and families could provide financial assistance for their members at difficult times.

I also run the regression for house transactions that only used silver and copper cash (qian) with an additional currency by decade interaction term, as shown in Equation, and estimate the exchange rate between the two currencies. The detailed exchange rates by decade are shown in Table 6.3 in the Appendix.

Table 5.1: Hedonic Regression of House and Land Transaction

	(1)		(2)	
	log(House Price)		log(Land Price)	
House Characteristics				
Sanjian	0.960 (1.45)	0.115 (1.17)		
Sanjian × Rooms	0.054* (2.06)	0.071* (2.19)		
Sihe	0.951 (1.43)	-0.005 (-0.01)		
Sihe × Rooms	0.081*** (5.41)	0.116* (2.36)		
Other × Rooms	0.323 (1.39)	0.073 (0.23)		
Seller/Buyer Characteristics				
Female Seller	-0.185 (-0.92)	-0.097 (-0.34)	-0.563 (-0.92)	0.033 (0.04)
Same Lineage	0.045 (0.34)	-0.135 (-0.79)	0.355 (1.05)	-0.169 (-0.29)
Same Family	0.056 (0.24)	0.431 (1.18)	0.431 (0.49)	-2.370 (-1.15)
Seller in Urgent Situation	1.327*** (4.08)	0.451*** (3.80)	0.270 (0.20)	3.113 (1.32)
Transaction Characteristics				
Next to Ancestral Hall	0.732** (2.86)	0.682* (1.80)	0.570 (0.55)	0.862 (0.61)
Next to Temple	-0.658 (-1.32)	-1.342* (-1.88)		
Old	0.290 (0.74)	0.141 (0.25)		
New	0.248 (0.83)	0.037 (0.09)		

Continued on next page.

Table 5.1 Cont'd: Hedonic Regression of House and Land Transaction

House Orientation				
East	0.752*	0.951*		
	(2.13)	(1.69)		
West	-0.328	0.487		
	(-0.80)	(0.80)		
South	0.831*	0.969*		
	(2.40)	(1.96)		
Transaction Methods	✓	✓	✓	✓
Currency	✓	✓	✓	✓
County FE	✓	✓	✓	✓
Decade FE	✓		✓	
Year FE		✓		✓
Currency × Decade FE	✓	✓	✓	✓
County × Decade FE	✓		✓	
County × Year FE		✓		✓
<i>N</i>	678	678	172	172

t statistics in parentheses

* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$

Note: This table presents the estimation results of equation 5.2 and 5.4. Panel (1) shows the result from the subsample with residential house transactions; panel (2) shows results from land transactions. The first column in panel (1) and the first column in panel (2) use decade as the basic time variable; the second column in panel (1) and the second column in panel (2) use year as the basic time variable. Currency × Decade FE is included across all columns as the current sample size cannot afford Currency × Year FE.

6 Conclusion and Future Work

The transaction records from the Huizhou Documents are a gold mine for economic history studies. They provide detailed information not only on the transaction itself, but also the social background of the sellers and buys. The homogeneity of Huizhou houses provides us with convenience on deciding house price per unit.

The immediate next step of this paper is to continue the transcription and extend the dataset. Upon obtaining a larger sample size, I plan to explore the determinants of house price in historic Huizhou housing market with more rigorous research designs that a larger sample could support. By exploring different types of transaction that include renting, leasing and selling, this paper will also shed light on the long term discount rate in this housing market, as well as the

exchange rates among different currencies used in last imperial China. Constructing a price index using those records is the very first step of learning about the Huizhou historic housing market. By combining other long-run data from the area, such as commercial activity, education, and occupation, I expect more fruitful studies.

7 Appendix

The following graph shows the house price index estimated for each county in each decade. The value plotted is $\gamma_1 + \gamma_3$, the sum of the decade fixed effects and the decade by county fixed effects, in Equation 5.2. Yi County is the reference county. Due to limited amount of observation in some decades, the estimation is missing for some decades and counties.

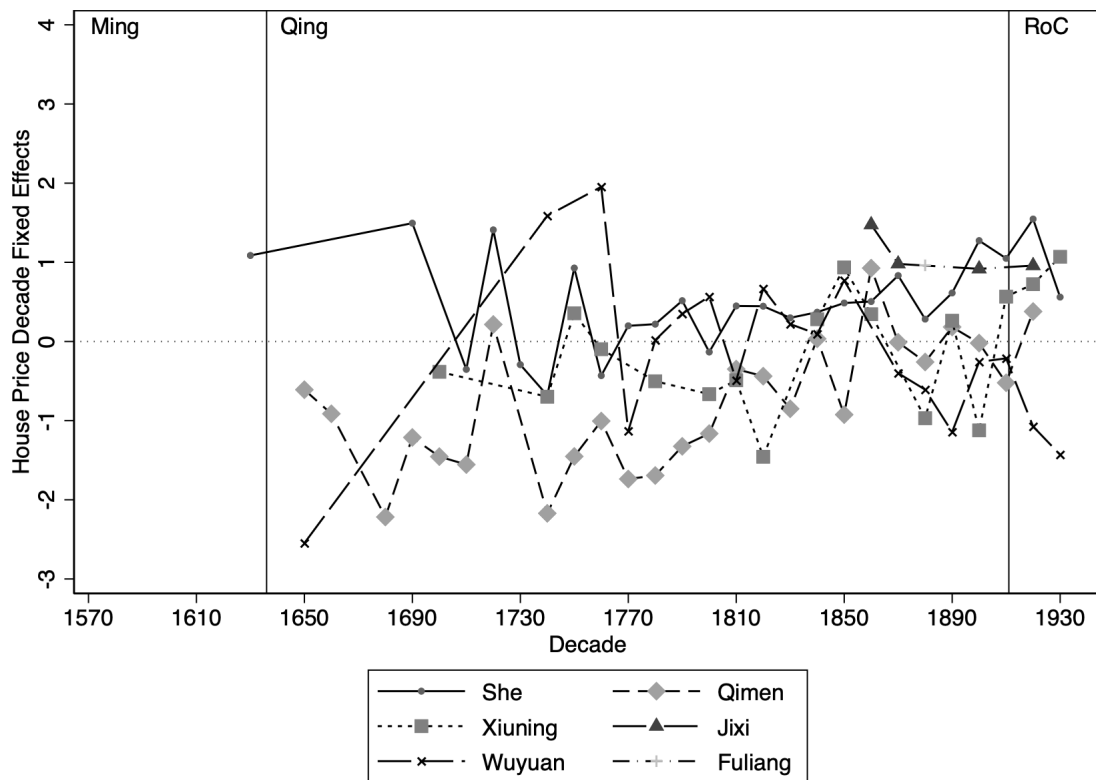


Figure 7.1: The Decade FE of House Price (log) by County

Table 7.1: The Historic Huizhou House and Land Price Indices (log)

Year	House Price	Land Price	Year	House Price	Land Price
1570	0.000	0.000	1770	1.588	2.056
1580	1.654		1780	1.579	2.574
1590	2.091	0.800	1790	1.749	2.479
1600	1.534		1800	1.929	3.027
1610	1.352	2.440	1810	2.059	3.211
1620	1.131	1.706	1820	2.129	3.461
1630	1.494	2.154	1830	2.213	3.623
1640	1.706	2.325	1840	2.093	3.453
1650	1.698	2.354	1850	2.090	3.259
1660	1.571	2.133	1860	2.181	3.301
1670	1.390	2.248	1870	2.160	3.152
1680	1.352	2.648	1880	2.130	2.937
1690	1.444	2.738	1890	2.139	2.744
1700	1.567	2.681	1900	2.239	2.849
1710	1.718	1.994	1910	2.373	3.084
1720	1.815	1.737	1920	2.394	3.498
1730	1.789	1.923	1930	2.328	4.192
1740	1.737	2.611	1940	3.635	5.489
1750	1.648	2.832	1949	5.920	8.372
1760	1.616	2.595			

Note: The numbers presented in this table are the numeric values of the local polynomial smoothing in Figure 5.3 and 5.4.

