Title
Essays on Spatial Development

Author
Rothenberg, Alexander David

Publication Date
2012

Essays on Spatial Development

by

Alexander David Rothenberg

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Economics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Bryan S. Graham, Chair
Professor Edward Miguel
Professor Patrick Kline
Professor Robert Helsley

Spring 2012

Essays on Spatial Development

Abstract

Essays on Spatial Development

by

Alexander David Rothenberg
Doctor of Philosophy in Economics

University of California, Berkeley

Professor Bryan S. Graham, Chair

This dissertation contains three essays on the relationship between the spatial distribution of economic activity and different types of public goods. In the first essay, I study how changes in transport infrastructure affect the location decisions of firms by examining how manufacturers responded to changes in road quality in Indonesia. Using new data, I document massive upgrades to Indonesia's highway networks during the 1990s, a period in which national transportation funding increased by 83 percent. I first show that these road improvements were accompanied by a significant dispersion of manufacturing activity, and that different industries responded in ways predicted by theory. To make better counterfactual predictions, I develop a structural model of location choice in which firms face a trade off: locating closer to demand sources requires firms to pay higher factor prices. The model predicts that some location characteristics relevant to firms are determined in equilibrium, necessitating the use of instrumental variables. I estimate a random coefficients logit model with endogenous choice characteristics and find significant differences in firms' willingness to pay for greater market access across different industrial sectors. Counterfactual policy simulations suggest that new toll roads connecting urban areas would cause a modest amount of industrial suburbanization. In contrast, upgrading rural roads would have little or no effect on equilibrium firm locations.

In the second essay, co-authored with Bryan Graham, we provide estimates of the implicit prices that consumers and firms pay for access to infrastructure in Honduras. Without credible estimates of the welfare effects of infrastructure improvements, it is impossible for policymakers to know whether the benefits of these investments outweigh their substantial costs. A major challenge with trying to understanding the welfare consequences of improved transport infrastructure is an identification problem: transport improvements are never randomly assigned. We overcome this identification program by exploiting variation from a novel natural experiment: Honduras' infestation with Panama disease. The Honduran railroad network was constructed by fruit companies to ship bananas from plantations to port cities, but because of an unpredictable outbreak of Panama disease, major plantations and

their associated railway infrastructure were abandoned. We argue that outbreaks of Panama disease were extremely difficult to predict, and because of this, conditional on the railway network that existed in the 1930s and a host of observable characteristics, the areas where railway lines were abandoned were randomly assigned. We use our identification strategy to uncover the implicit prices of access to infrastructure paid by consumers and the implicit production costs paid by firms.

In the third essay, co-authored with Rachel Glennerster and Edward Miguel, we study how spatial variation in ethnic diversity, which exists largely for historical reasons, affects the provision of local public goods in rural Sierra Leone. Scholars have pointed to ethnic divisions as a leading cause of underdevelopment, due in part to their adverse effects on public goods. We investigate this issue in post-war Sierra Leone, one of the world's poorest countries. To address concerns over endogenous local ethnic composition, we use an instrumental variables strategy relying on historical census data on ethnic composition. We find that local diversity is not associated with worse public goods provision across a variety of outcomes, specifications, and diversity measures, with precisely estimated zeros. We investigate the role that leading mechanisms proposed in the literature play in generating the findings.

To Sarah, with all of my love.

Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without the support and guidance of Bryan Graham, my primary adviser and mentor in this research. Our extensive discussions about these projects were extremely helpful for focusing my thoughts and providing direction. I am very grateful for the fact that he was willing to spend considerable amounts time reading drafts of my work and offering suggestions for improving it, even at very early stages. I feel that I've learned a great deal from Bryan and his way of thinking has been tremendously influential in this work.

I am also indebted to Edward Miguel, whose continued enthusiasm for these projects was contagious and extremely helpful when I needed support. Ted's ideas for framing this research and his big picture focus were invaluable when I was lost in the details. I am grateful for the time he spent on these ideas and his careful attention to detail, and his advice and suggestions have substantially improved the quality of this dissertation. I would also like to thank the other members of my committee, Patrick Kline and Robert Helsley, for their comments and advice on these projects, which are much better as a consequence of their valuable input. Several other professors at Berkeley should also be recognized, including David Card, Frederico Finan, Paul Gertler, and James Powell. This work has substantially benefitted from their advice and comments.

I would also like to thank all of my fellow graduate students, but most especially Maximilian Kasy, Sarath Sanga, Zach Liscow, Gee Hee Hong, Daniel Hicks, and Joan Hamory. Their comments and suggestions throughout the research process were amazingly helpful, as was their friendship. I also want to thank Glen Stringer for several invaluable detailed discussions about the IRMS dataset, as well as Ir. Taufik Widjoyono, Ir. Julius J. Sohilait, and Yohanes Richwanto at Departmen Pekerjaan Umum for generously granting me access to their data and answering my many questions about it.

This work was supported, in part, by fellowships and grants from several centers at U.C. Berkeley, including the Fischer Center for Real Estate and Urban Economics, the Center for Equitable Growth, the Institute of Business and Economic Research, and the Center of Evaluation for Global Action.

Finally, and most importantly, I owe a debt of gratitude to Sarah. The last six years have been challenging on multiple levels, and her continued support, faith, and patience through this process have been invaluable.

Chapter 1

Transport Infrastructure and Firm Location Choice in Equilibrium: Evidence from Indonesia's Highways

Abstract

Transport improvements can have two competing effects on firm spatial concentrations. By making it easier for firms to reach customers from a given site, lower transport costs can encourage agglomeration, but on the other hand, by expanding access to cheaper labor and land, lower transport costs make producing in more sites feasible and promote dispersion. To better understand how transport improvements affect the spatial distribution of economic activity, I study how the location choices of new manufacturers responded to changes in road quality in Indonesia. Using new data, I document massive upgrades to Indonesia's highway networks during the 1990s, a period in which national transportation funding increased by 83 percent. I first show that these road improvements were accompanied by a significant dispersion of manufacturing activity, and that different industries responded in ways predicted by theory. To make better counterfactual predictions, I develop a structural model of location choice in which firms face a trade off: locating closer to demand sources requires firms to pay higher factor prices. The model predicts that some location characteristics relevant to firms are determined in equilibrium, necessitating the use of instrumental variables. I estimate a random coefficients logit model with endogenous choice characteristics and find significant differences in firms' willingness to pay for greater market access across different industrial sectors. Counterfactual policy simulations suggest that new toll roads connecting urban areas would cause a modest amount of industrial suburbanization. In contrast, upgrading rural roads would have little or no effect on equilibrium firm locations.

1.1 Introduction

In many developing countries, investments in transport infrastructure are growing at an astonishing pace. China's total spending on transport projects increased from \$9.2 billion in 2000-2004 to \$26.4 billion in 2005-2009, while India's spending increased from \$2.9 billion to \$29.4 billion between the same periods.[1] The goal of these projects is to lower transport costs between different regions. But as regions become better connected, the spatial distribution of economic activity that emerges remains difficult to predict.

Better transportation networks might induce firms to locate outside of congested urban agglomerations, so that they can access cheaper land and labor. The possibility for dispersion is stressed by policymakers who believe that transport improvements can bring more jobs and firms to less developed regions. For instance, in national planning documents, Indonesia's government claims that transportation investments promote "the equitable distribution and dissemination of development efforts, penetrating the isolation and backwardness of remote areas".[2] As another example, when the Suramadu Bridge opened in 2009, connecting Surabaya, Indonesia's second largest city, to the less densely populated island of Madura, Indonesia's President Yudhoyono projected that "Madura will be much more developed as a result of the bridge".[3] Indeed, classic models in urban economics (Alonso, 1964; Mills, 1967; Muth, 1969) and economic geography (Helpman, 1998) suggest mechanisms through which lower transport costs induce a dispersion of firms and workers to more peripheral areas.

On the other hand, better roads make firms in existing cities more profitable by bringing them closer to other markets. Because of this, lower transport costs could intensify the self-reinforcing home market effects that cause agglomerations to form and grow. In the influential core-periphery model of Krugman (1991), reducing trade costs between two regions causes firms to agglomerate, pulling the entire manufacturing sector into one region. Thus, road improvements may actually exacerbate spatial inequalities instead of reducing them. Despite the prominent role that transport costs play in models of urban economics and economic geography, we currently have limited knowledge about their actual effects on firm location choices and, consequently, how they affect the growth paths of different regions. Moreover, we have few tools at our disposal to predict what equilibria would look like if new road programs were implemented, whether they be highways that connect major cities or upgrades to rural roads.

This paper makes several contributions to our understanding of how transport costs affect the spatial distribution of economic activity by exploiting unique data on a large road

[1] These figures, stated in constant 2000 U.S. dollars, are taken from the World Bank's Private Participation in Infrastructure (PPI) Project Database.

[2] This quotation is taken from a planning document describing transportation development objectives in Repelita VI (author's translation). Similar sentiments are echoed in other planning documents.

[3] This quotation is taken from Faisal, Achmad and Harsaputra, Indra "Suramadu bridge touted to boost economy, create jobs" *The Jakarta Post* 11 June 2009.

improvement program in Indonesia. During the 1970s and 1980s, quality paved highways in Indonesia consisted of only a few major arteries connecting provincial capitals and other large cities. However, in the early 1990s, there was an 83 percent increase in funding allocated for road improvements, and road networks throughout the archipelago were rapidly improved. Upgrading projects were not uniform over space or time, producing substantial variation in transport improvements that can be used to estimate their effects.

Using new panel data on the quality of major highways, I first present reduced form evidence suggesting that Indonesia's road improvements induced a moderate, statistically significant dispersion of manufacturing activity. During the same period in which road improvement projects were occurring, the spatial concentration of manufacturing employment fell by more than 20 percent. Interestingly, the amount of dispersion varied across industries in ways that are predicted from theory. For instance, the spatial concentration of producers of perishable goods, which deteriorate rapidly in transit and need to be consumed close to where they are produced, did not change significantly over the period, while it fell substantially for producers of durable goods. Although I see evidence of dispersion, new firms did not move to the most remote parts of Indonesia; instead, they suburbanized, locating increasingly in neighboring areas of existing agglomerations. Using a series of linear panel regressions, I estimate positive and significant average effects of road improvements on new manufacturing establishments and employment.

On its own this reduced form analysis only partially sheds light on the mechanisms behind these results, which hinders our ability to make counterfactual predictions. For instance, if firms move to new locations in response to better market access, their presence will drive up local wages and rents in these areas, and this will, in turn, affect the location choices of other firms. Predictions about what would happen if new roads were built may be inaccurate if these general equilibrium responses are not taken into account. To explain the relative importance of different mechanisms, and to make counterfactual predictions that incorporate these general equilibrium effects, I develop and estimate a structural model of firm location choice.

I present a multiple-region model of monopolistic competition and regional trade (e.g. Head and Mayer, 2004), designed to capture two sources of the costs and benefits of agglomerations. One key prediction of the model is that firm profits depend on a location's *market potential* (Harris, 1954), a weighted average of real regional incomes, where the weights decline with transport costs. This demand force pulls firms to locate in existing agglomerations. However, because local supply schedules for land and labor are upward sloping, locating in agglomerations is costly. Hence, firms face a tradeoff: those who locate closer to demand sources must pay higher factor prices for production. The model also allows for sectoral differences in the willingness to substitute between different location characteristics, motivated by the industry differences highlighted in the reduced form analysis. With some additional distributional assumptions on the unobserved components, I show how parameters of the model can be estimated with discrete choice techniques.

Identifying these parameters is challenging, since many characteristics that firms observe when determining where to operate (including local wages, rents, and access to other markets) are themselves affected by the decisions that other firms make, creating possible simultaneity problems. New road improvements may also be targeted to particular areas, and estimates of the effects of better market access may be confounded with the fact that areas with better roads were selected by policymakers, creating targeting bias. Moreover, without data on how location characteristics vary over time, it is impossible to distinguish features of firm profit functions that depend on these characteristics from those that depend on fixed natural productive amenities, many of which may be unobserved.

To overcome these identification problems, I combine the new panel data on road quality with techniques from industrial organization that allow researchers to estimate discrete choice models with endogenous choice characteristics (Berry et al., 1995). Panel data on road quality and market access enable me to control for time-invariant unobservables that may be correlated with the provision of infrastructure. For example, in Indonesia, long-term spatial plans dictated that certain areas would be targeted for road improvements. These plans were revised infrequently, and to the extent that they were adopted, controlling for location fixed effects enables me to remove the targeting bias from parameter estimates. Fixed effects also allow me to remove from parameter estimates the effects of other unobserved factors, such as time-invariant productive amenities.

To deal with simultaneity problems associated with identifying choice parameters, I combine location fixed effects with sequential moment restrictions. Under these restrictions, regional productivity shocks are innovations, unpredictable given past information, and lagged location characteristics can serve as instruments for current location characteristics. Although this identification strategy maintains certain assumptions, it strictly weakens the identification assumptions required for estimation with fixed effects alone.

After estimating the model, I discuss its predicted substitution patterns, showing that location pairs which are closer to one another along a range of distance measures have stronger cross-elasticities. The parameter estimates also suggest that there is substantial heterogeneity in firms' willingness to pay for greater market access across industrial sectors. For instance, food producers, textile firms, and sporting-goods manufacturers all have stronger preferences for locating closer to large markets than makers of wood products, which tend to locate closer to raw materials.

Finally, I use the model to predict what would have happened to industrial locations under two realistic counterfactual scenarios: the on-time construction of the Trans-Java Expressway and an improvement to certain rural roads. The Trans-Java Expressway is a series of proposed toll roads connecting cities along the northern coast of Java. Originally planned for operation in 1994, it has been mired in construction delays and remains incomplete. Using the model to simulate what would have happened to industrial locations, I find that the toll roads would have induced a moderate degree of increased suburbanization. With better roads, manufacturing activity would have moved further outside of existing urban centers, but firms would not have relocated to the remotest parts of Indonesia. In contrast, I find that upgraded rural roads have little if any significant effects on industrial locations, despite claims often made by policymakers to the contrary.

This paper contributes to a growing literature that studies the effects of transport infrastructure through the lens of trade theory (e.g. Michaels, 2008; Donaldson, 2010) and urban economics (e.g. Baum-Snow, 2007). While prior work has used models in which trade is driven by Ricardian comparative advantage or factor endowments, this paper uses a model that focuses on trade driven by increasing returns to scale and imperfect competition. This class of models is used frequently in economic geography, and this paper also contributes to a long-standing research program that focuses on testing such models (e.g. Davis and Weinstein, 2003; Redding and Sturm, 2008). Within this literature, there is a line of research that uses discrete choice models to estimate firm location choices, dating back to Carlton (1983). Most papers in this literature estimate choices for a single cross-section of firms (Coughlin et al., 1991; Head et al., 1995; Henderson and Kuncoro, 1996; Head and Mayer, 2004), and this paper builds upon prior work by using panel data, which allow me to distinguish between the effects of observed location characteristics and the effects of unobservable fixed factors.

Most importantly, to the best of my knowledge, prior work has largely not addressed the fundamental endogeneity problems associated with estimating firm location choices. The fact that a location's wages, rents, and market potential are determined in equilibrium necessitates the use of a model and conditional moment restrictions for identification.[4] By deriving the estimating equations from an explicit theoretical framework, constructing a time-varying measure of transport costs from a new dataset on road quality, estimating the model on a panel of new firms, and using structural econometric techniques to address the endogeneity of location characteristics, this paper aims to extend the empirical literature on firm location choices and transportation. While the results discussed here are undoubtedly specific to Indonesia, the model and empirical techniques advanced could be readily applied to examine the impact of regional policies on firms in other settings.

The rest of this paper is structured as follows: Section 1.2 describes Indonesia's road construction program and manufacturing activity in the late 1980s and 1990s. Section 1.3 describes a new dataset on road quality in Indonesia and discusses how these data are used to construct proxies for transport costs. It also discusses the data on newly entering manufacturing firms and location characteristics. Section 1.4 presents reduced form evidence on how road improvements induced greater dispersion of manufacturing activity. To obtain more accurate counterfactual predictions, Section 1.5 presents a structural model of monopolistic competition and regional trade, discussing how to identify and estimate its parameters. Section 1.6 presents parameter estimates from the choice model and discusses the predicted

[4] An exception is Liu et al. (2010), who study the location choices of firms investing in China between 1993 and 1996 and use a control-function approach to deal with unobserved heterogeneity across locations (**?**). The authors find, just as we do here, that not allowing for unobserved location characteristics causes researchers to substantially over-estimate wage coefficients.

locational substitution patterns. In Section 1.7, I use the model to predict what would have happened to industrial locations under various counterfactual scenarios, and Section 1.8 concludes.

1.2 Roads and Manufacturing in Indonesia

Although known for political repression, violence, and corruption, Suharto's regime in Indonesia (1967-1998) had an extraordinary development record. During the three decades in which he was in power, GDP grew by an average of 5% per year, and the poverty rate fell from 60% in the mid 1960s to around 10% in the early 1990s (Hill, 2000). One potential contributor to Indonesia's economic success was the government's investments in major public works programs, including improvements to transport infrastructure.

Indonesia's roads, many of which were built by the Dutch colonial regime in the 18th and 19th centuries, were left to crumble and deteriorate under the leadership of Indonesia's first president, Sukarno (1945-1967). After coming to power in 1967 as the second president of Indonesia, Suharto quickly recognized the need to improve the country's infrastructure, and he made road improvements a priority of his first two five-year development plans, Repelita I (1969-1974) and Repelita II (1974-1979).[5] However, funding was insufficient for broad transport improvements, and the projects undertaken involved upgrading connections between major urban centers.[6]

After the collapse of oil revenues in the late 1970s, spending on road infrastructure slowed considerably and was not a priority of either Repelita III (1979-1984) or Repelita IV (1984-1989). However, manufacturing began growing rapidly by the end of the decade, and roads that were improved in the 1970s required heavy maintenance. This encouraged a shift in development priorities during the 1990s. Table 1.1 shows large changes in allocations of funds for improving roads between Indonesia's fourth, fifth, and sixth five-year development plans. During Repelita IV, the total budget for road improvements was \$2.1 billion. This was increased by 84 percent in Repelita V (1989-1994), to a sum of \$3.9 billion.[7] Transportation investments were the single largest item of the budget during Repelita V, forming nearly 18 percent of total planned development expenditures. Funds for road improvements in Repelita VI (1994-1999) were planned to be kept at similar levels as the first half of the decade, but the Asian financial crisis of 1997-1998 and its concurrent political upheaval resulted in less spending than originally intended.

[5] In Bahasa Indonesia, the phrase *rencana pembangunan lima tahun* is literally translated as "five year development plan". In characteristic Indonesian fashion, this phrase is seldom spelled out but instead expressed by the acronym Repelita.

[6] Surveys of transportation improvements in Indonesia during this period are difficult to find in the literature, but Leinbach (1989) and Azis (1990) provide some useful discussion.

[7] These figures are all quoted in constant 2000 U.S. dollars.

During the 1990s, road improvements were substantial and aimed at a wider variety of projects than before. Explicit attention was given to connecting sparsely populated areas, and to infrastructure improvements outside of the major islands. The large increases in budgeted spending translated into huge improvements in the network. According to new data described in the next section, in 1990, only 16 percent of Sulawesi's roads were paved, but after a decade, 54 percent were paved. In Sumatra, only 32 percent of the network was paved in 1990, but by 2000, 70 percent of the network was paved.

Importantly, these road improvements were also designed to adhere to long-term national spatial plans. Such plans dictated that particular regions should receive infrastructure improvements, and they were revised very infrequently (approximately once a decade). This suggests that the road authorities did not regularly respond to changes in outcomes, and it also suggests that location fixed effects can remove much of the targeting bias.[8]

As the road network rapidly improved, Indonesia's manufacturing sector grew considerably. From 1985 to 1992, manufactured exports grew at an average annual rate of over 20 percent in real terms, while the share of labor intensive manufactures grew from 40 percent of exports in 1982 to over 60 percent in 1992.[9] However, after the Asian Financial Crisis, in which Indonesia experienced a massive exchange-rate depreciation that caused a financial crisis and political upheaval, spending on transport infrastructure slowed considerably. Moreover, local governments began to assert more authority during Indonesia's program of decentralization, and this involved transferring the maintenance of many national roads to local governments. Anecdotal evidence suggests that many local governments did not have the capacity to maintain the roads under their jurisdiction, and roads began to deteriorate (Davidson, 2010a).