Table 7.2: The Historic Huizhou House and Land Price Indices

Year	House Price	Land Price	Year	House Price	Land Price
1570	1.000	1.000			
1580	5.229		1770	4.895	7.816
1590	8.093	2.226	1780	4.848	13.119
1600	4.636	1.000	1790	5.751	11.934
1610	3.864	11.478	1800	6.885	20.631
1620	3.097	5.508	1810	7.836	24.812
1630	4.453	8.617	1820	8.410	31.835
1640	5.509	10.227	1830	9.139	37.463
1650	5.464	10.525	1840	8.106	31.596
1660	4.813	8.438	1850	8.083	26.013
1670	4.016	9.465	1860	8.851	27.129
1680	3.866	14.130	1870	8.674	23.372
1690	4.237	15.456	1880	8.417	18.859
1700	4.792	14.600	1890	8.494	15.543
1710	5.574	7.342	1900	9.380	17.263
1720	6.143	5.678	1910	10.733	21.849
1730	5.986	6.844	1920	10.954	33.043
1740	5.681	13.612	1930	10.256	66.135
1750	5.195	16.985	1940	37.918	241.940
1760	5.035	13.400	1949	372.468	4324.819

Note: The numbers presented in this table are the natural exponents of those in Table 7.1.

Table 7.3: Estimated Exchange Rate between Silver and Copper Cash (Qian)

Year	50g Silver = ___ Qian
1750	1.00
1760	24.57
1770	170.01
1780	225.55
1790	494.99
1800	237.93
1810	673.13
1820	906.94
1830	398.29
1840	314.41
1850	533.94
1860	1208.88
1870	834.17
1880	379.17
1890	7555.86
1900	4608.01
1910	1174.51
1920	301.18
1930	2621.19

Table 6.3 shows the estimation results of the following equation:

$$\ln(\text{Price}_{jt}) = \alpha Y_{jt} + \theta_1 T_j + \theta_2 \text{cash}_j + \theta_3 T_j * \text{cash}_j + \xi_{jt} \quad (7.1)$$

The exchange ratio between cash and silver is normalized to be 1 in the 1570s; i.e. 1 *liang* (50 grams) silver could be traded for 1 cash. The numbers shown in Table 6.3 are $\exp(\theta_3)$. They show the exchange rate in each decades relative to the exchange rate in 1570s.

Chapter 3

Historic Shocks and Lineage Performance

1 Introduction

In late imperial China, lineages are groups that organize people based on patrilineal descendants. Lineages were the center of social life for most people thanks to Confucianism's prominence and its emphasis on ancestor veneration. They provided basic religious, political, and economic services as an informal institution on the grass-root level. In a society with a historically low internal migration rate, many Chinese people remain living today in areas where their ancestors lived. This raises the significance of lineage in the society and possibly makes the effect of lineage more long-lasting comparing to countries with higher migration rate. The hypothesis that this paper tests is that different cultures of different lineages have immediate and long-lasting impacts on their people.

The role of lineages in history has drawn the attention of anthropologists, historians, and economists. There are two main aspects of the commonly discussed research questions. First, how long exogenous historical events' impacts persist for later generations within a lineage; second, how lineages, as organizations, change the effect of such historical shocks. Particularly for economists, some attempt to discover some of the "missing variables" in a conventional regression equation that explains people's education and labor market outcomes. This paper

contributes to understanding lineages using several unique datasets I construct based on ancient house and land transaction records, government records, and genealogies of lineages from the historic Huizhou prefecture in China. It aims to show the very long-run effect of family on the education and labor market outcomes of later generations, using historical shocks as natural experiments. Historical shocks are not only used natural experiments to test the impact of lineage characteristics but also used to answer the question "how long does a historical shock's influence persist."

Lineages may vary in many ways. A common lineage is a group of families living in one or a few adjacent villages with the same ancestors. In contrast, a larger lineage could spread across the entire lower reaches of the Yangtze River. Most lineages had male leadership, while a small portion had female leadership. Lineages also differ in main occupation, connection to the government, and total wealth. Using several unique datasets I constructed based on ancient house and land transaction records, government records, and genealogies of lineages from China, this paper contributes to understanding the role of lineage with individual-level information covering almost 400 years.

This paper currently does not use lineage- and individual-level data as the collecting process is yet to complete. For the first step, I use the housing information collected for the second chapter and show the effect of historical shocks, including natural disasters, political turmoil, public construction, famine, and industrial development, on housing prices. Note that in the ideal setup of this research question requires grouping the house transactions by lineage to calculate a lineage wealth as one of the lineage characteristics. The preliminary results show that in the first ten years following a shock, some natural disasters decrease house prices; national and local political turbulence has heterogeneous effects on house prices; famines drastically decrease house prices.

This paper is structured as follows. Section 2 introduces the literature on the economic impact of lineages both in the short and the long run, as well as studies that use historical genealogies as the primary data source. Section 3 presents the summary statistics of shocks and

the outcome variables. At the current stage, the data collection of Huizhou lineages' genealogies is still ongoing. As a result, the summary statistics of the education level and employment information on the individual level will be missing for now. Section 4 discusses the empirical strategies. Section 5 presents the preliminary results using house transaction price as the outcome variable. Due to the limitation of data at the current stage, I am not yet conducting analysis involving lineage- and individual-level information. Section 6 and 7 discusses the results and future work.

2 Literature Review

2.1 Historical Data on Families and Lineages

Historical sources telling the story of how people lived back in time have great value for eras before census data is available. A commonly used source is lineage genealogy, a register of members of a lineage across generations; a lineage was justified by a genealogy (Watson, 1982). Family is arguably the most basic and important unit in social life; in imperial China, lineage, a group of families that are related by common patrilineal descent, serves more complicated functions. It could be widespread in scale: some lineages spread across the whole country and played an important role in the Chinese economy. Although not as comprehensive as census data, genealogies still cover the majority of the population and research using lineage data demonstrated its ability to answer important questions. Shiue (2018) uses multi-generational data from 1298 to 1925 collected from genealogies of seven large lineages in Tongcheng City, Anhui and provides a picture of the long run local mobility in terms of education. This paper will use genealogies in a similar fashion.

2.2 The Importance of Lineages in Historical Studies

There are some influential economic history studies using Chinese data that investigate path dependence using region as the basic unit. For example, Chen et al. (2019) show that the prefectures with better results in the ancient Chinese imperial examination (the Keju exam) have better human capital outcomes that persists till today. They use each prefecture's proximity to areas with a dense population of pine and bamboo, which were the main materials for making paper and ink, as an instrument for the grades of the Keju exam.

There are also path dependence studies using family as the basic unit with both Chinese and non-Chinese data. Teso (2017) uses historical data from sub-Saharan countries and shows that women with ancestors that were more affected by the transatlantic slave trade, during which the male population decreased and women were forced to take on men's jobs, are more likely to participate in the labor force, have lower levels of fertility, and are more likely to participate in household decisions today. Bleakley and Ferrie (2016) tracked down three generations from the participants of Georgia's Cherokee Land Lottery of 1832, in which the winners unconditionally received a large amount of increase in wealth. They do not find evidence suggesting that higher wealth improve their children's or grandchildren's educational and financial outcomes.

Ever since the dominance of the Confucius ideology among the ruling class, the Chinese society had been modeled as a big family by both scholars and the Chinese government, and the family relationship predominates all other kinds of relationships (Freedman, 1961). The importance of the family-like structures in political and economic life justifies the idea of framing family/lineage as an institution. Li and Sato (2008) show that the descendants of families with the landlord classification have better educational outcomes today comparing to descendants of families with the poor lower-middle peasant classification¹, even with the destructive shock during the Great Leap Forward and the Cultural Revolution. They suggest that family contributes to the performance of its members as a cultural institution in the long run.

¹A state designated social class system used by the Chinese government from 1949 to 1979. During the Great Leap Forward and the Cultural Revolution, the landlord class was not able to receive proper education and was suffered from struggle sessions, a form of publicly organized humiliation.

It might appear that studying the economic history using lineage as the basic unit is easily substitutable with using data on the individual/family level, as a lineage is a group of people and families. However, studying history from the perspective of lineages would provide different insights. On the one hand, families in a lineage tend to have heterogeneous characteristics; a family from the slave farmer class and a family from the landlord class could belong to the same lineage. The importance of lineage as a way of organizing those families together should not be ignored. When facing a shock, learning about the interaction between the effect of lineage attributes and family attributes is not possible without lineage level information. On the other hand, families come and go within a lineage, some of them stopped existing after major natural disasters. Looking at the lineage level is more suitable in the context of a study that covers a very long period.

The effect of lineage' strength on the economic activity is usually thought to be positive. The more powerful a lineage is, the more resource it can provide to its members, thus lead to better achievements (McDermott, 2015). However, the strength of lineage might become an impediment when we came closer to the 20th century. As Bertrand Russell argued in his 1922 book *the Problem of China*:

Filial piety, and the strength of the family generally, are perhaps the weakest point in Confucian ethics, the only point where the system departs seriously from common sense. Family feeling has militated against public spirit, and the authority of the old has increased the tyranny of ancient custom. In the present day, when China is confronted with problems requiring a radically new outlook, these features of the Confucian system have made it a barrier to necessary reconstruction.

The mixed theory suggests different mechanisms that lineage might interfere with their members' response to shocks, and this paper will use data to illustrate a clearer picture.

3 Data Description

3.1 Historical Shocks

The shocks in the paper will be the major historical events that I collected from the local gazetteers and chorographies. I collected chorographies from the Huangshan City Archive Office and gazetteers from the Huangshan University Library. The chorographies are local history records created and maintained by the county-, city-, and province-level government.

Administrations at each level were also responsible for updating their local chorographies periodically. I specifically collected county-level chorographies for this paper as they contain the most detailed information. Gazetteers contain the names of administrative entities and the changes of administration boundaries over time. I used gazetteers as complementary information to confirm the location of each event if the records in chorographies raise confusion.

After reading and collecting the chorographies, I concluded several types of major events that tended to have universal impact on the local residents and use them as shocks. More specifically, I recorded historical events that were: natural disasters, such as earthquakes, floods and wildfires; construction of large public goods, such as roads, bridges and schools; political turmoils, such as wars, protests, riots, change in emperors/dynasties; public constructions, such as roads and dams; famines; and the founding of private businesses. The detailed definitions of each category are listed in Table 1. Using the land transaction documents, potentially combined with the official land registers, we can trace out an estimation of lineages' wealth over a long time period. Using the genealogies, we can observe the education and labor market outcomes for lineages members in the Huizhou area, as well as some characteristics of the lineage, such as size, existing years, the highest official rank that its member ever achieved, the gender of the lineage leaders, etc.

Figure 1 shows the distribution of recorded shocks across decades. It is obvious that there are more shocks during the Republic of China (RoC) era. This could be the result of a more complete records in nearer history, or the result of increased social conflicts at the time. For

example, Figure 2 shows the distribution of natural disasters across decades. The highest numbers are seen during the RoC era, which is likely to be the result of better recording quality of historical events. As we can see in Figure 3, the incidence of political turmoils also peaked in the RoC era, which is a fair reflection of the national wars and increased political tension in the country.

Table 3.1: The Definitions of General and Detailed Categories of the Historical Shocks

(1) General Category	(2) Total N.	(3) Detailed Category	(4) Definition	(5) Total N.
Natural Disaster	166	Earthquake	An earthquake that was recorded in the county	22
		Flood	A flood that was recorded in the county	47
		Drought	A drought that was recorded in the county	26
		Fire	A destructive urban fire that was recorded in the county	18
		Plague	An epidemic that was recorded in the county	16
		Extreme Weather	An unusual weather event that was recorded in the county, such as a severe thunderstorm	15
		Tiger invasion	Wild tigers invaded some areas of a county, causing casualties and economic damage	9
Political Turmoil	180	War	Military forces crossed fire on a county's territory	107
		Protest	A recorded protest, such as a student protest against the government; or a strike. The majority of such protests happened after 1911.	29
		Local Riot	A recorded riot in a county.	18
Public Constructions	120	Military	The completion of a military construction project	1
		Bridge	The construction of a new bridge	11
		Dam	The construction of a new dam	6
		Hospital	The construction of a new hospital	3
		Roads/Waterway	A new road or a waterway opened	9
		Memorial	The construction of a memorial site	6
		Other	Other public constructions	12
		Economic Policies	A new or a change in local policy concerning the economy; or a local population census	45
Famine	27	Caused by Flood	A recorded famine that happened due to a precedent flood	2
		Caused by Drought	A recorded famine that happened due to a precedent drought	15
		Caused by Locusts Plague	A recorded famine that happened due to a precedent locusts plague	3
Industrial Development	46	New Factory	The foundation of a modern factory	34

Note: There could be multiple shocks for a county in a year.

I collected historical shocks from the the Huangshan City Archive Office and gazetteers from the Huangshan University Library. I specifically collected shock variables from the chorographies at the county level; they are local history records created and maintained by the county-level government.