1.3 Data and Measurement

In this section, I first define kabupatens, the spatial unit of analysis used in my empirical work. Then, I present new data on Indonesia's highway improvements and their subsequent deterioration, and I explain how they are used to construct a panel of transport cost estimates between locations. Finally, I discuss Indonesia's *Survei Industri* (SI), an annual census of manufacturing firms with more than 25 employees.

[8] This idea comes from conversations with the highway authorities at DPU. Unfortunately, I do not have access to the exact national spatial plans that were used, but it is worth noting that in every planning and budgeting document I do have access to, no information is provided at levels below the province.

[9] For more details on the rise of labor-intensive manufacturing in Indonesia, see Hill (2000).

1.3.1 Spatial Unit of Analysis

Throughout the paper, I focus only on the islands of Java, Sumatra, and Sulawesi, since these are the three islands with the largest amounts of population and manufacturing activity, and I use Indonesia's *kabupatens* (districts) as the spatial unit of analysis. The kabupaten is the second administrative division in Indonesia, nested below the province. Because many kabupatens were divided and partitioned into new kabupatens after the fall of Suharto, I aggregate back to the 1990 definitions in order to achieve a consistent geographic unit of analysis. The sample contains 185 kabupatens, with a median land area of 1,498 square kilometers. This is slightly smaller than the size of U.S. counties, which have a median area of 1,595 square kilometers. Indonesia's major cities are also given separate identifiers, and these designations are also used in the analysis.[10]

1.3.2 Data on Road Quality

Many of the major roads used in Indonesia today have been around in some form for centuries, meaning that their effects can only be studied by using variation in quality over time. This type of variation is different from the spatial variation in infrastructure access used in prior work (e.g. Michaels, 2008; Donaldson, 2010). An understanding of the effects of road quality improvements should be very relevant for policymakers acting in developing countries, since it is generally cheaper to repair existing roads than to build new ones.

Data on the evolution of road quality come from a unique source: Indonesia's Integrated Road Management System (IRMS), maintained by the Department of Public Works (*Departemen Pekerjaan Umum*, or DPU). In the late 1980s, DPU began to conduct extensive annual surveys of its road networks, collecting data along the kilometer-post intervals of all major highways. Road quality surveys were conducted by a team of surveyors, who measured the surface type and width of road segments and also collected longitudinal data for computing the international roughness index (IRI).[11] The original dataset is extremely detailed, with more than 1.2 million kilometer-post-interval-year observations. Although some of the road-link identifiers changed as roads were upgraded and reclassified, it is possible to merge the kilometer-post interval data to shapefiles of the road networks. This yields a panel of

[10] Note that roughly 13 percent of the firm-year observations in my sample were reclassified by aggregating kabupaten codes. Most of the reclassified observations (35,901 observations) were due to collapsing the five separate Jakarta codes into a single code. An additional 6,660 observations (0.2 percent of the sample) were reclassified by aggregating adjacent rural regions with small amounts of manufacturing activity. See Appendix Section 1.C.2 for more details.

[11] The international roughness index (IRI) is a measure of road quality that was developed by the World Bank in the 1980s. It is constructed as the ratio of a vehicle's accumulated suspension motion (in meters), divided by the distance travelled by the vehicle during measurement (in kilometers). See Appendix Section 1.C.1 for more details on IRI is and how it was measured.

quality measures along major inter-urban roads from 1990 to 2007.[12] Figures 1.1, 1.2, and 1.3 depict the evolution of pavement along the highway networks of Java, Sumatra, and Sulawesi respectively. These show considerable spatial variation in the timing and extent of the improvements, and they also highlight the magnitude of the road improvement program.

1.3.3 Measuring Transport Costs

In the Indonesian context, measuring the cost of transporting goods between regions is extremely challenging. A common approach in the trade literature is to first estimate a gravity equation, using detailed data on regional trade flows, and to back out transport costs from parameter estimates.[13] Unfortunately, regional trade flow data have never been systematically collected in Indonesia, so this approach is infeasible. Another method involves backing out transport costs from price differences (e.g. Donaldson, 2010). This requires invoking an iceberg trade costs assumption (Samuelson, 1954), prior knowledge of where certain goods are produced, and observations of prices of that good in various locations. Although Indonesia's central statistical agency, *Badan Pusat Statistik* (BPS), collects detailed data on goods prices used in constructing the CPI, they do so only for a limited number of provincial capital cities, making it difficult to exploit much spatial variation. Moreover, many of these provincial capitals are also ports, so trade between them would not necessarily rely on using the road network. It is also difficult to pin down goods that are only produced in a single location.

Faced with these challenges, I construct a proxy for transport costs using the available data on road quality. The measure is based on road roughness: when faced with potholes, ragged pavement, or unpaved surfaces, drivers slow down, and this reduction in speed increases travel time and hence the cost of transport. Of course, there is not a one-to-one relationship between road roughness and speed, because drivers choose the speed at which they travel, and different preferences for ride smoothness or the desired arrival time might induce different choices of speed.

Yu et al. (2006) provide a mapping between subjective measures of ride quality and roughness at different speeds. This mapping can be used to determine the maximum speed that one can travel over a road with a given roughness level while maintaing a constant level of ride quality. Given this roughness-induced speed limit, it is straightforward to calculate travel times along network arcs and to compute the shortest path between different regions, using travel time as the single cost factor (Dijkstra, 1959). Note that the travel times on road sections were computed using speeds derived from the detailed kilometer-post-interval roughness data, which were then aggregated to form cost measures along the network arcs.[14]

[12] Appendix Section 1.C.1 provides more detail about the road quality data, particularly the process of merging the interval data to network shapefiles and the creation of variables.

[13] See Anderson and Van Wincoop (2004).

[14] See Appendix Section 1.C.1 for more details. Note that the travel time measure incorporates a continuous

In order to allow for travel between islands, I use data on the locations of major ports and estimate travel times between them, effectively linking all of the regions together in one transport cost matrix.[15]

Travel time is a useful way of measuring transport costs, because it is correlated with distance (and hence fuel consumption) and should also be related to drivers' wage bills. From surveys of trucking firms throughout Indonesia, the Asia Foundation (2008) found that fuel and labor costs were the largest contributors to vehicle operating costs, reinforcing confidence in the travel time measure.[16] While most of the variation in travel times comes from changes in the quality of existing roads, some new toll roads were also constructed during the period (mostly on Java), creating variation in physical distances (and speeds) that is also captured in the measure.[17]

Table 1.2 presents summary statistics of average transport costs between a given kabupaten and all other kabupatens on that island for Java, Sumatra, and Sulawesi, for the period 1990-2005. Physical distances did not change substantially, because only a few toll roads opened up over the period, and these new roads were confined exclusively to Java. The average distance falls very slightly, but travel times decrease significantly from 1990 to 2000 (17 percent). Although physical distances remained unchanged in Sumatra, travel times fell by 24 percent, on average, from 1990 to 2000. Similarly, average travel times in Sulawesi fell by 38 percent over the time period, despite any change in physical distances. The rapid deterioration of road quality from 2000 to 2005 is also quite apparent.

The average summary statistics presented here mask substantial geographic and temporal variation in the areas that received the largest reductions in transport costs. For instance, in Java, the largest reductions in travel times over the 1990-2000 period occurred for Central Java, while in Sumatra, the largest reductions occured for the provinces of Riau, Jambi, Bengkulu, and South Sumatra. For Sulawesi, the provinces that received the largest improvements in average transport costs were Gorontalo and South Sulawesi.[18]

measure of road quality, the international roughness index (IRI), rather than a simpler binary measure for whether not a road is paved. This was done to better match the transportation literature, but both measures are highly correlated.

[15] For more details, see Appendix Section 1.C.1.

[16] Fuel and labor costs amounted to 53 percent of vehicle operating costs on average, according to the survey. Other significant cost factors included lubricants and tires (13%), and other maintenance costs (4%), all of which should increase as cars are driven on rougher roads.

[17] Toll roads were coded with minimum levels of roughness when they are introduced. Because the fee for using toll roads is generally very small compared to the value of goods or services shipped, I ignore it when measuring transport costs.

[18] These trends are documented visually in Appendix Figures 1.C.1, 1.C.2, and 1.C.3.

1.3.4 Survey of Manufacturing Firms

The estimates of travel times between regions are combined with a plant-level survey: Indonesia's Annual Survey of Manufacturing Establishments (*Survei Tahunan Perusahaan Industri Pengolahan*, or SI). The SI is intended to be a complete annual enumeration of manufacturing plants with 20 or more employees. Administered by the Indonesia's central statistical agency (*Badan Pusat Statistik*, or BPS), the survey is extremely detailed, recording information on plant employment sizes, their industry of operation, cost variables, and measures of value added. Importantly for this work, enumerators recorded each plant's operating location at the kabupaten level, enabling me to link firms to data on transport costs and other location characteristics.[19]

While I use the entire panel of firms to construct measures of spatial concentration and location characteristics, I often treat the data as a repeated cross-section of new firms. In practice, firms in the dataset do not change their kabupaten of residence.[20] New firms are counted when they appear in the dataset having never appeared before. Occasionally firms were not surveyed during their first year of operation, but since enumerators record each firm's starting year, I can accurately time the entry of all firms in the sample.[21] Throughout the analysis, I dropped all firms that were coded as state-owned enterprises (less than 3 percent of all firm-year observations), since these firms are less likely to be governed by market forces.

The SI is also used to construct time-varying location characteristics, including wage rates, commercial land values, and indirect tax rates. A location's wage rate was constructed by taking the median wage rate for all manufacturing workers in that location and time. Commercial land values were taken by averaging the firm's book (or estimated, if book was not reported) value of land capital, then taking the median across all firm-level observations in a given location and year. Land values are difficult to measure in some cases, since only 54% of firm-year observations reported land values, and the lack of precisely estimated rental rates is a major caveat to the results. A location's indirect tax rate is defined as the median share of the value of a firm's output that is spent on indirect taxes, which include establishment license fees, building and land taxes, and sales taxes. Portions of these taxes are set independently by kabupaten governments and vary across space and time.

[19] Throughout the discussion, I use plants and firms interchangeably, because it is likely that less than 5% of plants in the dataset are operated by multi-plant firms (Blalock and Gertler, 2008).

[20] When firms do change locations, it is generally due to a coding error, since they typically switch back to their original location in the next year. Only 15 percent of firm-level observations had multiple kabupaten codes in the raw data, and only 5 percent had two or more kabupaten codes.

[21] Note that the starting year variable was not collected between 2001 and 2005. For some firms appearing in these years, the starting year was taken from the 2006 dataset, but in cases where it could not be obtained, entry is determined by the first year that the plant's unique identifier appears in the panel.

1.4 Trends in Industrial Location

Using these data, I now present reduced form evidence on how the locations of Indonesian manufacturing plants changed in response to changes in transport costs. Lower transport costs raise the profitability of existing cities and may be expected to further intensify agglomerations (Krugman, 1991). On the other hand, by giving firms access to cheaper factors of production, they might encourage firms to disperse (Helpman, 1998). To determine which prediction is more relevant empirically, I first examine how industrial concentration measures evolved over time for different industries. Next, I discuss trends in how new firms located in different types of regions. Finally, I link the changes in observed industrial concentrations to changes in market access.

1.4.1 Measures of Spatial Concentration

From 1985 to 1996, the manufacturing sector in Indonesia was marked by substantial growth in the number of new firms. As firms entered the market, they increasingly moved away from existing agglomerations, reducing industrial concentration across space. The literature provides several measures of industrial concentration, but my main results focus on the Ellison and Glaeser (1997) index.[22] This index measuring the spatial concentration of employment was constructed using plant-level data for every 5-digit industrial classification and year, and Panel A of Figure 1.4 depicts how the mean and median of this index evolved across industries over time.[23] The graph shows a striking reduction in the index, from an average of 0.058 in 1985 to 0.039 in 1996, a fall of over 30 percent.[24] To put this change in perspective, in 1985, the median concentration of manufacturing employment across industries

[22] Because the Ellison and Glaeser (1997) explicitly accounts for industrial concentration, it is useful for analyzing my dataset since many industries are dominated by a small number of large firms, and the plant size distributions change significantly over time, potentially skewing results. However, results using a simpler spatial Herfindahl can be found in Appendix Figure 1.C.4.

[23] Note that in selecting the sample of industries for the analysis, I dropped industries if they were missing too many firm-year observations to construct a consistent measure over time or if they had fewer than 10 firms throughout the period. In a few instances, similar ISIC codes were merged together in order to avoid dropping both from the sample.

[24] The change in average concentration is statistically significant at an $\alpha = 0.05$ level using a two-sample comparison of means ($t = 2.542$, 2-sided $p-$value $= 0.0126$, 1-sided $p-$value $= 0.0063$). Note that these results on dispersion are very different from Sjöberg and Sjöholm (2004), who argue that over the 1980-1996 period, spatial concentration remained more or less unchanged. There are several reasons for discrepancies between my analysis and theirs, but the most important, I suspect, is that I use kabupatens as my spatial unit of analysis, while they use provinces. Since much of the changes I observe take place within provinces, it is not surprising that they find mixed results, while I find evidence pointing towards dispersion. Moreover, I also use a finer level of industrial classification (5-digit ISIC) to compute these indices, and I have access to surveys for every year. Without proper cleaning, these indices are very sensitive to outliers (misreported employment), especially for industries with a smaller number of firms.

in Indonesia (0.044) was 70 percent larger than the U.S.'s median index in 1987 (0.026), according to Ellison and Glaeser (1997). By 1996, Indonesia's median concentration index fell to roughly equal that of the U.S. in 1987.

A variety of industries experienced reductions in concentration over the period. Other Manufacturing (ISIC 39), which includes the production of sporting goods (ISIC 39030) and toys (ISIC 39040), showed the largest reductions in concentration. The furniture and wood products industry (ISIC 33) also experienced dispersion, with major reductions for producers of wood veneer and excelsior (ISIC 33114), home furnishings (ISIC 33230), and handicraft and wood carving (ISIC 33140). Interestingly, textiles (ISIC 32) was the only industry group not to experience any overall reduction in concentration, with the median 5-digit industry experiencing a 10 percent increase in concentration over the period.[25]

However, within industries, there were substantial differences in concentration trends. For instance, while storable processed foods, such as coconut and palm oil (ISIC 31151) and canned, processed seafood (ISIC 31140) dispersed, more perishable food products, such as tofu and tempe (ISIC 31242) and ice (ISIC 31230) remained flat or experienced increases in concentration. One hypothesis suggested by this comparison is that, during a period of large transport improvements, producers of durable goods may experience reductions in concentration, while producers of highly perishable products will remain unaffected. Highly perishable products need to be produced very close to where they are consumed, while more durable goods can be produced farther away, provided that transport costs are sufficiently low.

Moreover, while finished metal, machines, and electronics (ISIC 38) experienced a modest reduction in concentration over the period, manufacturers of radios and television (ISIC 38320) and producers of optical and photographic equipment (ISIC 38500) experienced increases in concentration. These industries are more skill intensive than others and are probably more subject to Marshellian agglomeration economies than other manufacturers. This suggests that while road improvements might induce dispersion for producers of low skill goods, they may not affect industries in which strong external economies are important.

To explore differential trends in concentration measures across different types of industries, I first classified 5-digit industries into either durables or non-durables, based on their reported inventory shares of output.[26] Using these classifications, Panel B of Figure 1.4 depicts how, over the 1990s, both spatial concentration measures fell more rapidly for durable goods (high inventory shares) than for non-durables.[27] This largely confirms our predictions. If transport improvements enable firms to take advantage of cheaper access to land and labor in remote areas, durable goods should be more likely to relocate than non-durable goods,

[25] More detail on the changes in concentration can be found in Appendix Tables 1.C.1 and 1.C.2.

[26] Durable goods industries are classified as those goods whose inventories were, on an average of plant-years, greater than or equal to 10 percent of output, while non-durables have inventory shares less than 10 percent of output.

[27] The difference-in-difference and differential trends estimates are reported in Appendix Table 1.C.4.

since non-durables are perishable and must be consumed in close proximity to where they are produced.

1.4.2 Regional Trends

Another way of exploring the reductions in concentration is to investigate changes in how different types of regions received new plants. Figure 1.5 depicts the shares of new firms locating in cities (defined as of 1990), in kabupatens that are neighbors of cities, in neighbors of neighbors of cities, and in other kabupatens, classified as rural. In 1985, 40 percent of new firms located in cities, but by 1996, only 24 percent of new firms located in cities. Neighbors of cities experienced a 10 percentage point increase in the share of new firms (from 33 percent in 1985 to 43 percent in 1996), while neighbors-of-neighbors experienced a 9 percentage point increase (from 13 percent in 1985 to 22 percent in 1996). Rural shares are mostly flat, however, suggesting that transportation improvements might bring firms to areas near cities, but not to the remotest parts of Indonesia.[28]

As further evidence of the different trends in location across industries, Table 1.3 decomposes the changes of new firms into those resulting from durable goods and from non-durable goods. It is apparent that most of the changes in new firm shares come from durable goods firms, which should be the most responsive to changes in transport costs. For instance, of the 16 percentage point reduction in the share of new firms locating in cities between 1985 and 1996, 10 percentage points is attributable to firms with high inventory shares, while only 6 percent is attributable to firms with low inventory shares.

1.4.3 Regression Analysis

We can summarize the effects of road improvements on the activity of new manufacturing plants by fitting a series of regression functions with region-specific intercepts, exploiting variation in the timing and placement of road improvements across regions in Indonesia. Let $r = 1, ..., R$ index regions (kabupatens), let $j = 1, ..., J$ index industrial sectors, and let $t = 1, ..., T$ index years. Also, define MP_{rt} to be a region's *market potential* in year t (Harris, 1954):

$$MP_{rt} = \sum_{d=1}^{R} \frac{Y_{dt}}{T_{rdt}} \tag{1.1}$$

This is weighted average of each region's GDP, Y_{dt}, where the weights decline in travel times, T_{rdt}. To construct MP_{rt}, I use annual data on real non-oil gross domestic product for each kabupaten and the annual roughness-induced travel time measure of transport costs between kabupaten pairs. As road improvements bring region r closer to other larger markets, MP_{rt} increases. In Section 1.5 of this paper, a similar market potential variable emerges from

[28] Reclassifying kabupatens by physical centroid distance to the nearest 1990 city reveals similar trends.

a model of monopolistic competition and regional trade, and it captures all of the spatial interactions between firms in different regions.

In Table 1.4, we begin by estimating models of the following form:

$$y_{rjt} = \beta MP_{rt} + \varepsilon_{rjt} \tag{1.2}$$

where ε_{rjt} is an error term. The dependent variable, y_{rjt}, is the log of one plus the number of new firms (or new employees) appearing in a region-sector-year cell.[29] In the first column of both panels, we assume that the errors have the following form:

$$\varepsilon_{rjt} = \delta_r + \delta_j + \delta_t + \nu_{rjt}$$

where the ν_{rjt} are assumed to be strictly exogenous conditional on the unobserved effects and a sequence of market potential instruments:

$$\mathbb{E}\left[\, \nu_{rjt} \,|\, \delta_r, \delta_j, \delta_t, z_{r1}^{MP}, ..., z_{rT}^{MP} \,\right] = 0 \tag{1.3}$$

One concern with this regression function is that the outcome variable, y_{rjt}, might be simultaneously determined with local GDP, Y_{rt}, which is included in the construction of market potential, MP_{rt}. To address possible simultaneity bias, I use a base-weighted version of market potential, z_{rt}^{MP}, as instruments for actual market potential. The formula for z_{rt}^{MP} looks just like equation (1.1), except that the GDP weights are fixed to equal regional GDP in 1985 (i.e. $Y_{dt} = Y_{d,1985}$ for all t). Because I partial out all region-specific effects in these regressions, all of the variation in the predicted market potential comes from changes in transport costs. This specification controls for all time-invariant unobservables that influence outcomes for particular sectors and for particular regions, and it also controls for all national unobservables that affect outcomes in each year.