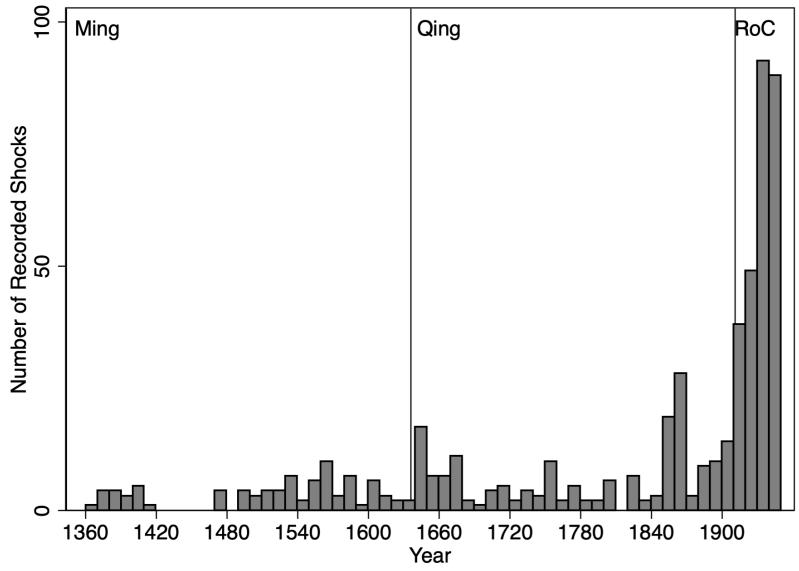


Figure 3.1: Number of Shocks by Decade

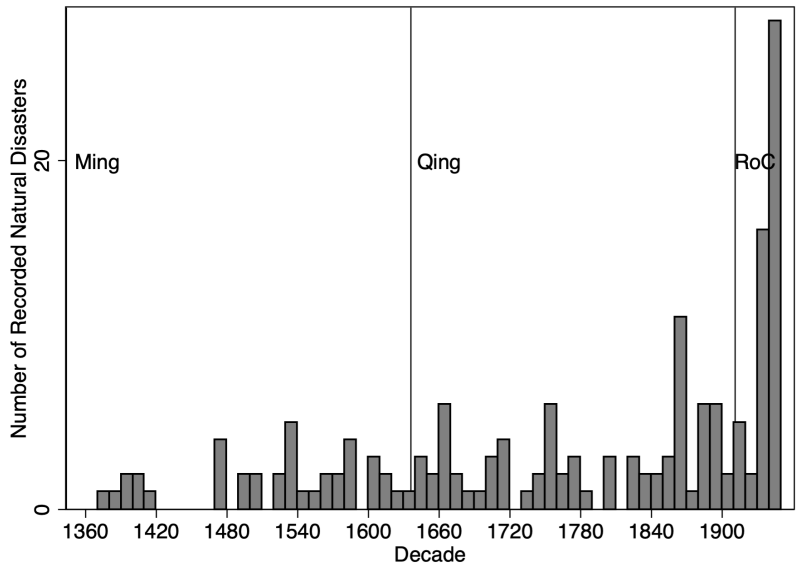


Figure 3.2: Number of Natural Disasters by Decade

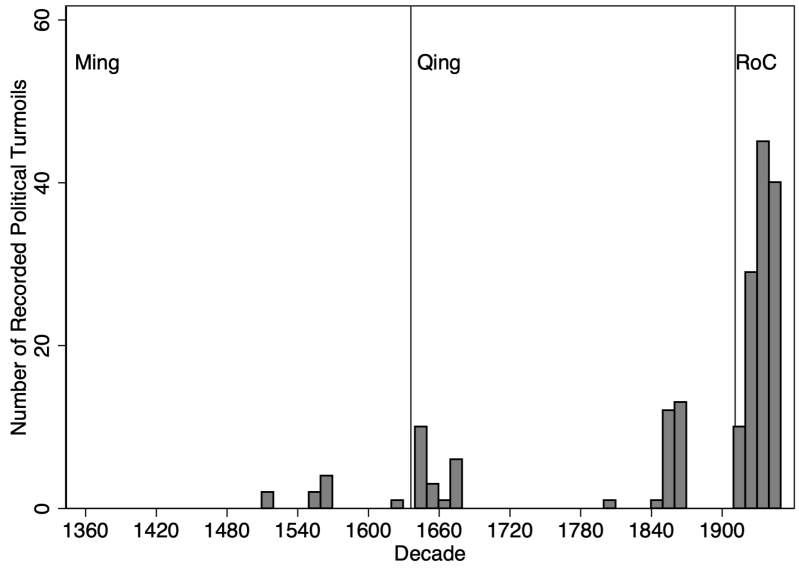


Figure 3.3: Number of Political Turmoils by Decade

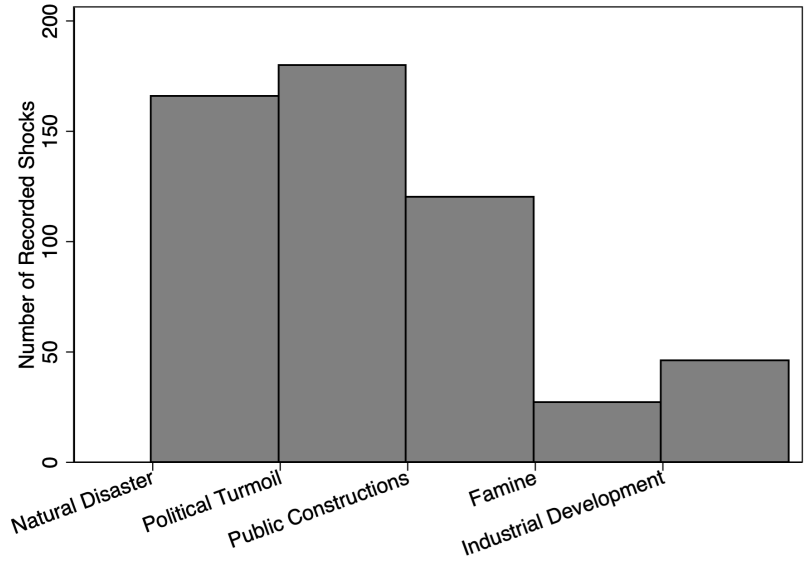


Figure 3.4: Number of Shocks by Detailed Categories

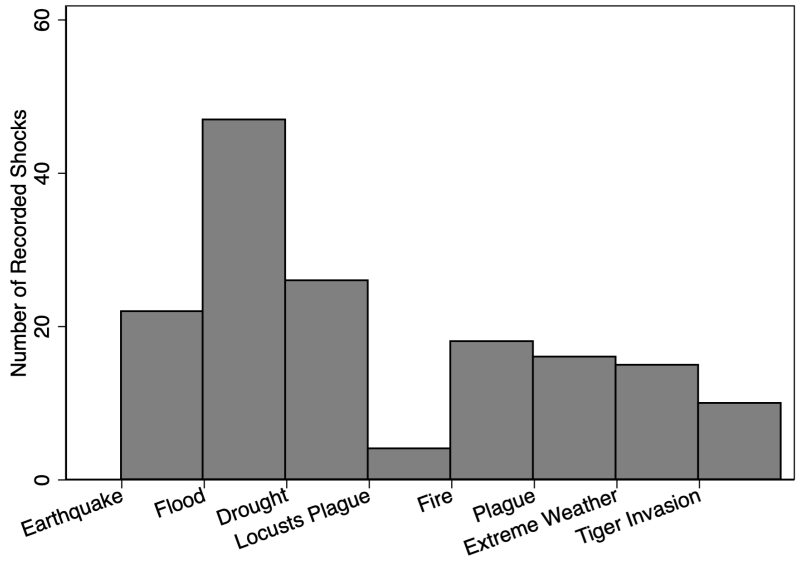


Figure 3.5: Number of Natural Disasters by Detailed Categories

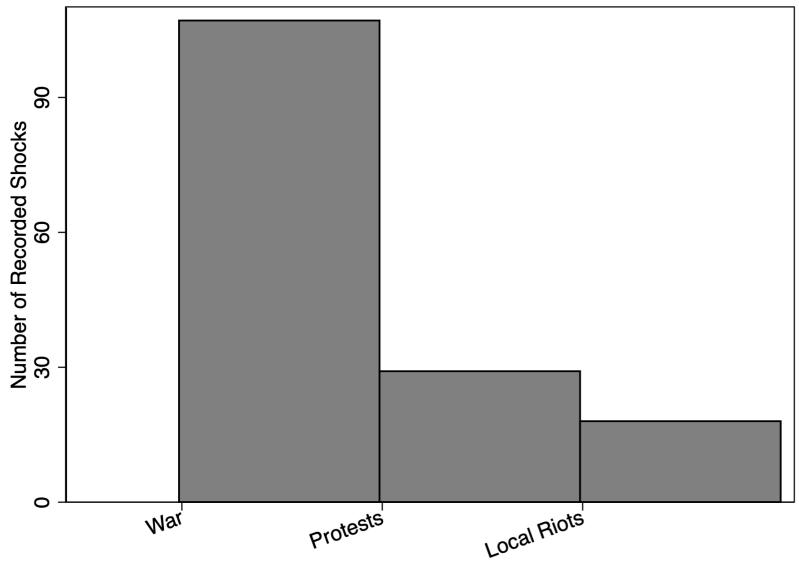


Figure 3.6: Number of Political Turmoils by Detailed Categories

Figure 4 shows the count of shocks by general categories. Figure 5 shows the count of different types of natural disasters; Figure 6 shows the count of different types of political turmoils.

3.2 Outcome Variables

The individual-level outcome variables will be collected from genealogies of Huizhou families that can be obtained from the Huangshan University Library and the Huangshan City Archive Office. A typical genealogy from the Huizhou area chronicles all members of a lineage, starting from a few generations before the first generation that migrated to Huizhou. Members' information include their names, relationships with other members, education level, occupation, residential location, and age of death. Members who achieved relatively higher social status are usually recorded with more information. Currently, based on samples of Huizhou genealogies that I have read and data that I have already collected, I conclude that the following information could be obtained and incorporated as outcome variables in my estimation.

Wealth measurement. My main data source is the land and house transaction data I have been collecting from Huangshan, Anhui Province in China since the summer of 2017. Those documents are usually written by an educated acquaintance of the seller and buyer. They are usually kept by buyers and they serve similar purposes as property ownership certificates. Each document follows a certain template that is more or less constant throughout the Ming Dynasty, the Qing Dynasty and the ROC era. A regular document typically states the name of sellers and buyers, location, the type of house layout, an indicator of the relationship between buyers and sellers (if they belong in the same lineage, if they are directly related), currency and price, date. Sometimes a document also includes more detailed information about the house, such as the house's orientation, area of the house, room count, etc. Based on the price index I develop in Chapter 2, I will be able to estimate the value of each recorded house in different years. The sample so far has more than 1000 house and land transaction records, and I expect the sample to grow while I collect more transaction documents.

Educational and Labor Market Outcomes. The data that measures the educational and labor market outcomes will be from the genealogies of local families. Genealogies record the major information of male lineage members and some basic information of their family. Information on male lineage members usually include the result of Keju (the Chinese civil

examination system) if a member ever participated in the exam. It also records the major achievements of its members, usually include the official positions a member held, the amount of financial support a member provided to the lineage, etc. Aggregated information on the lineage level will be used to construct my educational and labor market outcome variables.

3.3 Lineage Characteristics

Lineage characteristics will be similarly obtained from the collections of Huizhou genealogies at the Huangshan University Library and the Huangshan City Archive Office. The above-mentioned outcome variables are constructed on the individual or individual family level. I will link the house and land transaction data to the genealogies by the names of property owners and their locations. I then group the individual-level information to create lineage characteristics. These include the size of the lineage, the highest official rank in the government that a lineage's members have ever achieved, the number of lineage members that served in the government, the number of members that succeeded in the Keju exam, the gender of the lineage leader. A collective measure of a lineage's wealth level could be constructed by combining the house and land transaction data by lineage.

4 Methodology

To investigate the average effect each shock has on individuals, I first run a difference-in-difference model as shown in equation 4.1. Figure 4.1 is a illustration of the difference-in-difference design.

$$Outcome_{it} = \alpha_1^s Post_{it} + \sum_s \alpha_2^s ShockType_{it}^s + \sum_s \alpha_3 Post_{it} ShockType_{it}^s + \gamma Z_{it} + \varepsilon_{it} \quad (4.1)$$

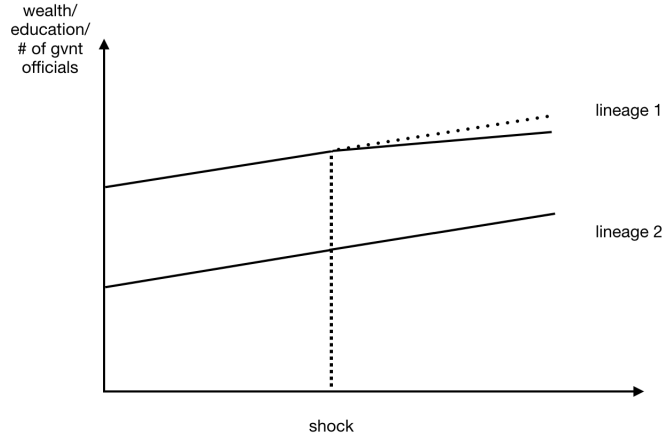


Figure 4.1: Illustration of a D-in-D Analysis

The basic unit of this regression is an individual (or an individual family) i living in time t ; s represents a certain type of shock. $Post_{it}$ is a dummy variable indicating if this observation is from before or after a shock. $ShockType_{it}^s$ is a vector of dummy variables indicating whether the observation is from a lineage that was affected by a certain type of shock s . Z_{it} is a vector of individual characteristics, as discussed in Section 3.2. The coefficient of interest is α_3 which captures the average effect of type s shock.

With detailed information on lineages available, I plan to employ a triple difference strategy to explore the differential effect of the shocks on individuals from different lineages. Equation 4.2 demonstrates the research design.

$$\begin{aligned}
Outcome_{ijt} = & \beta_1 Post_{ijt} + \sum_s \beta_2^s ShockType_{ijt}^s + \beta_3 LinChars_{jt} + \sum_s \beta_4^s Post_{ijt} * ShockType_{ijt}^s \\
& + \sum_s \beta_5^s LinChars_{jt} * ShockType_{ijt}^s + \beta_6 LinChars_{jt} * Post_{ijt} \\
& + \sum_s \beta_7^s Post_{ijt} * ShockType_{ijt}^s * LinChars_{jt} + \gamma Z_{it} + \eta_{ijt}
\end{aligned} \tag{4.2}$$

Again, the basic unit of this regression is an individual i living in time t , with additional

information of lineage j . s still represents the type of the shock dummies. Comparing to Equation 4.1, the triple difference strategy adds a new vector of variables that contains lineage-specific information; $LinChars_{it}$ is a set of lineage characteristics such as size, existing years, the highest official rank that its members ever achieved, the gender of the lineage leaders, etc. The coefficients of interest are β_4 and β_7 . β_4 estimates the average causal effect of a shock on an outcome, β_7 is expected to estimate how each lineage attribute affects the outcome variable after a shock. Note that to study the short run effect, the time window needs to be decided. To study the long run, the effect of the shocks needs to be aggregated for each lineage.

At the current stage, I do not have the lineage information, thus I cannot match each individual that I observe in the land and house transaction data with lineages. Instead, I will present the result from a simple model that examines the effect of different shocks on the housing price. The ordinary least square regression equation is as follows:

$$Price_{ht} = \sum_s \beta_1^s Post_{ht} * ShockType_{ht}^s + County_h + Year_t + HouseChars_{ht} + \varepsilon_{ht} \quad (4.3)$$

In Equation 4.3, each h represents a house transaction. β_1 is the coefficient of interest. $ShockType_{hjt}$ indicates whether a transaction happened within ten years after a shock of type j in decade t . β_1 thus shows the average effect of such a shock on the value of an affected property in the next ten years. $County_h$ is the county in which this transaction happened. $Year_{ht}$ is the year the transaction occurred. $HouseChars_{ht}$ are the characteristics of each house as recorded in the transaction dataset that I collect and use in Chapter 2. To account for the differential effect of the shocks in each year afterwards, I also show the results of the following equation:

$$Price_{ht} = \sum_s \beta_1^s Lag_{ht} * ShockType_{ht}^s + \sum_s \beta_2^s Lead_{ht} * ShockType_{ht}^s + County_h + Year_{ht} + HouseChars_{ht} + \varepsilon_{ht} \quad (4.4)$$

I estimate this equation separately for each type of shocks. $Lead_h$ and Lag_h represent how many years a transaction happened before or after a shock, within a 21-year window. The coefficients of interest here are β_1 and β_2 ; they show the house price each year before and after the shock.

5 The Effect of Shocks on Housing Price

The following table shows the regression results from Equation 4.3. They are estimated β_1 . Column (1) shows the coefficients of the shocks, defined generally; Column (2) shows the coefficients of the shocks, defined in a more detailed way (see Table 3.1).

Table 5.1: Results of the OLS Regressions

	(1) ln(Price)		(2) ln(Price)
Natural Disaster	-0.074 (-1.43)	Earthquake	-0.206 (-1.53)
		Flood	-0.297*** (-4.63)
		Drought	0.299 (2.09)
		Fire	-0.148 (-1.62)
Political Turmoil	0.039 (0.59)	War (crossfire in the county)	0.161* (2.04)
		Protests	-0.398*** (-4.04)
		Local Riots	1.439*** (9.87)
Public Construction	0.024 (0.39)	Infrastructure	-0.313** (-2.63)
		New Economic Policy	0.241** (3.19)
		Other Public Construction	-0.191 (-1.84)
Famine	0.065 (0.64)	Famine caused by Flood	-0.729** (-3.46)
		Famine Caused by Drought	0.012 (0.72)
New Private Business	-0.011 (-0.12)	New Factory	0.083 (0.81)

	Other New Business	0.033 (0.26)
House Characteristics	✓	✓
Year FE	✓	✓
County FE	✓	✓
_cons	0.733 (0.72)	-0.419 (-0.41)
<i>N</i>	678	678

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the regression results from Equation 4.3. The coefficients presented are estimated β_1 . Column (1) shows the coefficients of the shocks, defined generally; Column (2) shows the coefficients of the shocks, defined in a more detailed fashion (see Table 3.1 for the definitions of the historical shocks).