In the second and third columns, we weaken the restrictions on the error term by adding more fixed effects. The second column adds a full set of sector-year effects, controlling not only for time-invariant unobservables that affect regions, but also for any omitted variables that influence outcomes differently in different sectors over time. The third column allows for differential trends in the outcome variable across different provinces. In all specifications, robust standard errors are clustered at the region level, allowing for both serial correlation in the disturbances over time and also for arbitrary correlation between the disturbances affecting different industries in the same region.

Overall, the estimates show significant positive associations between market potential and new manufacturing activity. The dependent variable and explanatory variables are both expressed in logs, so that the coefficients can be interpreted as elasticities. A ten percentage

[29] Dropping observations with zeros does not substantially change estimated effect sizes or confidence intervals; see Appendix Table 1.C.5.

increase in a region's market potential results in an approximately 1.3 percent increase in new firms and a 4.6 percent increase in new employees. The effect sizes are smaller when we allow for province trends, but they remain positive and statistically significant.

Since market potential varies only at the region-year level, the above specifications cannot rule out the possibility that other time-varying, region-specific confounders might actually be driving the results. However, since we have industry-level data, and we know that certain industries (e.g. durable goods producers) are more likely to be influenced by better market access, we can exploit variation *within region-years* in the effects of market potential across sectors. In the fourth column, we estimate models of the following form:

$$y_{rjt} = \gamma \left(D_j \times MP_{rt}\right) + \varepsilon_{rjt}$$

where D_j is an indicator for whether or not the industry is a producer of durable goods (or low-skill goods), and ε_{rjt} is defined as follows:

$$\varepsilon_{rjt} = \delta_{rt} + \delta_{jt} + \upsilon_{rjt}$$

These specifications do not allow us to estimate the entire effect of improving market potential. Instead, they deliver estimates of the differential effect of market potential improvements on durable goods producers, relative to non-durables producers. The coefficients estimates are small, but still significant at conventional levels. Relative to non-durables producers, a ten percent increase in a region's market potential results in a 0.2 percent increase in new durable goods plants and a 0.7 percent increase in jobs for the durable sector.

As a further check on the potential endogeneity of road improvements, I conduct a placebo exercise, estimating the effects of unbuilt sections of the Trans-Java Expressway. If policymakers targeted areas for receiving road improvements based on region-specific, time-varying unobservables that affect firm location choices, then we would expect the unbuilt tollways to have spurious effects on new firms and employment.[30] Table 1.5, Columns 1-3, reports estimates of the effects of these unbuilt toll roads and finds no significant coefficient estimates. Moreover, the estimates of the effects of market potential, controlling for the unbuilt expressway lines, are nearly identical to those reported in Table 1.4.

Another potential problem is that the strict exogeneity assumption, (1.3), does not allow for any feedback between lagged unobservables and future regressors. But if policymakers targeted faster growing areas with better infrastructure, we would expect past unobservables, ν_{rjt-1}, to be correlated with the future history of transport cost variables, $z_{rt}^{MP}, z_{r,t+1}^{MP}, ..., z_{rT}^{MP}$.

[30] Negative project selection is clearly a concern, as it would invalidate the legitimacy of the placebo exercise. I argue in Section 1.7 that the Trans-Java Expressway was not built for idiosyncratic reasons, having more to do with the corrupt way the construction rights were auctioned off than any possible negative selection of the project.

To allow for feedback, I weaken (1.3) to a series of sequential moment restrictions:

$$\mathbb{E}\left[\, \nu_{rjt} \,|\, \delta_r, \delta_j, \delta_t, z_{r1}^{MP}, ..., z_{rt-1}^{MP} \,\right] = 0 \qquad t = 1, ..., T \tag{1.4}$$

This is a weak exogeneity assumption (Chamberlain, 1992), stating that the current values of ν_{rjt} are shocks, uncorrelated with the past regressors; however, the current values of ν_{rjt} are allowed to be correlated with future values of the regressors. It is also a strictly weaker identification assumption; if (1.3) holds, than so does (1.4), but the converse is not true. Sequential moment restrictions open up a variety of possible estimation strategies, but I choose to simply estimate the model in first differences, using lagged changes in market potential IVs $(z_{r,t-1}^{MP} - z_{r,t-2}^{MP})$ as instruments for the current change in market potential $(MP_{r,t} - MP_{r,t-1})$. Results are reported in Table 1.5, Columns 4-6. Although the point estimates are somewhat smaller than before, the estimates are still positive and significant at conventional levels.

1.4.4 Summary

Overall, this analysis indicates that during the sample period, Indonesia has experienced significant reductions in industrial concentration. Areas that received expanded market potential as a result of the road improvement program experienced a growth in manufacturing activity and employment, on average. Moreover, different industries responded to these road improvements in predictable ways. Taken together, this is strongly suggestive evidence in refutation of the predictions of Krugman (1991).

While this reduced form analysis has shed some light on the relevance of different theoretical predictions, it does not allow us to distinguish between different mechanisms driving the results, and it may not be useful for predicting the impacts of different road programs. The estimated regression coefficients were obtained using one specific source of policy variation, and they may not be invariant to different policy regimes. As some regions attract firms because of better market access, this will affect equilibrium factor prices (wages and rents) in those locations, altering the choices of other firms. Predictions of what would happen to firm locations that ignore these general equilibrium factor price responses may be inaccurate. To quantify the relative importance of different mechanisms and provide a richer set of counterfactual predictions that account for general equilibrium effects, in the next section I develop and estimate a structural model of firm location choice.

1.5 Structural Model

In this section, I extend the firm location choice model of Head and Mayer (2004) in several important ways. First, I explicitly allow for multiple industrial sectors, highlighting the importance of sectoral differences in location choice parameters. Next, I also allow for unobserved productive amenities, common to all firms and all industries, that shift marginal cost functions at particular locations. Because the model implies that unobservable amenities will be directly correlated with wages, rents, and other factors influencing marginal costs, identification of the choice model's parameters requires conditional moment restrictions and estimation becomes substantially more involved. Finally, in the original model, firms ignore the effect that their location choices have on wages and rents at chosen locations, but here I allow for upward sloping labor and land supply curves. After presenting the model, I discuss how to estimate its parameters.

1.5.1 Setup

There are R regions, indexed by $r = 1, ..., R$. As in Krugman (1991), there are also two sectors: a constant returns to scale agricultural sector, and an imperfectly competitive manufacturing sector. Each region r is endowed with a mass of workers, $\overline{L}_r$, and workers decide whether to work in agriculture or manufacturing, based on a heterogeneous taste parameter. Workers are perfectly mobile between sectors within a region, but they cannot move between regions. In this sense, the model is short-run, unlike many long-run spatial equilibrium models in urban economics that allow for labor mobility (e.g. Roback, 1982; Busso et al., 2010).[31]

1.5.2 Consumer Preferences

There are two types of goods consumed by individuals: manufactured and agricultural products. Manufactured goods are differentiated products produced in one of K_s industrial sectors, indexed by $k = 1, ..., K_s$. Let $\mathcal{N}_r^k$ denote the set of industry k varieties produced in region r. Consumers in region r choose varieties from each industry and region, and a quantity of the agricultural good, $\mathbf{A}$, to maximize the following utility function:

$$U = \frac{C}{\eta}\left(\prod_{k=1}^{K_s} \mathbf{M}_k^{\mu_k}\right)\mathbf{A}^{1-\mu} \qquad \text{where} \qquad \sum_{k=1}^{K_s} \mu_k + \mu = 1 \tag{1.5}$$

[31] In principle, the assumption of labor immobility can be weakened, for instance if we allow workers to have idiosyncratic tastes for living in particular locations (Moretti, 2010). Crucially, we need some degree of fixity to ensure that factor prices are locally upward sloping.

This utility function represents Cobb-Douglas preferences over both agriculture and CES aggregates of manufacturing varieties for each industry, $\mathbf{M}_k$, which are given by:

$$\mathbf{M}_k = \left[\sum_{d=1}^{R}\left\{\int_{i\in\mathcal{N}_d^k} q^k(j)^{\frac{\sigma_k-1}{\sigma_k}}dj\right\}\right]^{\frac{\sigma_k}{\sigma_k-1}} \qquad \sigma_k \geq 1\ ,\ k=1,...,K_s$$

where $q^k(j)$ is the quantity of industry k's variety j consumed, and σ_k is an industry-specific parameter governing the elasticity of substitution between an industry's varieties. As σ_k tends to 1, varieties in that industry become less substitutable for one another, weakening competition in the industry. As σ_k grows larger, the varieties in industry k become more substitutable, and competition grows more intense.

Note that the utility function contains a scale factor, C/η. While C is just a constant used to normalize the scale of indirect utility, η is a heterogenous taste parameter, reflecting an individual's disutility from working in manufacturing.[32] The scale factor does not factor into worker utility if the individual works in agriculture (i.e. η is set to 1); otherwise, η is continuously distributed over $[1,\infty)$, with c.d.f. $F_r(\cdot)$. A larger draw of η corresponds to a worker who does not like working in manufacturing and must require a larger wage to induce him to switch sectors.

We solve the consumer's optimization problem by first choosing optimal bundles within a given industry and then by determining how to distribute income across industries. Using this approach, it is straightforward to show that in region r, demand for variety j in industry k is given by:

$$q_r^k(j) = \frac{p_r^k(j)^{-\sigma_k}\mu_k\mathbf{Y_r}}{(\mathbf{P}_r^k)^{1-\sigma_k}} \tag{1.6}$$

where $\mathbf{Y}_r$ denotes region r's nominal income, and $\mathbf{P}_r^k$ is given by:

$$\mathbf{P}_r^k = \left[\sum_{d=1}^{R}\left\{\int_{i\in\mathcal{N}_r^k} p_r^k(i)^{1-\sigma_k}di\right\}\right]^{\frac{1}{1-\sigma_k}} \tag{1.7}$$

This represents region r's CES price index for industry k varieties.[33]

1.5.3 Agriculture

Depending on their draws of η, workers decide whether to work in agriculture or manufacturing. The agricultural good is freely traded across locations, and produced under constant returns to scale, with labor as its only factor of production. Hence, a worker's agricultural

[32] The formula for C is given by $C^{-1} = (1-\mu)^{1-\mu}\prod_{k=1}^{K_s}\mu_k^{\mu_k}$.

[33] For a derivation of (1.6) and (1.7), see Appendix Section 1.A.1.

wage is equal to his or her marginal product, and we can normalize $w_A = p_A \equiv 1$, so that the agricultural wage is the numeraire. A worker in region r with taste parameter η will work in manufacturing if and only if:

$$V_{r,M} = \frac{w_r}{\eta \prod_{k=1}^{K_s} (\mathbf{P}_r^k)^{\mu_k}} > \frac{1}{\prod_{k=1}^{K_s} (\mathbf{P}_r^k)^{\mu_k}} = V_{r,A}$$

This implies that the share of workers who opt to work in manufacturing in region r is given by:

$$s_r = Pr\left\{\eta \leq w_r\right\} \equiv F_r\left(w_r\right) \tag{1.8}$$

Hence, the supply of manufacturing workers in region r is given by $L_r = \overline{L}_r s_r = \overline{L}_r F_r\left(w_r\right)$. We assume that $F_r'\left(\cdot\right) \geq 0$ for all $r = 1, ..., R$, so that local labor supply curves are upward sloping.

1.5.4 Manufacturing and Trade

Manufacturing varieties are produced with increasing returns to scale under Dixit-Stiglitz imperfect competition. Conditional on operating in region r, the cost to produce a quantity $q_r^k(i)$ of variety i in industry k is given by:

$$c\left[q_r^k(i)\right] = F_r^k + m_r(i) q_r^k(i) \tag{1.9}$$

where F_r^k represents fixed costs of production in region r. The marginal cost of producing a unit of variety i is given by:

$$m_r(i) = A_{ir} w_r^{\delta_i} \mathrm{r}_r^{\gamma_i} \tag{1.10}$$

where w_r denotes local wages, r_r denotes local rents, and A_{ir} is a cost measure specific to each variety and location. Note that the entire marginal cost function is specific to the industry, operation region, and the variety, not least because the parameters δ_i and γ_i are allowed to vary across varieties.

Because of fixed costs, firms choose a single location in which to produce, shipping their products to all other locations. All firms face industry-specific iceberg transport costs, representing the amount that must be produced in region r in order to deliver one unit of the product to region d (Samuelson, 1954). This is denoted by $\tau_{rd}^k \geq 1$. Due to the transport technology, $(\tau_{rd}^k - 1)$ units of the good "melt away" while being transported, so that only 1 unit is delivered to the destination region. We make three assumptions about transport costs: first, that $\tau_{rr}^k = 1$ for all regions r, so that transport within a region is costless. Second, transport costs are assumed to satisfy a triangle inequality, so that $\tau_{rd}^k \leq \tau_{rs}^k \tau_{sd}^k$ for all $s = 1, ..., R$. This assumption rules out any cross-region arbitrage opportunities in transport. Finally, for simplicity, we assume that the transport cost for industry k is just an industry-specific constant times the travel time measure, T_{rd}:

$$\tau_{rd}^k = \eta^k T_{rd} \qquad \text{for all } \; k = 1, ..., K_s \tag{1.11}$$

Since T_{rd} denotes the travel time based on the quality of road infrastructure, this assumption allows for the products of industries to melt away at different rates while their goods are being shipped between locations.[34]

The form of the consumer's utility function implies that all consumers in all locations consume every variety of every industry. Conditional on locating in region r, a firm in industry k has gross profits (ignoring fixed costs) that are equal to the sum of profits obtained from shipping its output to consumers in all destination locations:

$$\Pi_r^k(i) = \sum_{d=1}^{R} \pi_{rd}^k(i) = \sum_{d=1}^{R} \left(p_{rd}^k(i) - m_r(i)\tau_{rd}^k\right) q_{rd}^k(i)$$

Firms are operating under Dixit-Stiglitz monopolistic competition, and they choose prices ignoring their effects on regional industry price indices, $\mathbf{P}_r^k$. From the structure of competition and consumer demands, we can show that the firm's optimal pricing formula is given by:

$$p_{rd}^k(i) = \left(\frac{\sigma_k}{\sigma_k - 1}\right) m_r(i)\tau_{rd}^k \tag{1.12}$$

This expression implies a mill pricing strategy, as $p_{rd}^k(i) = \tau_{rd}^k p_r^k(i)$.[35] Moreover, prices are just industry-specific markups over the firm's marginal cost, with the size of the markup is governed by the size of σ_k, the elasticity of substitution.

Note that a firm's gross profits from locating in region r and shipping to region d are given by:

$$\pi_{rd}^k(i) = \left(p_{rd}^k(i) - m_r(i)\tau_{rd}^k\right) q_{rd}^k(i)$$

Plugging in expressions for consumer demand (1.6), transport costs (1.11), and optimal pricing (1.12), we can rewrite this expression as:

$$\pi_{rd}^k(i) = \gamma_k \left(m_r(i)\right)^{1-\sigma_k} \mathbf{Y}_d \left(\frac{\mathbf{P}_d^k}{T_{rd}}\right)^{-(1-\sigma_k)}$$

where γ_k is a constant specific to industry k, and $\mathbf{P}_d^k$ denotes the price index for industry k's

[34] In principle, the relationship between travel times, T_{rd}, and transport costs, τ_{rd}^k, could be calibrated, for example by using international trade flow data (Head and Mayer, 2004).

[35] Here, $p_r^k(i) \equiv \sigma_k m_r(i)/(\sigma_k - 1)$ is the firm's local price, just a simple markup over marginal cost. A derivation of (1.12) can be found in Appendix Section 1.A.2.

products consumed in region d.[36]

Summing across destination locations, we obtain the firm's total gross profits from locating in region r:

$$\Pi_r^k(i) = \gamma_k \left(m_r(i)\right)^{1-\sigma_k} \left[\sum_{d=1}^{R} \mathbf{Y}_d \left(\frac{\mathbf{P}_d^k}{T_{rd}}\right)^{-(1-\sigma_k)}\right] \tag{1.13}$$

This expression tells us that a firm's profits from operating in region r depend an industry-specific constant, γ_k, marginal costs, as well as the expression in brackets which is defined as the industry-specific *real market potential*:

$$RMP_r^k \equiv \sum_{d=1}^{R} \mathbf{Y}_d \left(\frac{\mathbf{P}_d^k}{T_{rd}}\right)^{-(1-\sigma_k)}$$

This is a weighted sum of regional incomes, where the weights decline in transport costs and increase in the price index for that specific industry. In this model, market potential is the single variable that captures all of the spatial interactions between firms in different locations. It links firm profits from locating in region r to transport costs between that region and all others. As a location becomes closer to larger demand markets, RMP_r^k increases.

The industry-specific real market potential is closely related to another variable, nominal market potential, discussed in an older literature on economic geography and regional science (Harris, 1954):

$$NMP_r = \sum_{d=1}^{R} \left(\frac{\mathbf{Y}_d}{T_{rd}}\right)$$

The difference between NMP_r and RMP_r^k is that real market potential explicitly accounts for competition, through the inclusion of price indices.[37]

I use real non-oil gross domestic product data to proxy for $\mathbf{Y}_d \left(\mathbf{P}_d^k\right)^{-(1-\sigma_k)}$. To the extent that locally, gross domestic production is not equal to domestic incomes or that the statistical agencies are not using price indices that match those from the theory, the market

[36] The exact form of the constant γ_k is given by:

$$\gamma_k = \frac{1}{\sigma_k} \left(\frac{\sigma_k \eta^k}{\sigma_k - 1}\right)^{1-\sigma_k} \mu_k$$

This constant is depends on the industry's elasticity of substitution, transport cost parameters, and Cobb-Douglas budget share parameters for industry k.

[37] In the formula for real market potential, the price index can be thought of as a measure of the intensity of competition. Lower price indices correspond to locations with lower markups and fiercer competition, while higher price indices correspond to larger markups and weaker competition. Firms in industry k want to locate in regions that are closer to larger markets, but this preference is tempered by the competitiveness of those locations, reflected in the price indexes.

access variable used in the estimation will be mis-measured. In some specifications, I also allow the data to predict the relationship between travel times and $(\tau_{rd}^{k})^{(1-\sigma_k)}$.

1.5.5 Firm Location Choices

Firms locate in region r if and only if their expected operating profits minus fixed costs from operating in region r are greater than those of all other locations. Following Head and Mayer (2004), we assume that the fixed cost of locating in region r for a firm operating in industry k, F_r^k, is the same across all locations, i.e. $F_r^k = F^k$ for all $r = 1, ..., R$. Given this assumption, fixed costs do not play any role in location choices and can be ignored.

Define $V_r^k(i)$ to be firm i's *value function* for region r, a simple transformation of operating profits minus fixed costs:

$$V_r^k(i) \equiv \frac{\ln \Pi_r^k(i) - \ln \gamma_k - F^k}{\sigma_k - 1} = \frac{1}{\sigma_k - 1} \ln RMP_r^k - \ln(m_r(i))$$

Taking logs of (1.10), we have:

$$\ln(m_r(i)) = \delta_i \ln(w_r) + \gamma_i \ln(\mathrm{r}_r) + \ln(A_{ir})$$

Assuming that we can decompose the idiosyncratic portion of the cost function into a vector of observable cost shifters, c_r, a single unobserved component, ξ_r, and a firm-location specific error term, ε_{ir}, we can write:

$$\ln(A_{ir}) = c_r'\theta_i - \xi_r - \varepsilon_{ir}$$

It is useful to think of ξ_r as an unobserved productive amenity (e.g. average ability of the workforce, or quality of life in region r), which shifts marginal costs for all firms and all industries. The term ε_{ir} is an idiosyncratic, firm and region specific component of marginal costs, which we further assume is distributed i.i.d. type 1 extreme value across locations for each firm. Collecting all of the observable cost shifters into a single $(K \times 1)$ vector, $x_r = (\ln(w_r), \ln(\mathrm{r}_r), c_r')'$ and the idiosyncratic technology parameters into another $(K \times 1)$ vector, $\beta_i = (\delta_i, \gamma_i, \theta_i')'$, we can rewrite the log of marginal costs as:

$$\ln(m_r(i)) = x_r'\beta_i - \xi_r - \varepsilon_{ir}$$

Define D_i to be a $(L \times 1)$ vector of firm-specific observables, for example, a full set of indicators for whether or not firm i operates in particular industries. Also, let v_i^k denote a random valuation component for $x_{r,k}$, the k-th element of the vector x_r. More precisely, v_i^k is firm i's idiosyncratic sensitivity to marginal cost variable k, which we assume is normally distributed across firms and scaled to have zero mean and unit variance. Also, define $\alpha^k = 1/(\sigma_k - 1)$.