In general, we do not observe a significant impact from shocks on house prices if we use a general definition of shocks. However, results in column (2) show that a more narrow definition of the shocks had larger impacts on house prices. Within natural disasters, flood significantly decreases house prices in the ten years after the flood. Earthquakes, droughts, and fires do not have obvious impacts. The reason could be that the Huizhou area was more prone to severe floods than other natural disasters; thus floods brought more damage to the local economy. Floods that led to large casualties brought the housing demand down. Within the category of political turmoil, different subcategories show impacts with opposite directions. A military crossfire in a county and local riots tend to increase house prices in the following ten years. This could be the result of decreased housing supply if a large number of them were destroyed during the riots, or that the demand increased as people wanted homes as shelters during a dangerous time, or both reason combined. The record did show that burning down houses was a common strategy used by rioters. On the other hand, protests were mainly responses to a national event. During the years with frequent protests (the RoC era), the majority of them addressed unsatisfying government reactions to foreign invasions or unequal treaties. Protests tend to decrease house prices as a likely outcome of dim economic outlook. The dichotomy between riots and protests reflect differential effects of local and national shocks on the housing market. Within the category of public construction, infrastructure construction decreases house prices in the following ten years; new economic policies have the opposite effect. I include every local census under the economic

policy category. This could coincide with increase in the population; the increased house prices might then be the result of the growing population instead of the policies. Famines caused by floods significantly decreases house prices in the following ten years.

To show the results from estimating Equation (4.4), I plotted the coefficients β_1 and β_2 for each year before and after the shock in a 21-year window, with ten years before the shock (β_2) and ten years after the shock (β_1). The year of the shock is year 0 on the horizontal axis. The reference year is set to be the year before the shock, i.e. year -1 on the horizontal axis. The vertical dashed lines represent year 0; the vertical solid lines in Figure 5.1 - 5.5 represent year -1.

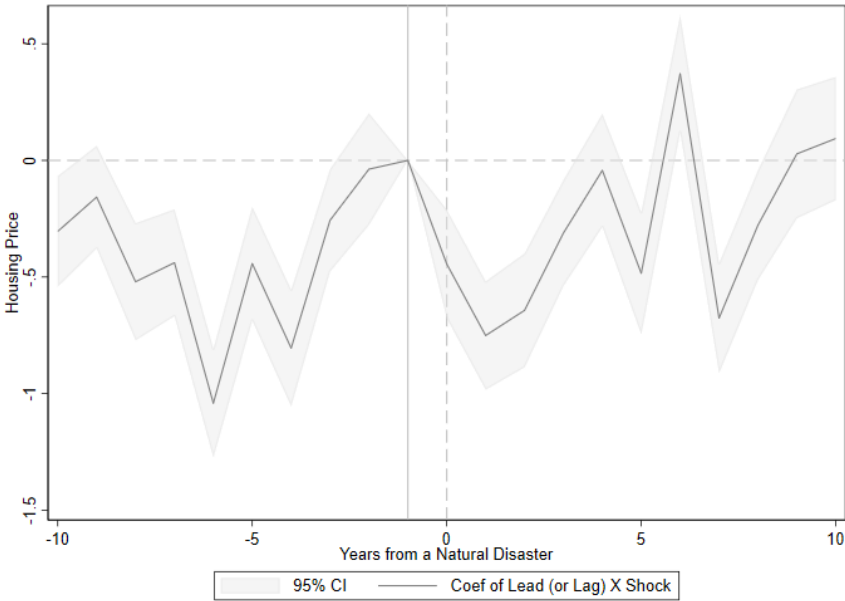


Figure 5.1: Impact on House Price: Natural Disasters

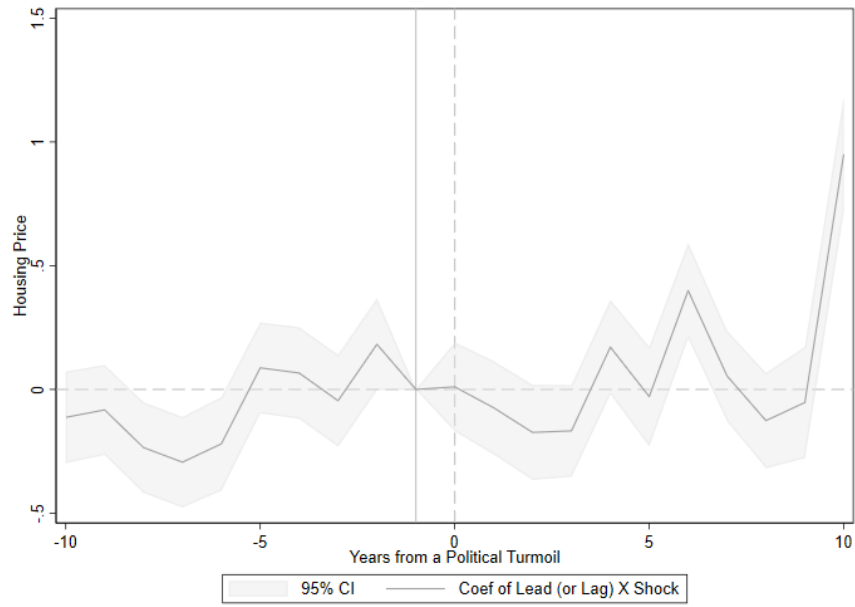


Figure 5.2: Impact on House Price: Political Turmoils

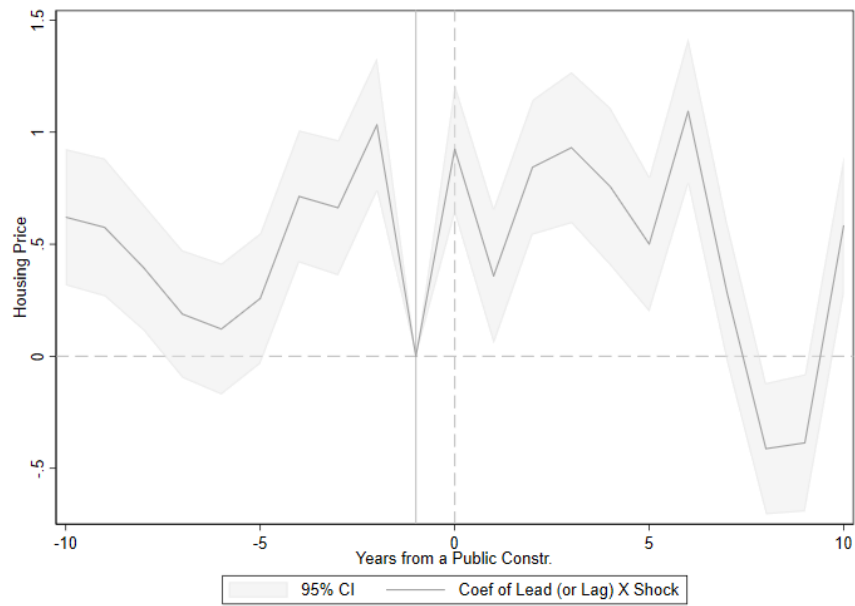


Figure 5.3: Impact on House Price: Public Constructions

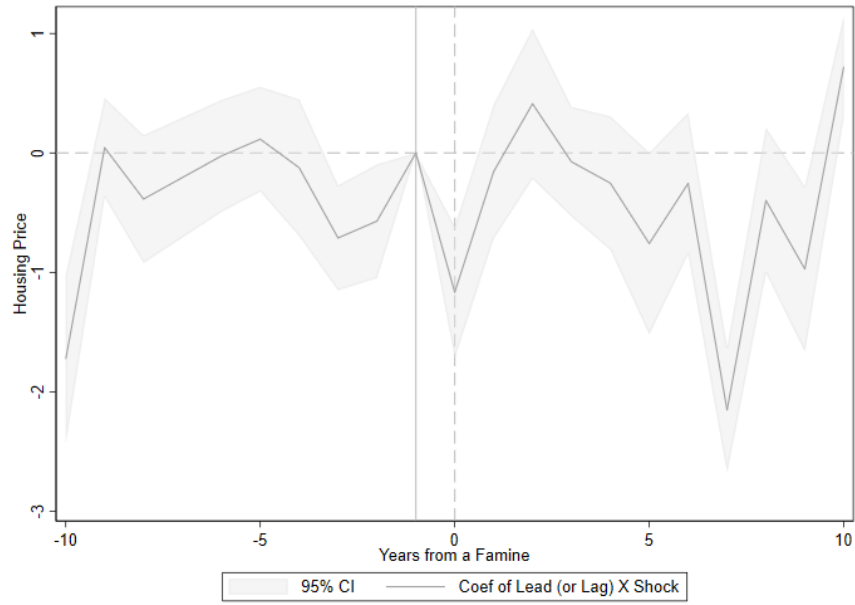


Figure 5.4: Impact on House Price: Famine

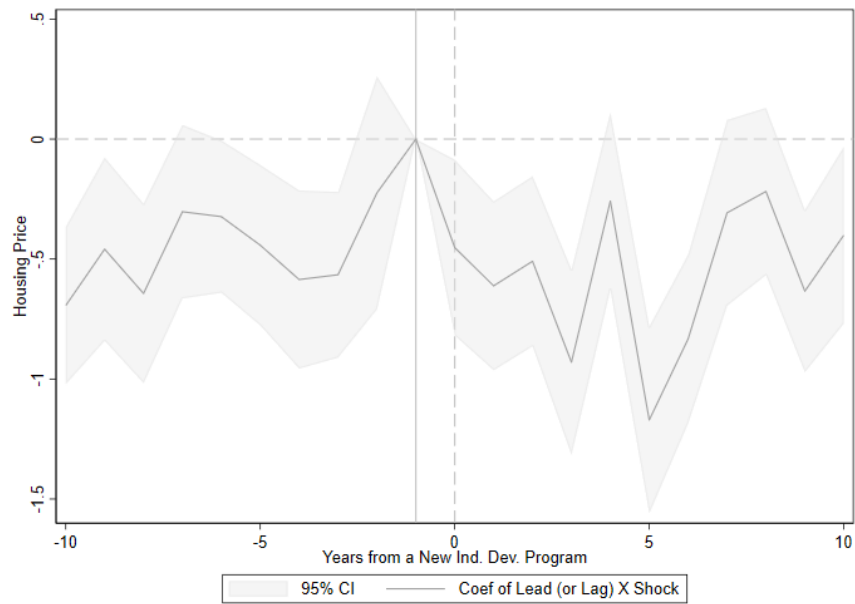


Figure 5.5: Impact on House Price: Industrial Development

It appears that house price tends to drop within the first two years after a major natural disaster event, then gradually increases (Figure 5.1). House price tends to drop within the first

three years after a political turmoil, then gradually increases (Figure 5.2). It also appears that the housing price takes a dip at the year of a famine (Figure 5.4). We do not observe clear trends in the impact of public constructions and industrial development on house price, as shown in Figure 5.3 and Figure 5.5; although there is an likely downward trend following a new local industrial development program. In Figure 5.3, the house price at year -1 is significantly lower than the years around it, which suggests that the public construction programs could be government's responses to events that were detrimental to the local economy

The preliminary results shown above do not contain any information on the lineage level; it is conducted with house transactions as the basic units. I expect more insightful results after obtaining lineage-level data.

6 Future Work

Due to restriction on currently available data data, this paper provides answers to a compromised research question: what is the impact of exogenous shocks on housing prices in the Huizhou area. This is similar to my preferred difference-in-difference research design (Equation 4.2), but without lineage identification and some of the outcome variables.

This paper will eventually answer the question of how natural, economic and political shocks effect people's lives in the short and long run, and how that effect interacts with various attributes of the lineages that they belonged to. The preferred research design is described in Equation 4.1 and Equation 4.2. This is made possible by the unique historical data set from the ancient Huizhou prefecture of China. The data that I will obtain in the near future will be genealogies from the Huizhou area that can be collected from the Huangshan University Library and the Huangshan City Archive. Upon the completion of data collection and obtaining estimation results from Equation 4.1 and Equation 4.2, this paper will contribute to the existing research on first, the impact of historical shocks on a larger set of outcome variables; second, the role that lineages played in how lineage responded and recovered from such shocks. This paper

adds to our understanding of the imperial Chinese economy on a rarely seen micro level and opens possibilities for other lineage related research.

To extend the research scope even further, I would also like to collect data on the current generation of the lineages from the Huizhou area to test whether such differential effect of historical shocks still persists today.

7 Conclusion and Discussion

With currently available data, this paper shows how each historical shock influenced the house prices in the following ten years. Flood and flood-led famines significantly decrease the house prices. Different political turmoils had different effects; different public construction programs also had different impacts.

However, to proceed with the next step, some problems need to be addressed. A potential endogeneity problem comes from confounding variables that contribute to both the outcomes and how badly one lineage might be affected by the shocks. A more powerful lineage is more likely to invest in precautionary measures such as building dams or creating mutual funds for its member, thus is less likely to be affected by the loss brought by a shock (McDermott, 2015). If we compare the difference in the outcomes using the variation in how badly lineages were suffered from the shock, we might be simply picking up the difference in outcomes between a more powerful lineage and one that is less so. As a result, how to accurately measure the pre-shock power of the lineage is an important task of this paper.

In the current research design, measuring wealth with land transaction data suffers from sample selection issue. Some lineages might have more or less transaction records preserved comparing to other lineages due to unobserved attributes. A possible solution is to use the change in total property value owned by lineages, instead of the absolute value each year.

Another potential problem comes from the spillover effect of the shocks. For example, when a village is hit hard with flood, it suffers from negative productivity shock as some

agricultural land might become nonarable. The subsequent population outflow to nearby villages might contribute to the growth in agricultural production at those nearby locations as labor is the most important input in intensive farming. If we compare its later performance with lineages from nearby villages, instead of the effect of shocks, we might be picking up the effect of the migration from one village to another. Apart from adding a distance variable that could possibly account for this effect, there are two arguments one can make. The first one is, if the potential spillover of a shock at one location has similar effect on the other location, the estimation of the effect of the shock would be biased but arguably conservative. The second argument is that this problem might be harmless depending on which research question I am asking. If we are asking about the effect of shock on the local level of economic situation as a whole, the spillover could be thought of as one mechanism that is contributing to the result.

8 Appendix

Hedonic variables included in the estimations but omitted from Table 5.1:

Table 8.1: Regression Coef. of the Hedonic Var. in Eq. 4.3

	(1)	(2)
	ln(Price)	ln(Price)
Sanjian \times Rooms	0.0692*** (6.46)	0.0615*** (5.82)
Sihe \times Rooms	0.0992*** (10.41)	0.0983*** (10.47)
Commercial Property	0.879*** (8.12)	0.854*** (7.98)
Female Seller	-0.200** (-3.00)	-0.289*** (-4.33)
Same Lineage	-0.0670 (-1.51)	-0.0841 (-1.93)
Same Family	0.514*** (6.40)	0.475*** (5.91)
Seller in Urgent Situations	0.659*** (5.42)	0.692*** (5.73)
New	-0.113 (-1.02)	-0.220* (-1.99)
Next to Ancestral Halls	0.236* (2.57)	0.277** (2.99)
Abandoned	-0.00305 (-0.02)	0.0317 (0.23)
Next to Main Roads	0.867*** (5.24)	0.870*** (5.32)
East	0.979*** (9.21)	1.031*** (9.81)
West	0.284* (2.16)	0.303* (2.30)

South	1.261*** (10.64)	1.186*** (9.96)
General Shocks	✓	
Detailed Shocks		✓
Currency	✓	✓
Transaction Methods	✓	✓
County FE	✓	✓
Year FE	✓	✓
_cons	0.323 (0.26)	1.073 (0.67)
<i>N</i>	678	678

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This table presents the coefficients of variables included in the regressions shown in Table 5.1 but were omitted from Table 5.1. The first column shows the results from the regression in which shocks are generally defined; the second column shows the results from the regression in which shocks are more detailedly defined.