Using this notation, we can write the firm's *value function* as:

$$V_r^k(i) = \alpha_i \ln RMP_r^k - x_r'\beta_i + \xi_r + \varepsilon_{ir} \tag{1.14}$$

where

$$\alpha_i = \overline{\alpha} + \sum_{l=1}^{L} \pi_{\alpha,l} D_{i,l} + \overline{\sigma}_\alpha v_i$$
$$\beta_{i,k} = \overline{\beta}_k + \sum_{l=1}^{D} \pi_{k,l} D_{i,l} + \overline{\sigma}_k v_i \qquad k = 1, ..., K$$

In this setup, $\pi_{k,l}$ is a coefficient measuring how $\beta_{i,k}$ varies with firm characteristics, while $\overline{\sigma}_k$ represents the standard deviation of firm valuations for $x_{r,k}$.

Given this setup, we can write the value (or transformed operating profits) a firm gains from choosing location r as follows:

$$\begin{aligned} V_{ri} &= \alpha_i \ln RMP_r^k + \sum_{k=1}^{J} x_{r,k}\beta_{ki} + \xi_r + \varepsilon_{ir}^k \\ &= \left\{ \overline{\alpha} \ln RMP_r^k + \sum_{k=1}^{J} x_{r,k}\overline{\beta}_k + \xi_r \right\} \\ &+ \left\{ \sum_{l=1}^{D} (D_{i,l}\pi_{\alpha,l} + \overline{\sigma}_\alpha v_i) \ln RMP_r^k + \sum_{k=1}^{K} \left(\sum_{l=1}^{D} (D_{i,l}\pi_{k,l} + \overline{\sigma}_k v_i)\, x_{rk} \right) \right\} + \varepsilon_{ir} \\ &= \delta_r + \mu_{ri} + \varepsilon_{ir} \end{aligned}$$

The first term in this expression, δ_r, is the mean valuation of choosing location r and is common to all firms in all industries. It depends on $(\overline{\alpha}, \overline{\beta'})'$, the mean technology parameters, as well as ξ_r, the unobserved productive amenity. The second term, μ_{ri}, represents mean-zero heteroskedastic deviations from the mean valuation, capturing the effects of the sectoral differences. Firm i in industry k chooses to operate in location r if $V_r^k(i) > V_d^k(i)$ for all other locations, d. This implicitly defines the set of observed and unobserved variables that lead to the choice of location r. Formally, we can denote this set by A_r:

$$A_r = A_r\left(\mathbf{x}, \xi_r, \delta_{\cdot}\,; \theta_2\right) = \left\{ \left(D_i, v_i, \varepsilon_{ir}\right) \middle| V_r^k(i) \geq V_d^k(i)\ \forall\ d = 1, ..., R \right\}$$

1.5.6 Identification of the Choice Model

In an ideal experiment for studying firm location choices, we would randomly assign locations with factor prices, infrastructure access, and exogenous geographic features, and we would record firms' location choice responses. However, in observational studies, market access and other cost shifters are not randomly assigned, and instead reflect a host of factors, such as the availability of commercial land for real estate, local supplies of labor and consumers, and other characteristics unobserved to researchers. Unobserved productive amenities will raise the profitability of locating in certain regions, which, ceteris paribus, increases the number of workers and firms who locate in certain regions, raising incomes. Hence, the model implies that market access, wages, and rents will be directly correlated with unobserved productive amenities. This necessitates the use of instrumental variables: variables that are correlated with the endogenous choice characteristics but uncorrelated with omitted factors explaining the choices of firms.

Distinguishing between between omitted factors, such as natural advantages, and other theories in understanding why agglomerations form is a classic identification problem in empirical urban economics (Ellison and Glaeser, 1999). While cross-sectional instruments are clearly useful, finding them is challenging and their exclusion restrictions are often difficult motivate. However, if unobserved natural advantages are constant over time, the use of panel data and fixed effects can help us distinguish between natural advantages and transport cost theories.

Panel data is useful for another reason: if firm cost-functions are time-invariant, it makes sense that as location characteristics change, with increases or decreases in wages, rents, and market access, the identifying power of our model improves. Although the parameters of the model could, in principle, be estimated from data on a single cross-section, with firms making only one choice, such an approach seems far removed from the ideal experiment of repeatedly assigning locations with different bundles of characteristics and observing responses (Nevo, 2000). Nevertheless, in most applications of discrete choice to location decisions, authors only study a cross-section of firm choices.

To improve the identifying power of the discrete choice model, I estimate the parameters using variation in location characteristics over time.[38] Abusing notation, collect all of the choice characteristics for location r at time t as $\mathbf{x}_{rt} = [\ln RMP_r, x'_{rt}]'$, and let $\boldsymbol{\beta}_i = (\alpha_i, \beta'_i)'$ collect all of the choice parameters. With multiple time periods, firm i's value function for location r at time t is the following:

$$V_{irt} = \delta_{rt} + \sum_l (D_{il}\boldsymbol{\beta}_l + v_i\sigma)'\mathbf{x}_{rt} + \varepsilon_{irt}$$

[38] One serious objection to this approach is that over time, firm technologies are changing, and to the extent that this is the case, panel data are not helpful. Note that this is also a problem in the literature on estimating production functions (e.g. Olley and Pakes, 1996). While I cannot rule out this possibility, Wie (2000) (and work cited therein) suggests that Indonesian manufacturing in the 1990s and early 2000s is characterized by a strong absence of technical progress. This is one reason why post-crisis manufacturing growth has been so slow.

where the mean valuation terms, δ_{rt}, are given by:

$$\delta_{rt} = \mathbf{x}'_{rt}\overline{\boldsymbol{\beta}} + \xi_r + \xi_t + \nu_{rt}$$

Here, ξ_r represents any time-invariant unobserved productive amenity for region r (e.g. favorable geography). The term ξ_t represents an aggregate time effect, separate for all urban or non-urban areas at year t. The term ν_{rt} can be thought of as an unobserved, time-varying productivity shock specific to location r at time t.

I make use of two different conditional moment restrictions on ν_{rt} to identify the choice parameters. The first is similar to a strict exogeneity condition in linear panel models (Chamberlain, 1984):

$$\mathbb{E}\left[\,\nu_{rt}\,|\,\xi_r, \xi_t, x_{r1}, ..., x_{rT}, z_{r1}^{MP}, ..., z_{rT}^{MP}\right] = 0 \tag{1.15}$$

In words, this restriction says that once we condition on the unobserved fixed factor, ξ_r, the productivity shocks are uncorrelated with the entire history of the location characteristics, $x_{r1}, ..., x_{rT}$, and the history of market potential instruments, $z_{r1}^{MP}, ..., z_{rT}^{MP}$. As in Section 1.4, I use market potential with fixed 1985 output weights, so that all of the variation in the predicted MP_{rt} comes from changes in transport costs. Making use of this restriction is a large improvement over existing work, but in practice it may not always hold. For instance, if policymakers were targeting more productive areas with better infrastructure, we would expect past productivity shocks, $\nu_{r,t-1}$, to be correlated with future market access, $x_{rt}, x_{r,t+1}, ..., x_{rT}$.

Motivated by these dynamic targeting concerns, a second conditional moment restriction relaxes the first:

$$\mathbb{E}\left[\,\nu_{rt}\,|\,\xi_r, \xi_t, x_{r1}, ..., x_{r,t-1}, z_{r1}^{MP}, ..., z_{r,t-1}^{MP}\right] = 0 \tag{1.16}$$

This is a weak exogeneity moment restriction (Chamberlain, 1992), stating that current productivity shocks are innovations, uncorrelated with all previous realizations of the x_{rt}'s and z_{rt}^{MP}'s. Note that this is a strictly weaker identifying assumption than (1.15), and if (1.15) holds, than so does (1.16).

1.5.7 Estimation of the Choice Model

The assumption on the joint distributions of v^s and ε_{irt} gives rise to an expression for the conditional probability that firms with i characteristics choose location r at time t:

$$P_{irt} = \int \frac{\exp\{\delta_{rt} + \sum_{k=1}^{K} x_{rt}^{k}\left(\overline{\sigma}_k v_i + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}}{1 + \sum_{d=1}^{R} \exp\{\delta_{dt} + \sum_{k=1}^{K} x_{dt}^{k}\left(\overline{\sigma}_k v_i + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}} dF\left(v^s\right) \tag{1.17}$$

where the value from choosing the outside option is normalized to zero in each period.[39]

I estimate the choice model using a two step procedure. In the first step, I estimate the δ_{jt}'s and θ_2 using maximum simulated likelihood. Although a full search over the δ_{jt}'s and θ_2 is possible, in practice, because of the large number of locations in the dataset and the multiple years over which those locations are observed, it is computationally difficult. Consequently, I maximize the simulated likelihood function only over θ_2. For each value of θ_2, I choose $\delta_{jt} = \delta_{jt}(\theta_2)$ to ensure that the mean valuation components satisfy a market share constraint (Berry, 1994).[40]

In the second step, to recover the linear parameters, I fit the following regression function, making use of conditional moment restrictions (1.15) and (1.16):

$$\widehat{\delta}_{rt} = x'_{rt}\beta + \xi_r + \xi_t + \upsilon_{rt} + \nu_{rt}$$

where $\upsilon_{rt} \equiv \widehat{\delta}_{rt} - \delta_{rt}$ denotes measurement error. Specific details, such as how to compute the gradient in the maximum likelihood step and how to work out standard errors, correcting for the fact that the $\widehat{\delta}_{rt}$'s are estimated, can be found in Appendix Section 1.B.

1.6 Results

1.6.1 Constant Coefficient Logit Results

Table 1.6 presents results from estimating a constant coefficient version of the random coefficients logit model. This effectively sets σ and π equal to zero in (1.17), and the mean technological parameters are estimated from linear regression (Berry, 1994). The exact form of the linear regression the following:

$$y_{rt} \equiv \ln\left(s_{rt}\right) - \ln\left(s_{0t}\right) = x'_{rt}\beta + \xi_r + \xi_t + \varepsilon_{rt}$$

where s_{rt} is the share of new firms in year t who locate in region r, and s_{0t} is the share who choose the outside option of locating in other regions in Indonesia.[41]

This specification is used to highlight some aspects of the methodology and contrast it with that used in prior work. Columns 1 and 2 present estimates of the mean technology

[39] Note that because I do not observe new firms in every location at every time period, the outside option (roughly, locating outside of kabupatens on Java, Sumatra, and Sulawesi) changes across years. However, the outside option is chosen on average by 6.1 percent of entrants in a given year, and it is never chosen by more than 10 percent of firms.

[40] This two-step estimation procedure is similar to that used in Langer (2010) in studying demographic preferences for new vehicles, although that study uses second-choice data.

[41] Locations in the model consist of all districts on Java, Sumatra, and Sulawesi, and the outside share consists of locations in the outer provinces (Bali, Kalimantan, Maluku, and Irian Jaya). On average, the outside share was chosen by approximately 6 percent of new firms.

parameters for a single cross-section of firms, here using all firms appearing in the 1990 survey to construct market shares. Column 1 includes no other control variables, while Column 2 adds several fixed controls (e.g. elevation, ruggedness, type of land). Though not always statistically significant, the signs on rent variables in columns 1 and 2 are positive, suggesting that firms are more profitable when they locate in places with larger land costs. These positive coefficients are not exclusively a feature of my dataset; for instance, Head and Mayer (2004) find significant positive wage coefficients in many of their specifications predicting the locations of Japanese car manufacturers in Europe. The problem is that the wage and rent variables are correlated with unobservable productive amenities, and without access to richer panel data, estimation on a single cross-section of firms cannot recover accurate parameter estimates. This is the same problem observed by (Berry et al., 1995) in their study of consumer demand for cars; a conditional logit gives a positive relationship between prices and demand, but this is because prices are correlated with unmeasured product quality.

Columns 3-8 use the entire panel of locations (from 1990-2005), market shares are constructed using new firms only, and market potential is instrumented using z_{rt}^{MP}, which is a market potential variable with fixed 1985 GDP weights. Estimation proceeds using 2-step GMM, and all specifications include location fixed effects and rural-urban year dummies, which should control for any time-invariant unobservable productive amenities, as well as any unobserved productive amenities that are common across rural and urban locations in each year. Robust standard errors are clustered at the location level, which allows for arbitrary serial correlation in the errors for each region over time (Arellano, 1987). Column 3 shows that the wage coefficient, which was previously imprecisely estimated, is now negative and statistically significant. Coefficients on rents are also negative and significant. The coefficient on market potential is large and statistically significant, and the ratio of the factor price and market potential coefficients suggests that firms would be willing to accept a 7.6 percent wage increase or a 14.4 percent rent increase for a 1 percent increase in a location's market potential.

In Column 4, I allow the effect of distance in the market potential variable to vary non-linearly. Recall that the market potential variable used in the analysis, MP_{rt}, was defined as

$$MP_{rt} = \sum_{d=1}^{R} \frac{\mathbf{Y}_{dt}}{T_{rdt}}$$

where $\mathbf{Y}_{dt}$ is real GDRP for region d at time t, and T_{rdt} is the roughness-based transport cost measure, measured as the travel time (in hours) between locations r and d at time t. In Column 4, I use a market potential variable that is defined differently:

$$\widetilde{MP}_{rt} = \sum_{d=1}^{R} \left(\frac{\mathbf{Y}_{dt}}{f\left(T_{rdt}\right)} \right)$$

where

$$f\left(T_{rdt}\right) = \delta_0 + \delta_1 T_{rdt} + \delta_2 T_{rdt}^2 + \delta_3 T_{rdt}^3$$

Estimation of this specification proceeds by using non-linear least squares. Column 4 shows large, statistically significant coefficients for a third-order polynomial. The implied distance-function and pointwise confidence bands are depicted graphically in Figure 1.7. From this figure, it appears that markets within 5-6 hours of reach are not discounted very heavily, but as travel times increase beyond 5-6 hours, the discounts grows rapidly. Thus, when choosing locations, firms seem to care much more about their access to nearby markets than they do about accessing farther away markets.

In Column 5, I replace the market potential variable with the (log) density of paved roads.[42] Road density is used frequently as a proxy for the quality of local infrastructure, and its coefficient is sometimes interpreted to reflect market access. Although both measures are positively correlated, when included in Column 5, the point estimate on road density is small and is not statistically significant. Examining the impact of road density on the locations of a single cross-section of Indonesian manufacturers, (Deichmann et al., 2005) find a similar non-result for eight out of the fifteen industries they studied. They claim that finding "suggest[s] that improvements in transport infrastructure may only have limited effects in attracting industry". Another possibility is that road density is a poorly measured version of market access, or that whether or not roads are paved is not as important as how rough they are, and whether they can be travelled over quickly.

To the extent that other infrastructure improvements were occurring at the same time as the road improvements, my estimates might be biased, picking up more than they should. In Column 6, in addition to the market potential variable, I include as a dependent variable the log of the median percentage of electricity consumed by firms in the region that is produced by the state electricity company, *Perusahaan Listrik Negara* (PLN). Electricity provision was improved dramatically over the sample period, but the coefficients on this variable, while large, are only significant at the 10 percent level. Moreover, the coefficient on market potential is only attenuated slightly when including electricity provision, suggesting that the electricity effects do not overwhelm the market potential effects. In Column 7, when we add the variable for indirect taxes, which has large effects and is statistically significant, the coefficient on electricity is no longer significant. Column 8 reports the preferred specification.

Column 9 adds a full set of province-year effects to the model, so that all of the variation in the explanatory variables comes from regional variation in wages, rents, taxes, and market access for a given province year. It is reassuring that coefficient estimates on wages, rents and taxes are largely similar, but the coefficient estimate on market potential nearly doubles in size. In Column 10, I estimate the model using the weak exogeneity moment restrictions. The rent coefficient is no longer significant, and the effects of market potential double (as well as their standard errors).

[42] The density of paved roads is measured as total km of paved roads per 100 km^2 of land.

1.6.2 Random Coefficient Logit Results

Table 1.7 displays results from the two-step BLP estimation on the full dataset of 17,684 new firms choosing one of over 100 locations over a 15 year period. The reported model is parsimonious: I've only included mean effects for wages, rents, and taxes, and interaction terms for the market potential. Overall, for six of the seven industrial categories, we find that market potential had a positive and statistically significant effect on location choice. The only industrial category that does not have a positive market potential effect is wood products (ISIC 33), which is most likely due to the fact that producers of wood products typically locate very close to forests, their sources of raw materials.[43] The largest coefficient was for other products, which were also the most likely to have experienced dispersion over the period (see Section 1.4).

Heterogeneity across firms in their willingness to substitute better market access for lower wages (or rents) is readily apparent from the positive standard deviation coefficient. The estimates of the mean parameters (and the single standard deviation estimate) imply that 99 percent of other products producers have positive valuations for market potential, while only 54 percent of wood products producers have positive valuations for market potential. Over 80 percent of each of the other industries had firms with positive market potential valuations.

Another way of evaluating the fit of the model is to determine whether or not its implied substitution patterns are reasonable. Locations that are more substitutable for one another should have stronger cross-elasticities, while those that are less substitutable should have smaller cross-elasticities. Define the *cross-market potential elasticity* between location k and j at time t, denoted $\eta^{MP}_{jk,t}$, as follows:

$$\eta^{MP}_{jk,t} = \frac{-\partial \log s_{jt}}{\partial \log MP_{kt}}$$

This elasticity tells us the percentage decrease in the share of new firms choosing location j that would result from a one-percent increase in market access for location k. We would expect $\eta^{MP}_{jk,t}$ to be positive; increasing market potential in location j should decrease firms' demand for location k. We would also expect that this elasticity would be larger for locations that are closer together physically, or in terms of various characteristics (e.g. GDP levels, population). For instance, if market access is improved in Jakarta, we would expect location shares of nearby regions in Western Java to be reduced more than locations that are farther away (i.e. remote kabupatens on Sulawesi).

Overall, estimates of the median cross-market potential elasticity across firms are positive

[43] This somewhat reconciles the fact that while wood producers were one of the largest firms to experience dispersion in the dataset, but we only found modest reduced form impacts of market potential on the locations of new firms. Better roads clearly do not explain all of the dispersion.

almost everywhere. In only 251 of the 33,241 location *i*-*j* pairs were the median cross elasticities positive, and in those cases they were extremely small.[44] To summarize the relationship between cross-market potential elasticities and various location characteristics, we estimate linear regressions in Table 1.8. These regressions take the following form:

$$\eta_{jk}^{MP} = \gamma_j + \gamma_k + \beta D_{jk} + \varepsilon_{jk}$$

where D_{jk} is a variable measuring the distance between j and k on some particular characteristic, and γ_j and γ_k are location-specific intercepts. The reported regression results show strong negative relationships between η_{jk}^{MP} and physical distance, differences in 1985 population levels, and differences in 1985 GDP levels. Hence, as locations grow closer together to one another along several dimensions, the cross market potential elasticity between those locations becomes larger. This suggests that our model is delivering the sort of rich substitution patterns that we would expect.

1.7 Counterfactual Simulations

One advantage of estimating the structural parameters of the model is that it can now be used to predict what would have happened to industrial locations had different road improvement programs been undertaken. The first counterfactual simulation involves the on-time construction of the Trans-Java Expressway, a planned road program that has yet to be fully completed. I contrast results from this simulation with those from implementing a rural roads program, designed to upgrade and improve highways in more remote parts of Java, Sumatra, and Sulawesi.

1.7.1 Overview: Trans-Java Expressway and Rural Road Upgrades

The Trans-Java Expressway was planned in the early 1990's under Suharto, as part of Repelita V. A map of all sections of the proposed expressway is depicted in Figure 1.8; finished sections are depicted in thick black lines, while unfinished sections are in red. The expressway was intended to be a contiguous tollway spanning approximately 1,100 km, linking Jakarta to Surabaya along Java's densely populated North coast. This would strengthen the connection between major cities along the coast, providing high-speed access and allowing for much faster transport.

Instead of tendering the expressway as a single contract, Suharto divided the project into 18 separate concessions, and in an episode emblematic of the corrupt practices of his regime,

[44] Note that the unit of analysis is the median cross-elasticity, where the median is taken over each of the years in the dataset. That there were a total of 183 locations, giving us 33,489 pair observations, since it is not necessarily the case that $\eta_{jk}^{MP} = \eta_{kj}^{MP}$. However, in some cases a location was not observed in the same year as others, so that our total location pairs in the analysis are 33,241.

auctioned off those concessions to companies owned by his friends and family.[45] During the Asian Financial Crisis, construction was suspended and many of the companies that held concessions to build different sections of the Trans-Java Expressway collapsed into default. Concessions often passed to different owners, creating construction delays. Difficulty in acquiring the land to build these roads and reduced state power to enforce eminent domain have also slowed progress, especially in the post-Decentralization period (Davidson, 2010a).[46] To predict what would have happened to industrial locations had this highway actually been built, I first construct the tollway in 1994 (when it was supposed to have been finished) and then recalculate transportation costs between regions.

This type of road program, involving the connection of major cities, is very different from programs that aim to improve rural roads. Since it is likely that firm locations will be affected differently by different types of road programs, I contrast the predictions of the Trans-Java Expressway simulation with those of a project that upgrades roads in rural areas. This involves bringing over 11,000 km of roads in rural kabupatens up to the average roughness levels for roads of the same function class.[47] After the roads are improved in the dataset, transportation costs are recalculated between regions as before.

With these two different counterfactual transport cost matrices, I make predictions for firm location choices using three different techniques. As a baseline, I first make predictions of what would have happened using a simple reduced form regression of firm location choice on market access. Next, I use the model to provide an upper bound on industrial relocation, essentially ignoring the general equilibrium factor price responses. Finally, I make predictions using the full structural model.

Note that when I conduct these simulations, I make two crucial simplifying assumptions. The first is that the process of entry is exogenous; the same set firms that actually entered over the 1994-2005 period also enter in the counterfactual simulations. One could imagine that large road improvement programs could affect entry directly, but my model and estimating framework are not equipped to allow for this possibility. Hence, the results can only speak to a reshuffling of existing firms between locations over time.

Additionally, I also assume that the share of new firms who choose to locate outside of Java, Sumatra, and Sulawesi, s_{0t}, remains unchanged during the counterfactuals. This may seem innocuous, but it has implications. For instance, if the Trans-Java Expressway is constructed, this will lower transport costs to the affected regions, raise every location's

[45] For instance, the rights to build a 35 km section connecting Kanji to Pejagan were sold to PT Bakrie & Brothers. Aburizal Bakrie had old ties to Suharto, starting in the late 1970s with several joint ventures, including major real estate projects, rubber plantations, and a scheme involving illicit sales of Pertamina's crude oil to foreign investors.

[46] Moreover, some concessionaires may never have intended to build and only hung on to their rights so that they could be resold. For instance, Davidson (2010b) explains how the concession for a 116 km section between Cikampek and Palimanan was successfully flipped to a Malaysian company in 2008 by a group of investors, including then vice-president Jusuf Kalla.

[47] For more on road classifications in Indonesia, see Appendix Section 1.C.1.

market potential, and improve profitability everywhere. If we allowed s_{0t} to change, it would fall rapidly and all "inside" location shares, s_{rt}, would increase. The problem with this is that it ignores the fact that market potential in the outside locations will also be improved, raising the profitability of choosing the outside option. Because the characteristics affecting the choice of the outside option are not explicitly specified or used in the choice model, there is no way to allow for this sort of response. Although I cannot rule out the possibility that some firms who chose the outside option would now choose to locate in Java, Sumatra, and Sulawesi, again the model is not suited for examining this.

1.7.2 Reduced Form Prediction

For a baseline prediction for the location choices that would have resulted from new road improvements, I estimate a simple linear relationship between log market potential and the inverted market shares (Berry, 1994):

$$y_{dt} = \alpha_d + \alpha_t + \beta MP_{dt} + \varepsilon_{dt}$$

After estimating this relationship, I predict what would have happened if market potential were constructed using current GDP weights but new transport costs, T^c_{odt}:

$$MP^c_{ot} = \sum_{d=1}^{r} \frac{Y_{dt}}{T^c_{odt}}$$

This reduced form prediction ignores several features predicted by the full model. First, changes in transport costs will change firms' location choices. When firms move to new locations, they will produce different levels of output, and hence the equilibrium GDP weights in MP^c_{ot} will change; here, we fix weights to their actual levels. Second, new firm locations will shift factor prices in different regions, and these responses are ignored. Nevertheless, this reduced form prediction provides a benchmark that I can use to compare with predictions based on the model.

1.7.3 Model-Based Upper Bound

If we ignore factor price changes, firms will move to areas with better market access, but they won't suffer the consequences of higher production costs. Hence, a model-based prediction that ignores factor price responses should provide an upper bound on the home-market effect induced agglomeration caused by road improvements. Under this scenario, we should see maximal increased industrial concentration as a result of the road improvements. To implement this prediction, I do the following:

Step 1: Take draws for the current simulated history, $\varepsilon_{ijt} \sim EV(1)$ and $v^s \sim N(0,1)$. Note that firms get individual, independent draws for ε_{ijt}, but the industrial draws of v^s are the same for each industry throughout all years of the current simulation.

Step 2: For a given year $t = 1994, ..., 2005$, I construct a starting value ($s = 0$) counterfactual market potential, using the current output and simulated transport cost measures.

$$MP^0_{rt} = \sum_{d=1}^{r} \frac{\mathbf{Y}_{dt}}{T^c_{rdt}}$$

Step 3: Given the simulated draws, counterfactual market potential, and current factor prices, I predict new location choices.

Step 4: Next, conditional on MP^0_{rt}, I use the model to predict each firm's output at their newly chosen location, q^c_{it}.[48] Firms' new outputs will be larger in higher market potential locations, but lower in places with higher factor costs.

Step 5: After predicting firms' counterfactual outputs, we construct new output weights for each location by adding the total output for the set of firms who choose that location, $\mathcal{C}_d$, to that location's actual output and subtracting the lost outputs from the set of firms who move away, $\mathcal{D}_d$:

$$\mathbf{Y}^c_{dt} = \mathbf{Y}_{dt} + \sum_{i \in \mathcal{C}_d} q_{dt}(i)^c - \sum_{i \in \mathcal{D}_d} q_{dt}(i)^c$$

This gives us a new counterfactual market potential for each location.[49]

Step 6: Using the new market potential variable, we feed this into Step 2 and repeat steps 2-5 until "convergence". For this exercise, I stop when less than 5 percent of firms have chosen different locations than were chosen in the previous iteration.

An attractive feature of this simulation is that all of the required parameters have been estimated from the choice model; there is no need for additional calibration. Since the model I develop may have multiple equilibria for certain parameter values, this algorithm for equilibrium selection will typically choose an equilibrium that is close to the actual one.

1.7.4 Full Structural Prediction

This is the same as above, except that we use the model to obtain expressions for factor demands. After specifying labor and land supply functions, we can recompute factor market equilibria after firms relocate. Hence, we insert a step in the algorithm from Section 1.7.3 as follows:

[48] For precise details on how this is done, see Appendix Section 1.A.3.

[49] Note that in this step, I am using output and income interchangeably, assuming that larger firm outputs correspond to larger worker incomes.

Step 4A: Conditional on the current iteration's market potential, MP^0_{rt}, and factor prices $(w^0_{rt}, \mathrm{r}^0_{ot})$, we predict each location's new wages and rents. From the model's cost function, it is easy to show that firm i's demand for labor is given by:[50]

$$L^* \left(q_r(i), w, \mathrm{r}\right) = \alpha_i A_{ir} w^{\alpha_i - 1} \mathrm{r}^{\beta_i}$$

After adding up individual labor demands at each location, we equate the location's demand for labor with local labor supply and solve for new equilibria. In practice, we use a Taylor approximation to linearize firms' individual labor demands, so that they can be computed and added together rapidly. A location's labor supply is specified as:

$$L^s_r = \gamma_r + \eta w_r$$

We try different values of $\eta = 1, 0.5$, and for each of these values, we choose γ_r so that initially, the labor market is in equilibrium. In theory, we could also solve for equilibrium land prices in each iteration, but because there are so many missing land values, we simply use a reduced form hedonic prediction to allow land values to respond to changes in market potential (the regression equation is shown in Table 1.9, Column 4). The hedonic prediction and the approximate solution to labor market equilibria give us a new set of factor prices at each location.

The full structural simulation emphasizes the fact that when firms move to locations, their demands for factors will drive up the prices of land and labor. This will, in turn, affect the location decisions of other entrants. Hence, it incorporates the full set of agglomeration and dispersion forces in the structural model, and should therefore give more realistic predictions than before.

1.7.5 Simulation Results

Table 1.10 reports the actual and counterfactual new firm counts across the two different scenarios, by province and simulation method. Column 1 reports the actual new firm counts, and columns 2-4 report counterfactual new firm counts for the Trans-Java Expressway simulations. Each simulation was run 1000 times, and 95 percent confidence intervals for changes in new firms were constructed using the empirical distribution of location outcomes across simulations.

Overall, the reduced form simulation (column 2), the model-based upper bound (column 3), and full simulation results (column 4) all tell a similar story: building the Trans-Java Expressway would have induced a small number of firms to locate away from Sumatra and Sulawesi, into the areas in Java that were most affected by the highway. However, the

[50] For more details on this step, see Appendix Section 1.A.4

precision of these estimates and their magnitudes varies depending on whether we focus on the reduced form or the structural predictions. In particular, the reduced form predictions suggest that firms would have moved from all over Sumatra and Sulawesi to relocate in Java, but the structural predictions are much noisier and suggest that only two provinces in Sumatra (North Sumatra and Riau) and one province in Sulawesi (South Sulawesi) would have been adversely affected. The small size of the predicted effects is likely due to the fact that the location fixed effects drive a substantial amount of variation in observed location choices.[51]

Columns 5-7 report counterfactual new firm counts for the rural road upgrading simulation. Again, the overall direction of the reduced form and model-based upper bound predictions is similar, but significance and magnitudes vary. For instance, the reduced form simulation predicts that all provinces in Java would have suffered significant losses in firms to areas in Sulawesi and northern provinces in Sumatra. However, the model-based upper bound predicts that industrial relocation would have only occurred between Jakarta, West Java, and Riau; changes for the other provinces are insignificant. However, the full structural prediction does not yield any significant differences between the actual and counterfactual new firm counts. Overall, the absence of large, statistically significant effects from this simulation suggest that rural road programs may not affect the location choices of firms.

Note that the model based upper bound simulations (columns 3 and 6) show more relocation than the full simulations (columns 4 and 7). This is expected because the model based upper bound ignores factor price responses, and once these are incorporated in the full simulation, this tends to mitigate the effects of increased market potential.

1.8 Conclusion

This paper has made several contributions to our understanding of how road improvements affect the location decisions of firms and, hence, the spatial distribution of economic activity. Using new data that documents a large road improvement program in Indonesia, I provide reduced form evidence showing that better market access for regions near cities is associated with a dispersion of manufacturing firms. Lower transport costs affected different industries in ways predictable from theory; for instance, durable goods producers were much more prone to dispersion than perishable goods producers, who need to locate very close to their sources of demand. These dispersion effects may have resulted from specific features of the road program, or the fact that land is so scare in Indonesia.

Next, I develop a structural model of monopolistic competition and regional trade, in which firms face a tradeoff between greater market access and higher production costs. To

[51] Note that while Java appears unaffected under the full simulation, this is partly due to aggregation; no provinces experienced significant increases in new firms, but two kabupatens on the northern coast of Java (Cirebon (3211) and Jepara (3320)) experienced positive increases in firms.

estimate the model's parameters, I use techniques from industrial organization that allow researchers to estimate discrete choice models with endogenous choice characteristics. I find significant differences between firms willingness to pay for improved market access across different industrial sectors, and I find that the model demonstrates rich patterns of substitution between different locations.

Finally, I use the model to predict what would have happened to industrial location decisions had two different transportation projects actually been undertaken: the on-time construction of the Trans-Java Expressway, and an upgrade to rural roads. My predictions suggest that the Trans-Java Expressway would have caused a modest number of firms to relocate from Sumatra and Sulawesi to the places in Java that were most affected by the toll roads. However, the rural roads program did not induce a statistically significant relocation of firms between provinces. Thus, despite claims made by politicians about the job creating effects of road improvements, I find that industrial locations would be largely stable in response to rural road programs.

This paper has focused on using the model to make counterfactual predictions, leaving aside important questions about social welfare for future research. While rural roads might not have substantially altered firm location choices, they should clearly bring important consumption benefits to rural areas, affecting welfare in ways that the current model cannot capture. A full welfare analysis would also incorporate heterogeneous labor mobility and determine whether road improvements bring spatial surpluses to affected areas, driving jobs away from unaffected regions and lowering welfare in these places, or if they have national productive effects that compensate potential losers.

A major limitation of the paper is that the model developed is static in nature, but it is estimated and simulated dynamically. Using panel data for estimation considerably weakens the identifying restrictions required for estimation, but this comes at a cost: namely, a looser correspondence between the model and how it is estimated. Future research should endeavor to extend the structural model to a full dynamic setting.

Table 1.1: Transportation Budgets for Indonesia's 5-Year Development Plans

(*billions of constant 2000 USD*)	**Repelita IV** **FY 1984-89**	**Repelita V** **FY 1989-94**	**Repelita VI** **FY 1994-99**
Roads	2.1	3.9	3.9
Railways and Freight	0.8	0.8	0.7
Ports and Shipping	1.0	0.7	0.5
Airports and Aircraft	0.7	0.8	0.7
Total	4.6	6.2	5.8
Transport as a Percentage of Total Allocations	11.6	17.6	18.8

Source: Various planning documents for Indonesia's five year development plans (*Rencana Pembangunan Lima Tahun*, abbreviated as *Repelita*). The table reports billions of U.S. dollars allocated to spending on transportation. Budget figures were converted to 2000 USD using OECD data on annual CPI indices and exchange rates.

Table 1.2: Transport Cost Summary Statistics

Java	**1990**	**1995**	**2000**	**2005**
Segment Length	375.44	375.44	374.87	374.87
	(233.79)	(233.79)	(233.32)	(233.32)
	$N = 5671$	$N = 5671$	$N = 5671$	$N = 5671$
Roughness-based Travel Time	4.59	4.16	3.81	4.51
	(2.68)	(2.54)	(2.27)	(2.69)
	$N = 5671$	$N = 5671$	$N = 5671$	$N = 5671$
Paved Road Share	0.46	0.77	0.79	0.80
	(0.50)	(0.42)	(0.41)	(0.40)
Sumatra	**1990**	**1995**	**2000**	**2005**
Segment Length	725.83	725.83	725.83	725.83
	(436.00)	(436.00)	(436.00)	(436.00)
	$N = 2145$	$N = 2145$	$N = 2145$	$N = 2145$
Roughness-based Travel Time	10.74	9.49	8.12	9.62
	(6.24)	(5.58)	(4.82)	(5.79)
	$N = 2145$	$N = 2145$	$N = 2145$	$N = 2145$
Paved Road Share	0.32	0.56	0.70	0.71
	(0.46)	(0.50)	(0.46)	(0.46)
Sulawesi	**1990**	**1995**	**2000**	**2005**
Segment Length	683.69	683.69	683.69	683.69
	(494.97)	(494.97)	(494.97)	(494.97)
	$N = 561$	$N = 561$	$N = 561$	$N = 561$
Roughness-based Travel Time	13.77	10.59	8.50	8.89
	(10.33)	(7.03)	(5.78)	(6.08)
	$N = 561$	$N = 561$	$N = 561$	$N = 561$
Paved Road Share	0.16	0.33	0.54	0.55
	(0.36)	(0.47)	(0.50)	(0.49)

Source: IRMS and author's calculations. For segment length and roughness-based travel times, the unit of observation is a pair of kabupatens on the same island. For percentage paved roads, estimates are taken from the detailed kilometer-post-interval data. Standard deviations in parentheses.

Table 1.3: Changes in New Firm Shares: 1985-1996

	Δ Share of New Firms	**Change Corresponding To ...**	
	1985 – 1996	**Durable**	**Non-Durable**
Cities	-0.159	-0.103	-0.056
Neighbors of Cities	0.095	0.071	0.024
Neighbors of Neighbors	0.099	0.052	0.046
Rural	-0.022	-0.006	-0.017
Other	-0.012	-0.003	-0.009

Source: SI and author's calculations. A total of 51 out of 218 kabupatens were classified as Cities in 1990. "Neighbors of Cities" are kabupatens that share a border with 1990 cities; there were 60 kabupatens in this category. "Neighbors of Neighbors of Cities" are kabupatens that share a border with kabupatens who share a border with 1990 cities; there were 78 kabupatens in this category. The remaining 29 kabupatens are categorized as "Rural". In classifying, some kabupatens fit into multiple categories, and when this occurred, the kabupaten was assigned to the group closest to cities as possible.

Table 1.4: Reduced Form Regressions

	IV			
Panel A: New Firms	**(1)**	**(2)**	**(3)**	**(4)**
MP_{rt}	0.122	0.122	0.076	
	(0.031)***	(0.031)***	(0.015)***	
$\text{MP}_{rt} \times \text{Durable}_j$				0.019
				(0.008)**
Adj. R^2	0.304	0.307	0.313	0.290
N	50320	50320	50320	50320
F-Statistic	15.822	15.832	25.858	6.000
Kabupaten FE	Yes	Yes	Yes	.
Year FE	Yes	.	.	.
Sector FE	Yes	.	.	.
Sector-Year FE	.	Yes	Yes	Yes
Province Trends	.	.	Yes	.
Kabu-Year FE	.	.	.	Yes

	IV			
Panel B: Employment	**(1)**	**(2)**	**(3)**	**(4)**
MP_{rt}	0.432	0.432	0.271	
	(0.089)***	(0.089)***	(0.070)***	
$\text{MP}_{rt} \times \text{Durable}_j$				0.067
				(0.027)**
Adj. R^2	0.301	0.303	0.307	0.275
N	50320	50320	50320	50320
F-Statistic	23.256	23.270	15.153	6.061
Kabupaten FE	Yes	Yes	Yes	.
Year FE	Yes	.	.	.
Sector FE	Yes	.	.	.
Sector-Year FE	.	Yes	Yes	Yes
Province Trends	.	.	Yes	.
Kabu-Year FE	.	.	.	Yes

Unit of observation is a region-industry-year. Robust standard errors in parentheses, clustered at the kabupaten level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level. The adjusted R-squared reported is the *within* R-squared, taken from the analogous reduced form regression.