Bibliography

- [1] Alexander, D. and Currie, J. (2017). Is it who you are or where you live? residential segregation and racial gaps in childhood asthma. *Journal of health economics*, 55:186–200.
- [2] Banzhaf, H. S. and Walsh, R. P. (2008). Do people vote with their feet? an empirical test of tiebout. *American Economic Review*, 98(3):843–63.
- [3] Barreca, A., Clay, K., Deschenes, O., Greenstone, M., and Shapiro, J. S. (2016). Adapting to climate change: The remarkable decline in the us temperature-mortality relationship over the twentieth century. *Journal of Political Economy*, 124(1):105–159.
- [4] Bayer, P., Keohane, N., and Timmins, C. (2009). Migration and hedonic valuation: The case of air quality. *Journal of Environmental Economics and Management*, 58(1):1–14.
- [5] Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262.
- [6] Bishop, K. C. and Timmins, C. (2018). Using panel data to easily estimate hedonic demand functions. *Journal of the Association of Environmental and Resource Economists*, 5(3):517–543.
- [7] Blau, F. D., Ferber, M. A., Winkler, A. E., and Winkler, A. E. (1992). *The Economics of Women, Men, and Work*. Prentice-Hall Englewood Cliffs, NJ.
- [8] Bleakley, H. and Ferrie, J. (2016). Shocking behavior: Random wealth in antebellum georgia

- and human capital across generations. *The Quarterly Journal of Economics*, 131(3):1455–1495.
- [9] Carbone, J. C., Hallstrom, D. G., and Smith, V. K. (2006). Can natural experiments measure behavioral responses to environmental risks? *Environmental and Resource Economics*, 33(3):273–297.
- [10] Centers for Disease Control & Prevention et al. (2013). Asthma factsdc’s national asthma control program grantees. *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention.*
- [11] Chattopadhyay, S. (1999). Estimating the demand for air quality: new evidence based on the chicago housing market. *Land Economics*, pages 22–38.
- [12] Chay, K. Y. and Greenstone, M. (2003). The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. *The quarterly journal of economics*, 118(3):1121–1167.
- [13] Chen, T., Kung, J. K.-s., and Ma, C. (2017). Long live keju! the persistent effects of china’s imperial examination system.
- [14] Chen, T., Kung, J. K.-s., and Ma, C. (2019). Long live keju! the persistent effects of china’s civil examination system. *The Economic Journal*.
- [15] Coffey, B. (2003). A reexamination of air pollution’s effects on infant health: Does mobility matter? *Working paper*.
- [16] Currie, J., Davis, L., Greenstone, M., and Walker, R. (2015). Environmental health risks and housing values: evidence from 1,600 toxic plant openings and closings. *American Economic Review*, 105(2):678–709.
- [17] Currie, J. and Neidell, M. (2005). Air pollution and infant health: what can we learn from california’s recent experience? *The Quarterly Journal of Economics*, 120(3):1003–1030.

- [18] Dahl, G. B. (2002). Mobility and the return to education: Testing a roy model with multiple markets. *Econometrica*, 70(6):2367–2420.
- [19] Decker, S. and Schmitz, H. (2016). Health shocks and risk aversion. *Journal of Health Economics*, 50:156–170.
- [20] Deschênes, O., Greenstone, M., and Shapiro, J. S. (2017). Defensive investments and the demand for air quality: Evidence from the nox budget program. *American Economic Review*, 107(10):2958–89.
- [21] Drees, M., van de Minne, A., et al. (2016). Do the determinants of house prices change over time? evidence from 200 years of transactions data. Technical report, European Real Estate Society (ERES).
- [22] Eichholtz, P. M. (1997). A long run house price index: The heregracht index, 1628–1973. *Real estate economics*, 25(2):175–192.
- [23] Freedman, M. (1961). The family in china, past and present. *Pacific Affairs*, 34(4):323–336.
- [24] Freeman, R., Liang, W., Song, R., and Timmins, C. (2019). Willingness to pay for clean air in china. *Journal of Environmental Economics and Management*, 94:188–216.
- [25] Gerking, S. and Stanley, L. R. (1986). An economic analysis of air pollution and health: the case of st. louis. *The review of economics and statistics*, pages 115–121.
- [26] Giglio, S., Maggiori, M., and Stroebel, J. (2014). Very long-run discount rates. *The Quarterly Journal of Economics*, 130(1):1–53.
- [27] Guarnieri, M. and Balmes, J. R. (2014). Outdoor air pollution and asthma. *The Lancet*, 383(9928):1581–1592.
- [28] Hill, R. J. (2013). Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. *Journal of economic surveys*, 27(5):879–914.

- [29] Ito, K. and Zhang, S. (2016). Willingness to pay for clean air: Evidence from air purifier markets in china. *National Bureau of Economic Research*, (No. w22367).
- [30] Li, S. and Sato, H. (2008). Class origin, family culture, and intergenerational correlation of education in rural china. *China Economic Quarterly*, 7(4):1105–1130.
- [31] Liu, T. (2013). *Traditional Building Skills of Huizhou Folk Houses*. Anhui Science and Technology Publishing House.
- [32] Mansfield, C., Johnson, F. R., and Van Houtven, G. (2006). The missing piece: Valuing averting behavior for children’s ozone exposures. *Resource and Energy Economics*, 28(3):215–228.
- [33] Mazumdar, S. (2001). Rights in people, rights in land: Concepts of customary property in late imperial china. *Extrême-Orient Extrême-Occident*, pages 89–107.
- [34] McDermott, J. P. (2013). *The Making of a New Rural Order in South China: Volume 1: I. Village, Land, and Lineage in Huizhou, 900–1600*, volume 1. Cambridge University Press.
- [35] McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record*, (673).
- [36] Moretti, E. and Neidell, M. (2011). Pollution, health, and avoidance behavior evidence from the ports of los angeles. *Journal of human Resources*, 46(1):154–175.
- [37] Nicholas, T. and Scherbina, A. (2013). Real estate prices during the roaring twenties and the great depression. *Real Estate Economics*, 41(2):278–309.
- [38] Peng, W. (2007). *The History of Chinese Currencies (in Chinese)*. Shanghai People’s Press.
- [39] Pope, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., Heath, C. W., et al. (1995). Particulate air pollution as a predictor of mortality in a prospective study of us adults. *American journal of respiratory and critical care medicine*, 151(3):669–674.

- [40] Russell, B. (1922). *The problem of China*. Century.
- [41] Shiller, R. C. (2005). *Irrational exuberance*. Princeton, N.J.: Princeton University Press.
- [42] Shiue, C. H. (2016). Social mobility in the long run: An analysis with five linked generations in china, 1300-1900. Technical report, Working paper.
- [43] Smith, V. K., Carbone, J. C., Pope, J. C., Hallstrom, D. G., and Darden, M. E. (2006). Adjusting to natural disasters. *Journal of Risk and Uncertainty*, 33(1-2):37–54.
- [44] Smith, V. K., Taylor Jr, D. H., Sloan, F. A., Johnson, F. R., and Desvousges, W. H. (2001). Do smokers respond to health shocks? *Review of Economics and Statistics*, 83(4):675–687.
- [45] Teso, E. (2014). The long-term effect of demographic shocks on the evolution of gender roles: Evidence from the transatlantic slave trade. *Journal of the European Economic Association*.
- [46] Triplett, J. (2004). Handbook on hedonic indexes and quality adjustments in price indexes.
- [47] Watson, R. S. (1982). The creation of a chinese lineage: the teng of ha tsuen, 1669–1751. *Modern Asian Studies*, 16(1):69–100.
- [48] White, E. N., Snowden, K., and Fishback, P. (2014). *Housing and mortgage markets in historical perspective*. University of Chicago Press.
- [49] World Health Organization et al. (2016). Ambient air pollution: A global assessment of exposure and burden of disease.
- [50] Zhang, T. (Forthcoming). Land law in chinese history. *Routledge Companion to Chinese Legal History*.
- [51] Zhu, Y. (2011). On the chinese real estate corporations in shanghai during the anti-japanese war. *Historical Research in Anhui*, (3):7.

[52] Zivin, J. G. and Neidell, M. (2009). Days of haze: Environmental information disclosure and intertemporal avoidance behavior. *Journal of Environmental Economics and Management*, 58(2):119–128.

Vita

Siyu Pan was born and raised in Southern Anhui Province, China. She studied international trade at Renmin University of China as an undergraduate student, and then attended the Graduate Program in Economic Development at Vanderbilt University in Nashville, Tennessee. She went on to earn a Ph.D. in economics from Georgia State University in Atlanta, Georgia.

At Georgia State University, Siyu studied environmental economics, economic history, health economics and urban economics.

In the Fall of 2020, Siyu will join the Department of Economics at Oberlin College as a two-year visiting assistant professor.