Table 1.5: Reduced Form Regressions: Robustness

	IV w/ Placebo			IV Sequential Moments		
Panel A: New Firms	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
MP_{rt}	0.120	0.120	0.083	0.045	0.045	0.046
	(0.031)***	(0.031)***	(0.015)***	(0.019)**	(0.019)**	(0.018)**
Unbuilt toll road	-0.003	-0.003	0.015			
in kabu. r ($t \geq 1994$)	(0.011)	(0.011)	(0.013)			
Adj. R^2	-0.004	-0.003	-0.004	-0.000	-0.000	-0.000
N	50320	50320	50320	44030	44030	44030
F-Statistic	8.395	8.400	14.960	5.678	5.679	6.250
Kabupaten FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	.	.	Yes	.	.
Sector FE	Yes	.	.	Yes	.	.
Sector-Year FE	.	Yes	Yes	.	Yes	Yes
Province Trends	.	.	Yes	.	.	Yes
Lagged Diff MP-85 IV	.	.	.	Yes	Yes	Yes

	IV w/ Placebo			IV Sequential Moments		
Panel B: Employment	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
MP_{rt}	0.430	0.430	0.300	0.234	0.234	0.238
	(0.092)***	(0.092)***	(0.066)***	(0.081)***	(0.081)***	(0.082)***
Unbuilt toll road	-0.004	-0.004	0.062			
in kabu. r ($t \geq 1994$)	(0.046)	(0.046)	(0.050)			
Adj. R^2	-0.004	-0.003	-0.004	-0.000	0.000	-0.000
N	50320	50320	50320	44030	44030	44030
F-Statistic	11.749	11.756	10.210	8.293	8.296	8.374
Kabupaten FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	.	.	Yes	.	.
Sector FE	Yes	.	.	Yes	.	.
Sector-Year FE	.	Yes	Yes	.	Yes	Yes
Province Trends	.	.	Yes	.	.	Yes
Lagged Diff MP-85 IV	.	.	.	Yes	Yes	Yes

Unit of observation is a region-industry-year. Robust standard errors in parentheses, clustered at the kabupaten level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.6: Constant Coefficient Logit Results

	OLS	**Panel GMM (1990-2005)**						
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
wage_rate	0.080 (0.088)	0.076 (0.086)	-0.145 (0.050)***	-0.143 (0.053)***	-0.138 (0.050)***	-0.140 (0.050)***	-0.116 (0.046)**	-0.193 (0.088)**
land_value	0.177 (0.058)***	0.182 (0.058)***	-0.076 (0.028)***	-0.075 (0.030)**	-0.072 (0.029)**	-0.075 (0.028)***	-0.079 (0.033)**	-0.042 (0.048)
indTaxRate	-3.423 (4.915)	-3.789 (4.817)	-6.383 (2.512)**	-6.391 (2.642)**	-7.059 (2.529)***	-6.102 (2.484)**	-4.687 (2.208)**	-7.479 (2.589)***
MP	0.603 (0.110)***	0.580 (0.105)***	1.100 (0.461)**	0.384 (0.182)**		1.007 (0.462)**	1.939 (0.862)**	2.114 (0.775)***
pavedDensity					0.079 (0.073)			
sharePLN						0.189 (0.126)		
δ_0				0.818 (0.071)***				
δ_1				1.126 (0.064)***				
δ_2				0.118 (0.012)***				
δ_3				0.099 (0.007)***				
Adj. R^2	0.241	0.229	0.715	0.743	0.714	0.715	0.729	0.312
N	2093	2093	2093	2093	2093	2093	2093	1937
F Statistic	13.770	13.497	14.159	21.476	14.551	13.837	4.791	15.395
Rural-Urban Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Market Potential IV	.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Kabupaten FE	.	.	Yes	Yes	Yes	Yes	Yes	Yes
Province-Year FE	.	.	.	.	.	.	Yes	.
Dynamic Panel IVs	.	.	.	.	.	.	.	Yes
WTP for MP with wages			7.60**	2.69*	0.57	7.21*	16.71*	10.95
WTP for MP with rents			14.45*	5.14	1.09	13.48*	24.68*	50.46
WTP for MP with taxes			0.17	0.06	0.01	0.16	0.41	0.28*

Robust standard errors in parentheses, clustered at the kabupaten level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level. In all columns except column 1, the adjusted R-squared reported is taken from the analogous reduced form regression.

Table 1.7: Random Coefficients Logit Results: Fixed Effects

		Means for Industrial Sectors							
	Overall Mean	**Foods & Beverages**	**Textiles & Clothing**	**Wood Products**	**Chemicals & Oil Prods.**	**Ceramics, Glass, & Non-Metals**	**Finished Metal Products**	**Other Products**	**Standard Deviations** σ
WAGE_RATE	-0.156 (0.053)***								
LAND_VALUE	-0.066 (0.030)**								
MP		0.954 (0.405)**	1.075 (0.416)***	0.093 (0.407)	1.539 (0.407)***	1.778 (0.551)***	1.385 (0.415)***	2.078 (0.405)***	0.953 (0.155)***
INDTAXRATE	-6.491 (2.688)**								

The model is estimated on the full sample of the new firms dataset. There are 17,684 firms across all years choosing locations, and given the variation in the choice set across years, there are a total of 2,442,084 observations. The first step mixed logit model was estimated with 100 scrambled Halton draws for each industry. The estimated simulated log-likelihood was equal to −19410.56, and the simulated likelihood ratio index is equal to $\rho \equiv 1 - SLL(\widehat{\beta})/SLL(0) = 0.7769$. Standard errors in parentheses, computed using asymptotic GMM results and the delta method (see Appendix 1.B.3 for more details). * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.8: Cross Market Potential Elasticity Regressions

	Median η_{jk}		
	(1)	(2)	(3)
Physical Distance	-0.001 (0.000)***		
Abs. Population Difference		-2.586 (0.406)***	
Abs. GDP Difference			-0.001 (0.000)***
Adj. R^2	0.831	0.835	0.833
N	33241	33241	33241
Region j FE	Yes	Yes	Yes
Region k FE	Yes	Yes	Yes

The unit of analysis is a location j-k pair, and the dependent variable is 1000 times the median η_{jk}^{MP}, where the median is taken over all years in which both locations were chosen by firms. The rescaling was used to make the parameter estimates reasonably sized. Robust standard errors in parentheses, clustered at the region j level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.9: Hedonic Regressions

	Wages		Land Values	
	(1)	(2)	(3)	(4)
MP	0.583 (0.274)**	0.577 (0.277)**	0.808 (0.386)**	0.760 (0.385)**
Adj. R^2	0.889	0.888	0.767	0.767
N	2960	2960	2960	2960
F Statistic	340.316	402.648	52.372	55.575
Kabupaten FE	Yes	Yes	Yes	Yes
Year FE	Yes	.	Yes	.
Rural-Urban Year FE	.	Yes	.	Yes

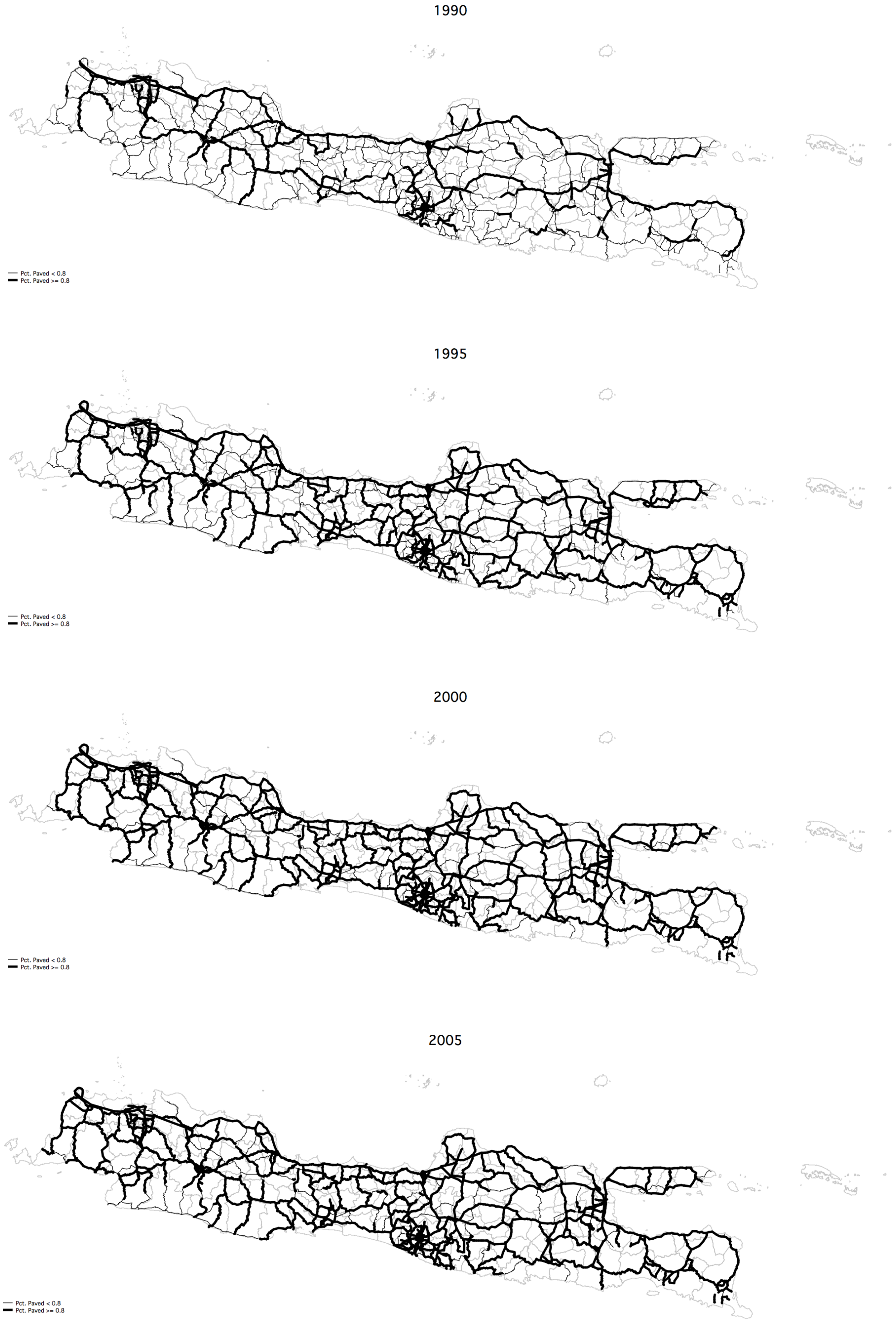
The unit of analysis is a region-year. Robust standard errors in parentheses, clustered at the region level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.10: Actual and Counterfactual New Firms, 1994-2005

		Δ Trans-Java Highway			Δ Rural Road Upgrades		
	Actual	**RF**	**MBUB**	**Full**	**RF**	**MBUB**	**Full**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(6)**
Sumatra							
Aceh	45	-1.0**	-0.4	-0.3	0.3**	1.1	0.3
North Sumatra	375	-9.7**	-10.7**	-26.3**	1.2**	49.0	-17.4
West Sumatra	72	-1.6**	-1.0	-3.9	0.0	3.8	-3.0
Riau	436	-1.2	-11.8**	-31.6**	18.0**	107.8**	-14.5
Jambi	44	-0.8**	-0.2	-2.6*	-0.0**	1.9	-2.1
South Sumatra	124	-2.7**	-2.7	-6.1	-0.3**	1.9	-4.1
Bengkulu	20	-0.5**	-0.4	-0.7	-0.1**	0.3	-0.4
Lampung	60	-1.1**	-1.0	-1.6	-0.2**	0.1	-0.9
Java							
Jakarta	996	-1.5**	-13.8	61.8	-3.9**	-34.7**	59.0
West Java	3591	-7.3**	-30.4	107.8*	-13.8**	-106.2**	104.5
Central Java	1793	32.5**	50.2**	6.6*	-2.6**	-22.7	-19.0
Yogyakarta	301	1.9**	4.2	-6.9	-0.9**	-3.4	-8.6
East Java	2617	4.4**	20.8	-62.0	-4.1**	-52.5	-67.1
Sulawesi							
North Sulawesi	140	-3.2**	-0.2	-7.8	0.5**	12.0	-6.0
Central Sulawesi	45	-0.6**	-0.0	-1.5	1.4**	3.2	-0.8
South Sulawesi	229	-5.4**	-2.1**	-18.0	-0.2**	15.6	-14.7
Southeast Sulawesi	98	-2.1**	-0.5	-7.2	4.8**	22.8	-5.2

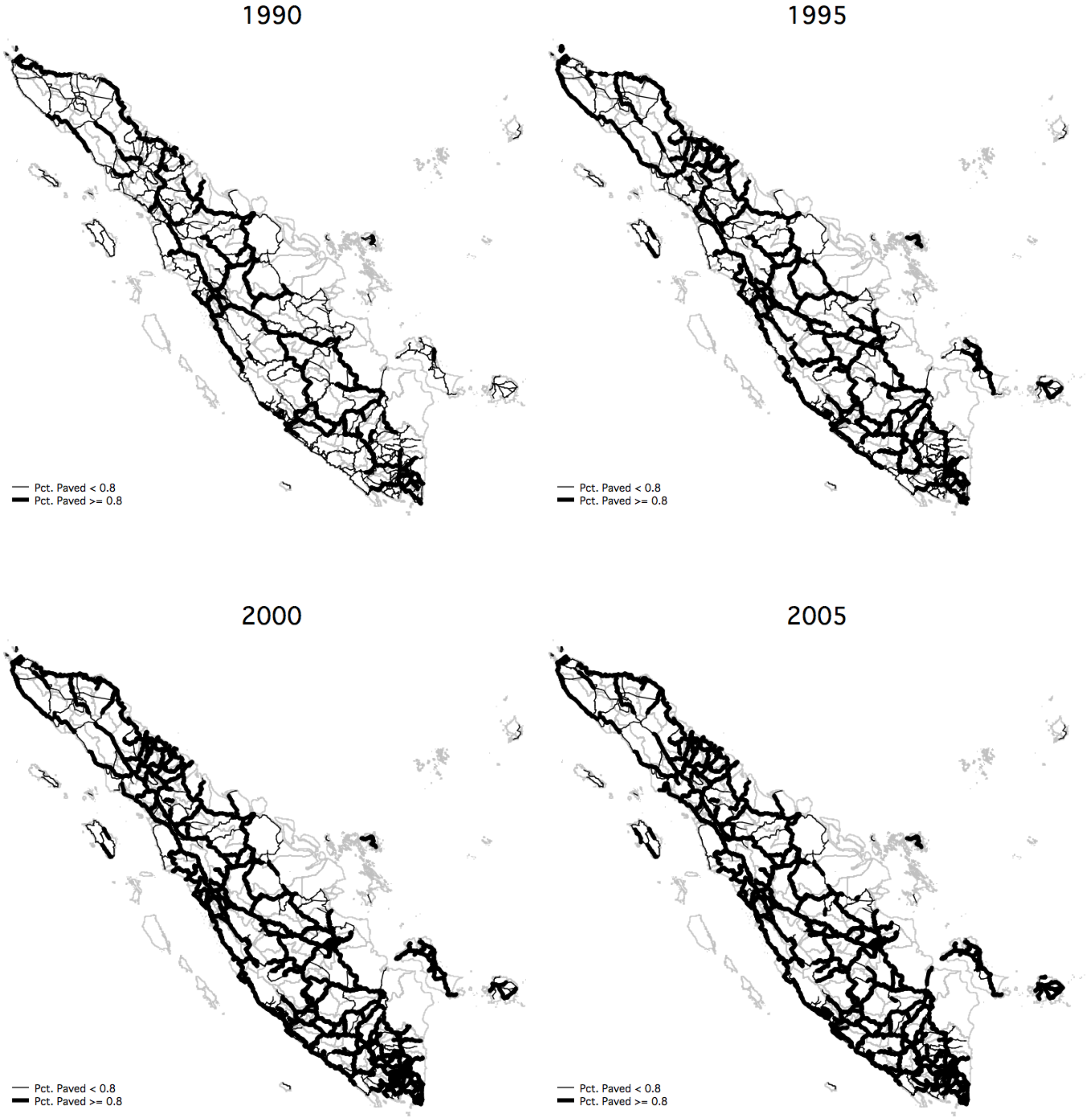
Source: Authors' calculations. Column 1 reports the actual number of new firms who located in each province. Columns 2-4 report the mean change in the counterfactual number of new firms if the Trans-Java Expressway had been constructed, and Column 4-6 reports the mean change in the number of new firms for the upgraded rural roads scenario. For each scenario, each of the three simulation methods (reduced form (RF), model-based upper bound (MBUB), and full structural prediction (Full)) was conducted 1000 times. For each counterfactual and simulation method, 95 percent confidence intervals were constructed using the empirical distribution of location outcomes for all simulations. The ** symbol denotes a statistically significant change in the number of new firms, relative to the actual number, while the * denotes that while the total for the province was not significantly different from zero, kabupatens within that province had statistically significant changes.

Figure 1.1: Evolution of Pavement on Java's Road Network



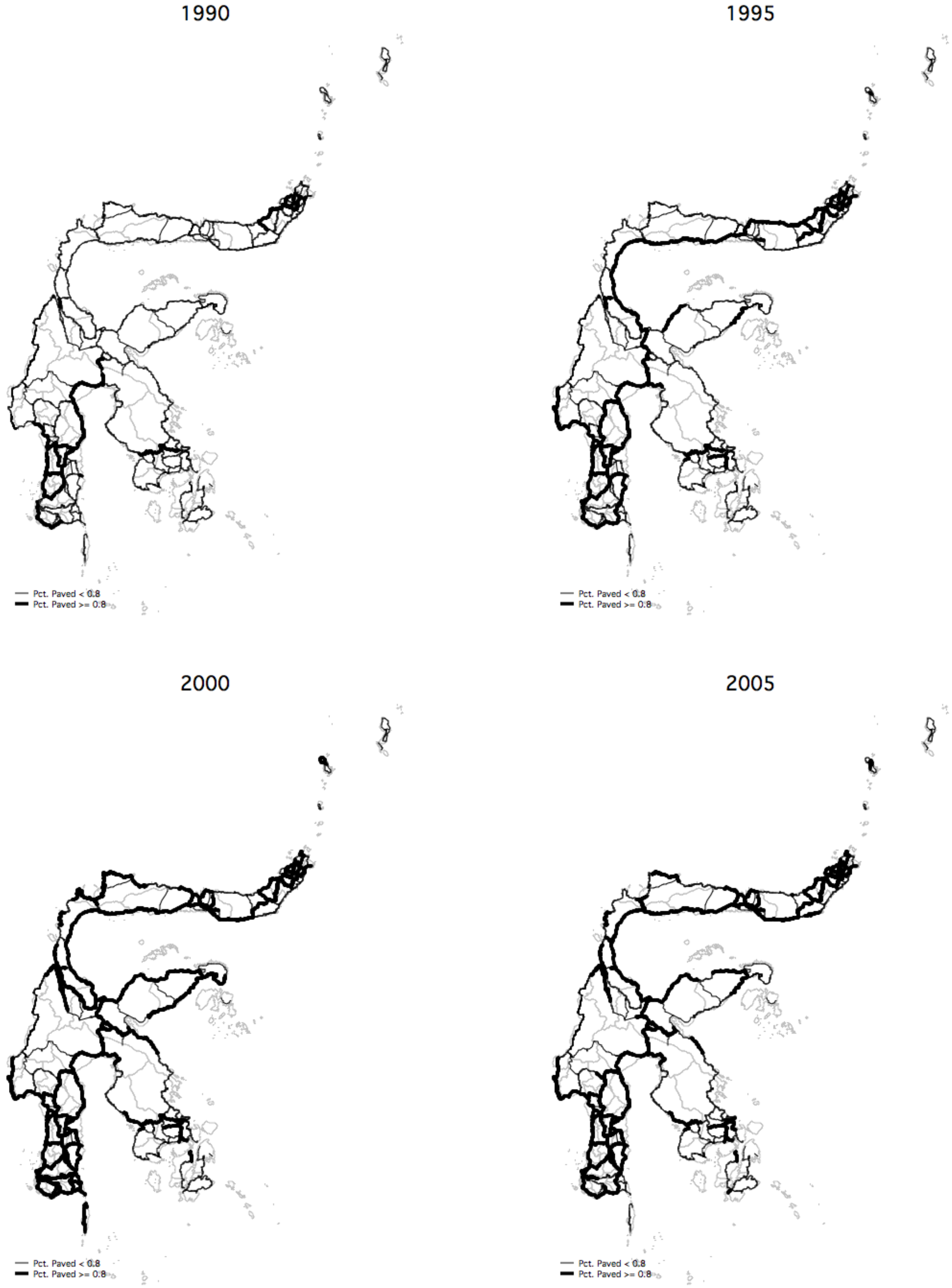
Source: IRMS and author's calculations. Thick black lines correspond to road sections that are 80 percent paved or greater, while thin black lines correspond to road sections that are less than 80 percent paved.

Figure 1.2: Evolution of Pavement on Sumatra's Road Network



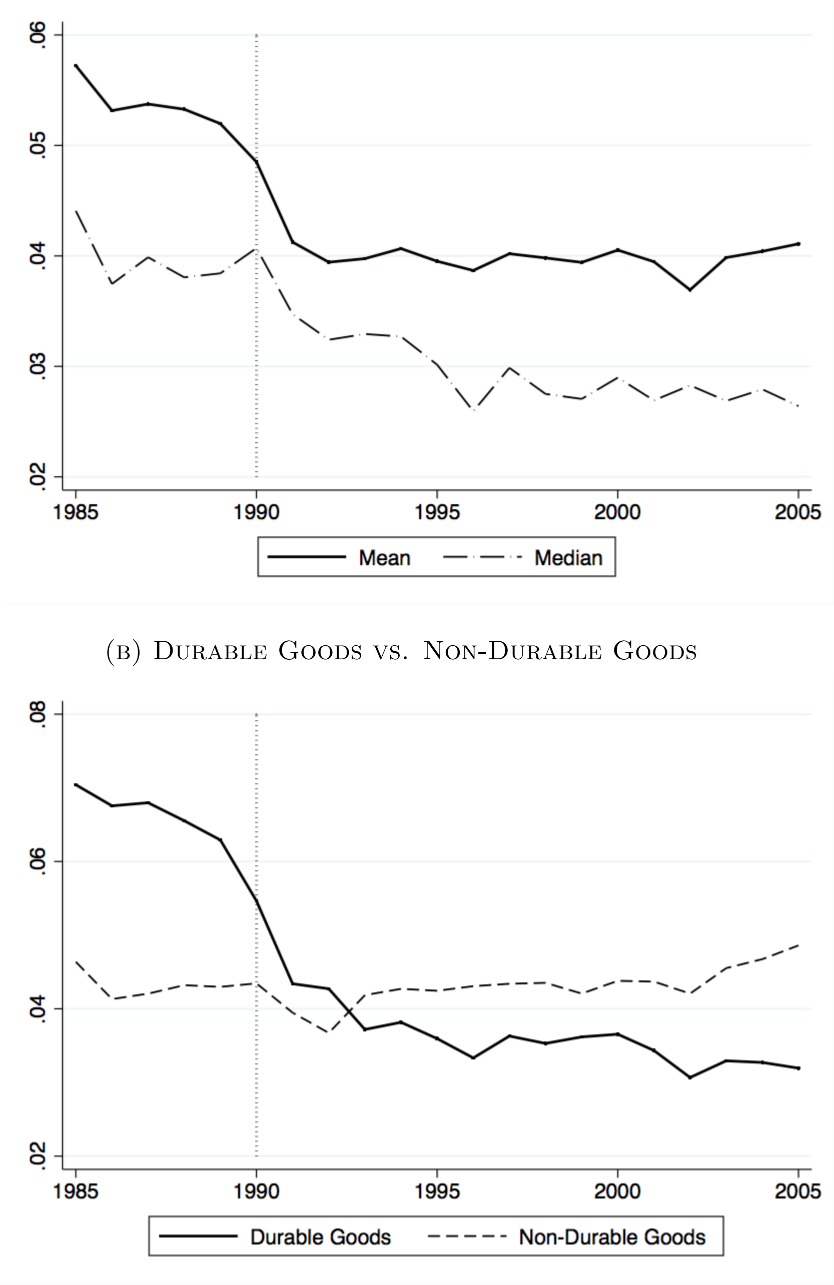
Source: IRMS and author's calculations. Thick black lines correspond to road sections that are 80 percent paved or greater, while thin black lines correspond to road sections that are less than 80 percent paved.

Figure 1.3: Evolution of Pavement on Sulawesi's Road Network



Source: IRMS and author's calculations. Thick black lines correspond to road sections that are 80 percent paved or greater, while thin black lines correspond to road sections that are less than 80 percent paved.

Figure 1.4: Trends in the Ellison and Glaeser (1997) Index

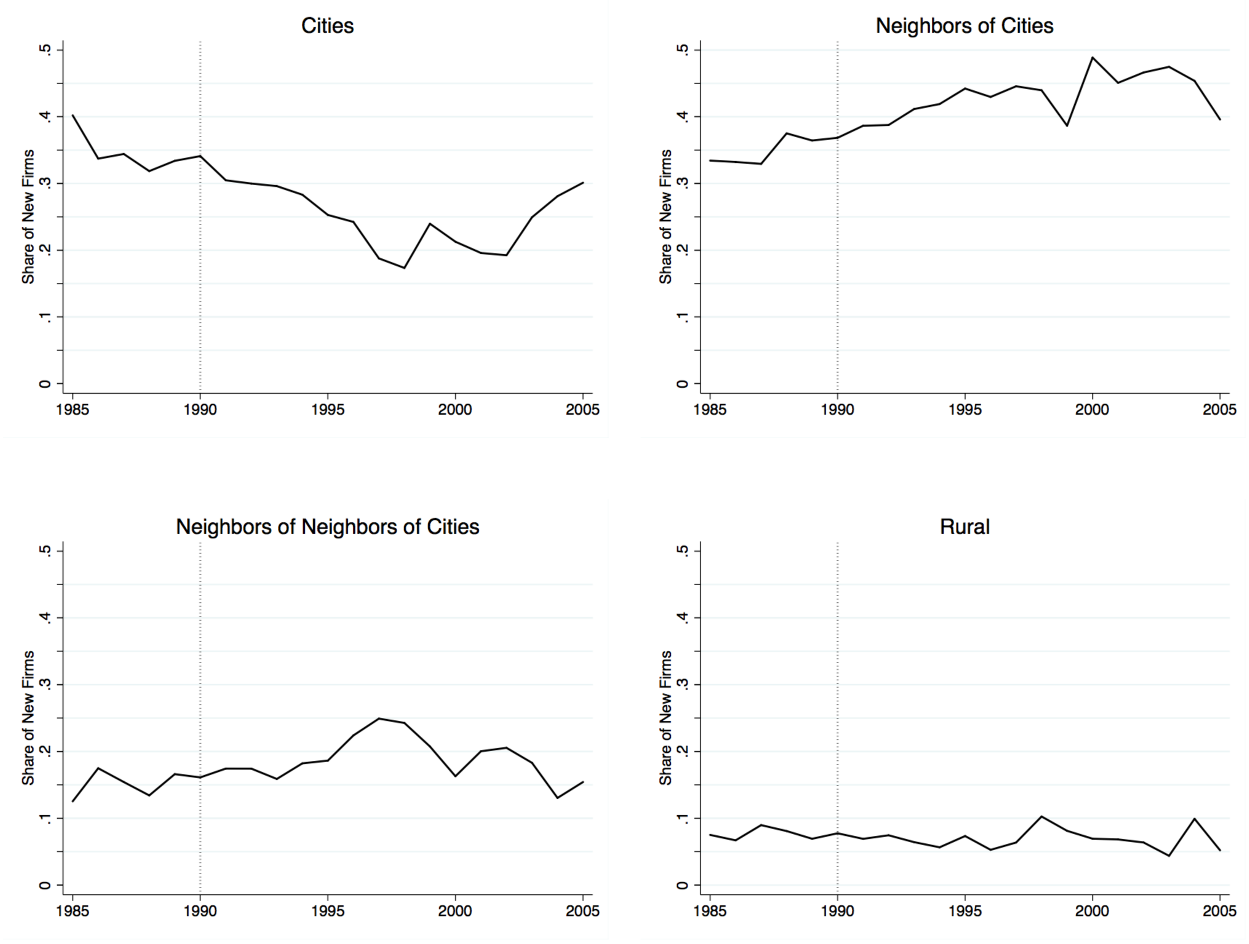


(a) All industries

(b) Durable Goods vs. Non-Durable Goods

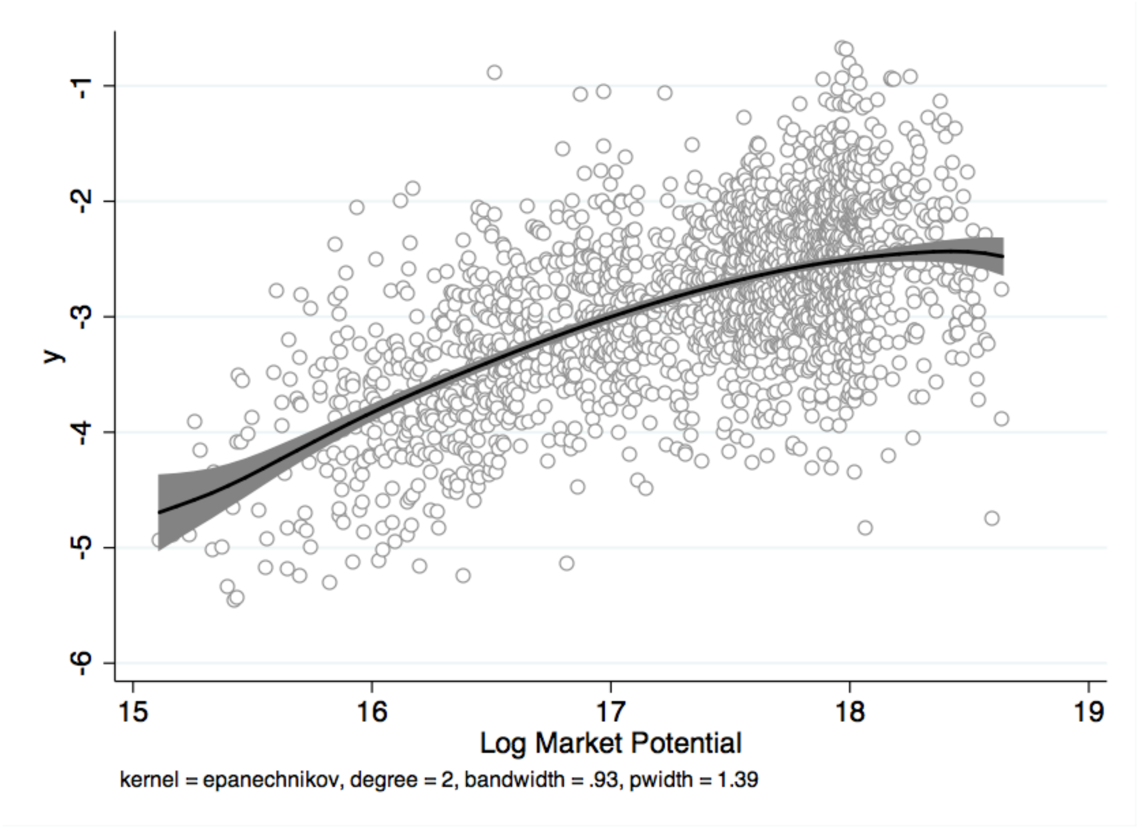
Source: SI data and author's calculations. Lines depict annual means or medians of different indices of industrial concentration across 5-digit industries, as well as means by industry type. Grey bar denotes crisis period (1997-1999). Regressions of industrial concentration measures across industry years on a set of year dummies (or a trend) indicate that the observed reductions are statistically significant, beginning in the Ellison and Glaeser (1997) index (see Table 1.C.3). From difference-in-difference regressions (see Appendix Table 1.C.4), the change in the Spatial Herfindahl for durable goods industries, relative to non-durable goods, was -0.05 (s.e. 0.019). Similar magnitudes for difference-in-difference estimates can be found for the Ellison and Glaeser Index, though the estimates are noisier.

Figure 1.5: Share of New Firms Locating in Different Types of Kabupatens



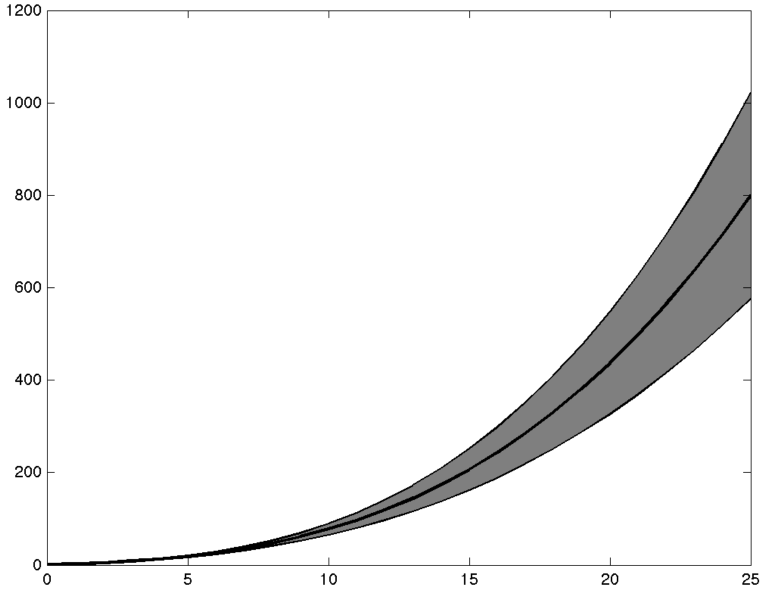
Source: SI data and author's calculations. Lines depict shares of new firms locating in different types of kabupatens within Java, Sumatra, and Sulawesi. Grey bar denotes crisis period (1997-1999). A total of 51 out of 218 kabupatens were classified as Cities in 1990. "Neighbors of Cities" are kabupatens that share a border with 1990 cities; there were 60 kabupatens in this category. "Neighbors of Neighbors of Cities" are kabupatens that share a border with kabupatens who share a border with 1990 cities; there were 78 kabupatens in this category. The remaining 29 kabupatens are categorized as "Rural". In classifying, some kabupatens fit into multiple categories, and when this occurred, the kabupaten was assigned to the group closest to cities as possible.

Figure 1.6: Partially Linear Regression



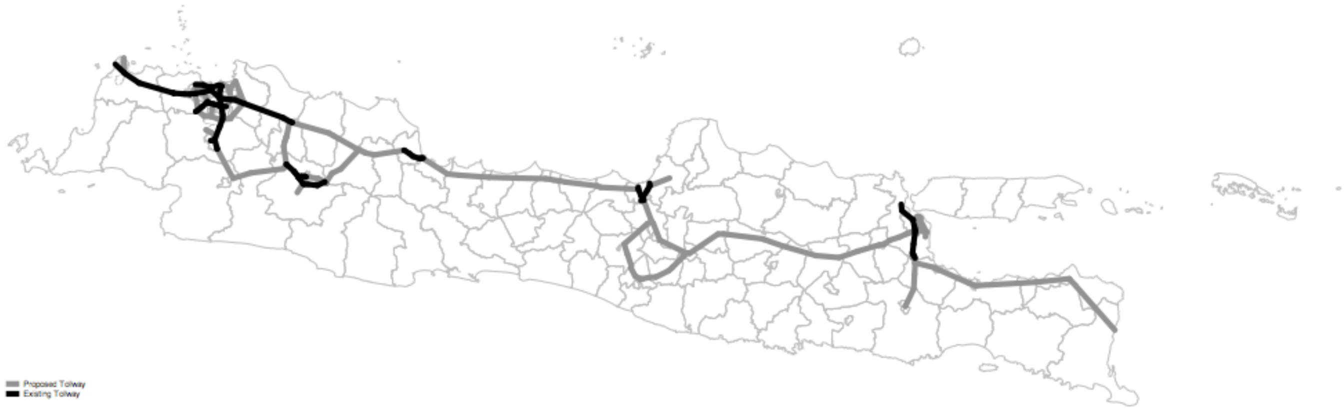
Partially linear regression was implemented using the sorting and differenced-based procedure discussed in Yatchew (1997), using first-order differencing. All regressions have kabupaten-specific intercepts and rural-urban-year-specific intercepts; these intercepts form the linear portion of the regression. Following Yatchew (1997), we tested the null hypothesis that $H_0 : f(MP) = \gamma$. This involves computing $V = \sqrt{T}(s^2_{res} - s^2_{diff})/s^2_{diff}$, where s^2_{res} is the residual variance of the full partially linear model, and $s^2_d iff$ is the residual variance under the null hypothesis. Under H_0, $V \sim N(0, 1)$ and it was calculated to be $\widehat{V} = 2.050$, hence the null hypothesis was rejected (p-value = 0.020).

Figure 1.7: Non-Linear Distance Function



The x axis is τ (roughness-based travel time, in hours), and the y axis is $\widehat{f}(\tau) = \widehat{\delta}_0 + \widehat{\delta}_1\tau + \widehat{\delta}_2\tau^2 + \widehat{\delta}_3\tau^3$, where the δ's are estimated in Table 1.6, Column 5. Pointwise 95 percent confidence bands are depicted in grey.

Figure 1.8: Map of the Trans-Java Expressway



Source: *Departemen Pekerjaan Umum.*

1.A Derivations for the Model and Counterfactuals

1.A.1 Consumer Demands

To derive the consumer demands for individual varieties, (1.6), first let E_k represent consumer expenditures on industry k. To choose optimal bundles of varieties from industry k, we setup the following Lagrangian:

$$\begin{aligned}\mathcal{L}_k &= \mathbf{M}_k + \lambda_k \left(E_k - \int_0^1 p^k(j) q^k(j) dj \right) \\ &= \left(\int_0^1 q^k(j)^{\frac{\sigma_k - 1}{\sigma_k}} \right)^{\frac{\sigma_k}{\sigma_k - 1}} + \lambda_k \left(E_k - \int_0^1 p^k(j) q^k(j) dj \right)\end{aligned}$$

Taking the derivative of this with respect to $q^k(j)$, we have:

$$\begin{aligned}\frac{\partial \mathcal{L}_k}{\partial q^k(j)} = \left(\frac{\sigma_k}{\sigma_k - 1} \right) \left(\int_0^1 q^k(j)^{\frac{\sigma_k - 1}{\sigma_k}} \right)^{\frac{1}{\sigma_k - 1}} \left(\frac{\sigma_k - 1}{\sigma_k} \right) q^k(j)^{\frac{-1}{\sigma_k}} - \lambda_k p^k(j) &\overset{\text{set}}{=} 0 \\ \implies \quad \left(\int_0^1 q^k(j)^{\frac{\sigma_k - 1}{\sigma_k}} \right)^{\frac{1}{\sigma_k - 1}} q^k(j)^{\frac{-1}{\sigma_k}} &= \lambda_k p^k(j)\end{aligned}$$

Rearranging terms, we have:

$$\begin{aligned}\left(\int_0^1 q^k(j)^{\frac{\sigma_k - 1}{\sigma_k}} \right)^{\frac{1}{\sigma_k - 1}} &= \lambda_k p^k(j) q^k(j)^{\frac{1}{\sigma_k}} \\ \left(\int_0^1 q^k(j)^{\frac{\sigma_k - 1}{\sigma_k}} \right)^{\frac{\sigma_k}{\sigma_k - 1}} &= \lambda_k^{\sigma_k} p^k(j)^{\sigma_k} q^k(j) \\ \implies \quad \mathbf{M}_k \lambda_k^{-\sigma_k} p^k(j)^{-\sigma_k} &= q^k(j) \qquad (1.18)\end{aligned}$$

Now, multiplying both sides by $p^k(j)$ and integrating over the set of varieties, we have:

$$\begin{aligned}\mathbf{M}_k \lambda_k^{-\sigma_k} p^k(j)^{1-\sigma_k} &= p^k(j) q^k(j) \\ \mathbf{M}_k \lambda_k^{-\sigma_k} \left(\int_0^1 p^k(j)^{1-\sigma_k} dj \right) &= \int_0^1 p^k(j) q^k(j) dj \equiv E_k\end{aligned}$$

So, rearranging, we have:

$$\mathbf{M}_k \lambda_k^{-\sigma_k} = \frac{E_k}{\int_0^1 p^k(j)^{1-\sigma_k} dj} \qquad (1.19)$$

Plugging (1.19) into (1.18), we arrive at the following expression:

$$q^k(j) = \frac{p^k(j)^{-\sigma_k} E_k}{\int_0^1 p^k(j)^{1-\sigma_k} dj} = \frac{p^k(j)^{-\sigma_k} E_k}{(\mathbf{P}^k)^{1-\sigma_k}}$$

where $\mathbf{P}^k$ is the price index defined in (1.7).

All that remains is to determine E_k, the share of the budget spent on manufacturing varieties from industry k. But, note that (1.5) is just a Cobb-Douglass utility function over the CES manufacturing indices. Hence, the budget shares are determined by the λ_k's, and $E_k = \lambda_k \mathbf{Y}$.

1.A.2 Firm Pricing

To derive the profit-maximizing prices that firms charge for varieties, note that a firm's profits from operating in region o and shipping goods to region d are given by:

$$\pi^k_{od}(i) = \left(p^k_{od}(i) - m^k_o(i) w \tau^k_{od}\right) q^k(i)$$

Note that expression takes into account the iceberg transport costs assumption, that in order to deliver one unit of the variety to region d, τ^k_{od} units must be produced.

Taking the derivative of this profit function with respect to $p^k_{od}(i)$, we have:

$$\frac{\partial \pi^k_{od}(i)}{\partial p^k_{od}(i)} = q^k(i) + \left(p^k_{od}(i) - m^k_o(i) w \tau^k_{od}\right) \frac{\partial q^k(i)}{\partial p^k_{od}(i)}$$

Setting this expression equal to zero and rearranging, we have:

$$\begin{aligned}
q^k(i) + p^k_{od}(i)\left(\frac{\partial q^k(i)}{\partial p^k_{od}(i)}\right) &= \left(m^k_o(i) w \tau^k_{od}\right) \frac{\partial q^k(i)}{\partial p^k_{od}(i)} \\
p^k_{od}(i)\left(\frac{q^k(i)}{p^k_{od}(i)} + \frac{\partial q^k(i)}{\partial p^k_{od}(i)}\right) &= \left(m^k_o(i) w \tau^k_{od}\right) \frac{\partial q^k(i)}{\partial p^k_{od}(i)} \\
p^k_{od}(i)\left(1 + \frac{\partial q^k(i)}{\partial p^k_{od}(i)} \frac{p^k_{od}(i)}{q^k(i)}\right) &= \left(m^k_o(i) w \tau^k_{od}\right) \left(\frac{\partial q^k(i)}{\partial p^k_{od}(i)} \frac{p^k_{od}(i)}{q^k(i)}\right) \qquad (1.20)
\end{aligned}$$

We compute $\left(\partial q^k(i)/\partial p^k_{od}(i)\right)\left(p^k_{od}(i)/q^k(i)\right)$ using the consumer's demand function, (1.6), and noting that because of the Dixit-Stiglitz structure of competition, firms ignore the effect that their prices have on the price index for their industry in region d, $\mathbf{P}^k_d$. This gives us:

$$\begin{aligned}
\left(\frac{\partial q^k(i)}{\partial p^k_{od}(i)} \frac{p^k_{od}(i)}{q^k(i)}\right) &= \left(\frac{-\sigma_k p^k_{od}(i)^{-\sigma_k - 1} \mu^k \mathbf{Y}_d}{\left(\mathbf{P}^k_d\right)^{1-\sigma_k}}\right) \frac{p^k_{od}(i)}{q^k(i)} \\
&= -\sigma_k \left(\frac{p^k_{od}(i)^{-\sigma_k} \mu^k \mathbf{Y}_d}{\left(\mathbf{P}^k_d\right)^{1-\sigma_k}}\right) \frac{\left(\mathbf{P}^k_d\right)^{1-\sigma_k}}{p^k_{od}(i)^{-\sigma_k} \mu^k \mathbf{Y}_d} = -\sigma_k
\end{aligned}$$

Plugging this result into (1.20), we have:

$$p_{od}^{k}(i)\left(1-\sigma_k\right)=\left(m_o^k(i)w\tau_{od}^k\right)\left(-\sigma_k\right)$$

from which (1.12) follows immediately.

1.A.3 Firm Outputs at Counterfactual Locations

We first need an expression for each firm's total output. Remembering that firms have to overproduce to satisfy export demands, their total output is given by the following:

$$q_o(i)=\sum_{d=1}^{R}\tau_{od}q_{od}^{*}(i)$$

Using the firm's optimal pricing formula, (1.12), and the demand function, (1.6), the equilibrium production quantities for exports to region d from region o are given by:

$$q_{od}^{*}(i)=\frac{\left[\frac{\sigma_k}{\sigma_k-1}\tau_{od}^k m_o^k(i)\right]^{-\sigma_k}\mu_k\mathbf{Y}_d}{(\mathbf{P}_d^k)^{1-\sigma_k}}$$

Plugging this expression into the one above and simplifying, we obtain the following:

$$\begin{aligned}q_o(i)&=\sum_{d=1}^{R}\tau_{od}^k\left\{\frac{\left[\frac{\sigma_k}{\sigma_k-1}\tau_{od}^k m_o^k(i)\right]^{-\sigma_k}\mu_k\mathbf{Y}_d}{(\mathbf{P}_d^k)^{1-\sigma_k}}\right\}\\&=\sum_{d=1}^{R}\left(\tau_{od}^k\right)^{1-\sigma_k}\mu_k\left(\frac{\sigma_k}{\sigma_k-1}\right)^{-\sigma_k}\left[m_o^k(i)\right]^{-\sigma_k}\frac{\mathbf{Y}_d}{(\mathbf{P}_d^k)^{1-\sigma_k}}\\&=\theta_k\left[m_o^k(i)\right]^{-\sigma_k}\sum_{d=1}^{R}\tau_{od}^{1-\sigma_k}\frac{\mathbf{Y}_d}{(\mathbf{P}_d^k)^{1-\sigma_k}}\\&=\theta_k\left[m_o^k(i)\right]^{-\sigma_k}RMP_o^k\end{aligned}$$

where θ_k is an industry-specific constant, given by:

$$\theta_k=\mu_k\left(\frac{\sigma_k}{\sigma_k-1}\right)^{-\sigma_k}\left(\eta^k\right)^{1-\sigma_k}$$

and η^k is the industry-specific multiplier for transport costs. Hence, the model implies that a firm's total output in equilibrium is given by:

$$q_{ot}(i)=\theta_k m_{ot}(i)^{-\sigma_k}RMP_{ot}\tag{1.21}$$

The parameter σ_k is the industry-specific elasticity of substitution parameter, which we can get by inverting the coefficient on market potential (from the choice model estimation):

$$\widehat{\sigma_k} = 1 + \frac{1}{\widehat{\beta}_{MP,k}}$$

In order to avoid estimating θ_k, it will be useful to first take the a ratio of counterfactual outputs in new location n to actual outputs, then to take logs:

$$\ln\left(q_{nt}(i)^c\right) = \ln\left(q_{ot}(i)\right) - \sigma_k \ln\left(m^k_{nt}(i)^c / m^k_{ot}(i)\right) + \ln\left(RMP^c_{nt}/RMP^c_{ot}\right) \tag{1.22}$$

We have everything we need to compute this expression, except for a measure of the firm's total marginal costs, $m_{ot}(i)$. To get this, recall that the firm's value function (whose parameters we estimate in the choice model) is defined by:

$$V_{oit} = \frac{1}{\sigma_k - 1}\ln\left(RMP_{ot}\right) - \ln\left(m^k_{ot}(i)\right)$$

In Section 1.5.5, I show how we can write this value function as the sum of a mean profit term, δ_{ot}, common to all firms and industries, a mean-zero heteroskedastic deviation from this mean profit term, μ_{oit}, and an idiosyncratic error term:

$$V_{oit} = \delta_{ot} + \mu_{oit} + \varepsilon_{oit}$$

This implies that we can write the log of firm marginal costs as:

$$\ln\left(m^k_{ot}(i)\right) = \frac{1}{\sigma_k - 1}\ln\left(RMP_{ot}\right) - \delta_{ot} - \mu_{oit} - \varepsilon_{oit}$$

Note that in the special case where μ_{oit} contains only an industry-dummy interaction and random coefficient on the market potential variable, we can write:

$$\ln\left(m^k_{ot}(i)\right) = \left[\overline{\beta}_{MP} - \beta(i)\right]\ln\left(RMP_{ot}\right) - \delta_{ot} - \varepsilon_{oit}$$

This implies that we can write a firm's new output in new counterfactual location, n, as the following:

$$\begin{aligned}\ln q_{nt}(i)^c = \ln q_{ot}(i) + \left(1 - \sigma_k\left[\overline{\beta}_{MP} - \beta(i)\right]\right)\left\{\ln\left(RMP^c_{nt}\right) - \ln\left(RMP_{ot}\right)\right\} \\ + \sigma_k\left(\delta^c_{nt} - \delta_{ot}\right) + \sigma_k\left(\varepsilon_{int} - \varepsilon_{iot}\right)\end{aligned}$$

Using this formula, we can predict counterfactual outputs in the new locations.

1.A.4 Counterfactual Factor Demands

In the model, we are essentially estimating the parameters of a cost function that looks like the following:

$$C\left(q_o(i), w, r, t\right) = F + A_o w_o^{\alpha_i} r_o^{\beta_i} q_o(i)$$

where w_o is the local wage (in levels), r_o is the local land price (in levels), and A_o is the local productive amenity. The parameters α_i, β_i, and γ_i are firm-specific and estimated in the BLP routine.

From Sheppard's Lemma, we know that the partial derivative of the cost function with respect to the factor price will give us the cost-minimizing input demand functions:

$$\begin{aligned} L^*\left(q_o(i), w, r, t\right) &= \frac{\partial C}{\partial w} = \alpha_i A_o w_o^{\alpha_i - 1} r_o^{\beta_i} q_o(i) \\ &= \alpha_i \left[\frac{m_o(i)}{w_o}\right] q_o(i) \\ T^*\left(q_o(i), w, r, t\right) &= \frac{\partial C}{\partial r} = \beta_i A_o w_o^{\alpha_i} r_o^{\beta_i - 1} q_o(i) \\ &= \beta_i \left[\frac{m_o(i)}{r_o}\right] q_o(i) \end{aligned}$$

Plugging in the expression for output, (1.21), we can rewrite these demand functions as follows:

$$L^*\left(q_o(i), w, r, t\right) = \alpha_i \theta_k \left[\frac{m_o(i)^{1-\sigma_k}}{w_o}\right] RMP_o \tag{1.23}$$

$$T^*\left(q_o(i), w, r, t\right) = \beta_i \theta_k \left[\frac{m_o(i)^{1-\sigma_k}}{r_o}\right] RMP_o \tag{1.24}$$

A similar ratio trick to (1.22) will allow us to eliminate the constant, θ_k, from these expressions.

1.B Logit Model Estimation

In the paper, I develop and estimate a choice model that allows for endogenous choice characteristics, as in the usual random coefficients logit framework (Berry et al., 1995). Each firm i is indexed by an industrial sector s and chooses one of $j = 1, ..., J_t$ locations at time $t = 1, ..., T$. Locations are either urban or non-urban locations, and this feature is indexed by $u(j) \in \{0, 1\}$. There are $i = 1, ..., N_s$ firms in each sector s. For each sector s, I take $R = 100$ scrambled Halton draws from a $N(0, 1)$ distribution to compute the random coefficients component of the choice probabilities. In practice, I use the 5-digit ISIC codes as sector identifiers. Draws are taken once for each industry and used for all firms in that industry, so that they are the same across industry-years.

Conditional on a realization of $\mathbf{v}_s = (v_{1s}, ..., v_{Ks})'$, the probability that firm i in sector s chooses location j at time t is given by:

$$\widetilde{P}_{isjt} = \frac{\exp\{x'_{jt}\beta + \xi_{jt} + \sum_{k=1}^{K} x^k_{jt}\left(\sigma_k v_{ks} + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}}{1 + \sum_{m=1}^{J_t} \exp\{x'_{mt}\beta + \xi_{mt} + \sum_{k=1}^{K} x^k_{mt}\left(\sigma_k v_{ks} + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}} \tag{1.25}$$

where we normalize the value of choosing the outside option to zero, and the unobserved choice component, ξ_{jt}, is given by:

$$\xi_{jt} \equiv \xi_j + \xi_{u(j)t} + \upsilon_{jt}$$

I have access to a census of manufacturing firms, and so assuming there is no sampling error, I have both the macro data (the total probability that firms choose a particular location at time t) as well as the micro data.

Noting that the terms $x'_{jt}\beta + \xi_j + \xi_{u(j)t} + \upsilon_{jt}$ are common to all individuals, we can write:

$$\delta_{jt} \equiv x'_{jt}\beta + \xi_j + \xi_{u(j)t} + \upsilon_{jt}$$

Crucially, the unobserved component of the mean valuation, υ_{jt}, which creates all of the estimation problems in usual random coefficient discrete choice models, is entirely subsumed within the δ_{jt}'s.

Let $\theta_1 = (\beta', \xi')'$ denote the *linear* parameters of the model, which are subsumed within the δ_{jt}'s, and let $\theta_2 = (\pi', \sigma')'$ denote the *non-linear* parameters of the model, including the coefficients on the demographic interactions as well as the standard deviation terms. To estimate $\theta = (\theta'_1, \theta 2')'$, I make use of the following 2-step estimation routine:

1. **Step 1**: Estimate the δ_{jt}'s and θ_2 using maximum simulated likelihood.

    - Although full maximum simulated likelihood is theoretically possible, in practice it is computationally infeasible. My dataset has over 100 locations, each of which are observed for possibly 15 years, so the δ_{jt} parameter space is way too large to search over. Consequently, I maximize the simulated likelihood function only over θ_2. For each value of θ_2, I choose $\delta_{jt} = \delta_{jt}(\theta_2)$ to ensure that the mean valuation components satisfy a market share constraint.

2. **Step 2**: To recover the linear parameters, θ_1, we estimate the following regression using 2SLS/GMM:

    $$\widehat{\delta_{jt}} = x'_{jt}\beta + \xi_j + \xi_{u(j)t} + \upsilon_{jt}$$

    where we use instruments for the endogenous x_{jt}'s. The method of moments estimator of θ_1 solves the sample analogues of the following moments:

    $$\mathbb{E}[\mathbf{Z}'(\delta - \mathbf{X}'\beta)] = 0$$

    where $\mathbf{Z}$ is a matrix of M instruments.

1.B.1 Interactions

Note that in order for the procedure to work, we need consistent estimation of the δ_{jt}'s, which are the *mean valuation* parameters. When constructing the interaction terms, care must be taken to ensure that the δ_{jt}'s accurately reflect mean valuation and not the value of the omitted group.

In the estimation, I created indicators for each group (2-digit industry) as follows. Index sectoral groups by $d = 1, ..., D, D+1$, and let the last group, $D+1$, denote the omitted sector group. Define $\widetilde{D_{id}}$ as an indicator for whether firm i belongs to industrial sector d. Then, define:

$$D_{id} = \widetilde{D}_{id} - \overline{\widetilde{D}_{id}} = \begin{cases} 1 - \frac{1}{N}\sum_{i=1}^{N} D_{id} & i \text{ is in group } d \\ -\frac{1}{N}\sum_{i=1}^{N} D_{id} & \text{else} \end{cases}$$

This just amounts to demeaning the group indicators. To see why this works, it is helpful to consider a linear model. For plants in the included group d, the expected value of an outcome variable y_{ijt} is given by:

$$\begin{aligned} \mathbb{E}[\, y_{ijt} \,|\, \mathbf{x}, i \in d\,] &= \delta_{jt} + \sum_{k=1}^{K} \left(\pi_{k,1}\mathbb{E}[\, D_{i1} \,|\, i \in d\,] + ... + \pi_{k,D}\mathbb{E}[\, D_{iD} \,|\, i \in d\,]\right) x_k \\ &= \delta_{jt} + \sum_{k=1}^{K} \left(\pi_{k,d} - \pi_{k,1}\mu_1 - ... - \pi_{k,D}\mu_D\right) x_k \\ &= \delta_{jt} + \sum_{k=1}^{K} \left(\pi_{k,d} - \mu'\pi_{\mathbf{k}}\right) x_k \end{aligned}$$

where $\pi_{\mathbf{k}} = (\pi_{k,1}, \pi_{k,2}, ..., \pi_{k,D})$ is a $(D \times 1)$ vector of coefficients on the included sectoral interaction terms for variable k, and $\mu = (\mu_1, \mu_2, ..., \mu_D)$ is a $(D \times 1)$ vector collecting the probabilities that firms are members of each group. Hence, to uncover mean parameters for included sector group d, we simply add the mean parameters β to the interaction term $\pi_{k,d}$, then subtract $\mu'\pi_k$.

For the omitted group $D + 1$, note that in expectation, $D_{id} = -\mu_d$ for all included groups, $d \neq D$, so we have:

$$\mathbb{E}[\, y_{ijt} \,|\, \mathbf{x}, d = D + 1\,] = \delta_{jt} - \sum_{k=1}^{K} \left(\mu'\pi_{\mathbf{k}}\right) x_k$$

So, to uncover the mean parameters for the omitted group, we subtract $\mu'\pi_k$ from the mean parameters, β.

The goal of this exercise is not to estimate parameters on the demeaned interaction terms, π_k, but to instead estimate parameters for each industry:

$$\gamma_{kd} \equiv \begin{cases} \beta + \pi_{kd} - \mu'\pi_k & d \in \{1, ..., D\} \\ \beta - \mu'\pi_k & d = D + 1 \end{cases}$$

To construct these parameters and perform inference on the γ's, we make use of the so-called Delta method. Specifically, we will show later that the estimated parameters, $\widehat{\beta}_k = (\beta_k, \pi_{k,1}, ..., \pi_{k,D-1})'$, are asymptotically normal:

$$\sqrt{N}(\widehat{\beta}_k - \beta_k) \xrightarrow{d} N(0, \mathbf{V})$$

Define:

$$\underset{(D\times D)}{\mathbf{R}} = \begin{bmatrix} 1 & (1-\mu_1) & -\mu_2 & \dots & -\mu_{D-1} \\ 1 & -\mu_1 & (1-\mu_2) & \dots & -\mu_{D-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -\mu_1 & -\mu_2 & \dots & (1-\mu_{D-1}) \\ 1 & -\mu_1 & -\mu_2 & \dots & -\mu_{D-1} \end{bmatrix}$$

It is easy to see that $\mathbf{R}\widehat{\beta}_k = \widehat{\gamma}_k$, the vector of mean utilities plus sectoral group parameters, which are the estimates we care about.

1.B.2 Details: Step 1

Let $\mathbf{j}(s) = (j(1), j(2), ..., j(N_s))$ denote a sequence of choices for firms $i = 1, ..., N_s$ in sector s. Since we take the same draws for each sector, the unconditional probability of observing a sequence of choices, $\mathbf{s}$, for firms in sector s is given by:

$$P_{\mathbf{j}(s)} = \int \prod_{i=1}^{N_s} \left(\frac{\exp\{\delta_{j(i)t} + \sum_{k=1}^{K} x_{j(i)t}^{k} \left(\sigma_k v_{ks} + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}}{1 + \sum_{m=1}^{J_t} \exp\{\delta_{mt} + \sum_{k=1}^{K} x_{mt}^{k} \left(\sigma_k v_{ks} + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}} \right) dF(\mathbf{v}_s)$$

where $F(\mathbf{v}_s) = F(v_{1s}, ..., v_{Ks})$ is the joint CDF of the distribution of the unobserved components. This integral is not computed analytically, but is instead approximated by simulation. For each industry, we draw $r = 1, ..., R$ values of $\mathbf{v}_s$, and each vector of draws is denoted by $\mathbf{v}_s^r = (v_{1s}^r, ..., v_{Ks}^r)$. In practice, we take $R = 100$ Halton sequence draws from a standard normal distribution for each element of the vector. We then approximate each sector's sequence of choice probabilities by the following:

$$\begin{aligned} \widetilde{P}_{\mathbf{j}(s)} &= \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{i=1}^{N_s} \left(\frac{\exp\{\delta_{j(i)t} + \sum_{k=1}^{K} x_{j(i)t}^{k} \left(\sigma_k v_{ks}^r + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}}{1 + \sum_{m=1}^{J_t} \exp\{\delta_{mt} + \sum_{k=1}^{K} x_{mt}^{k} \left(\sigma_k v_{ks}^r + \pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)\}} \right) \right] \\ &= \frac{1}{R} \sum_{r=1}^{R} PROD_{\mathbf{j}(s)}^{r} \end{aligned}$$

where $PROD_{\mathbf{j}(s)}^{r}$ is the probability of a sequence of choices conditional on vectors of draws $\mathbf{v}_s^r$ for each variable.

The simulated log-likelihood function is formed in the usual way:

$$SLL\big(\theta_2, \delta(\theta_2)\big) = \sum_{s=1}^{S} \ln \widetilde{P}_{\mathbf{j}^*(s)}\big(\theta_2, \delta(\theta_2)\big)$$

where $\mathbf{j}^*(s)$ denotes the vector of location choices that were actually chosen by firms in sector s. Note that the number of individuals choosing at time period t, and the choice set of locations at time period t, J_t, varies over time. I'm estimating the choice model on the sample of new firms each year, and not all locations are chosen each period, which is why the size of choice sets and the number of firms change each year.

I maximize the simulated likelihood function over $\theta_2 = (\pi', \sigma')'$, but at each iteration, I first calculate the predicted market shares:

$$\widehat{S}_{jt}(\theta, \delta) = \frac{1}{N_t} \frac{1}{R} \sum_{i=1}^{N_t} \sum_{r=1}^{R} \frac{\exp\{\cdot_{ijt}(\theta, \delta, \mathbf{v}^r)\}}{1 + \sum_{m=1}^{J_t} \exp\{\cdot_{imt}(\theta, \delta, \mathbf{v}^r)\}}$$

Then, I solve for the δ_{jt}'s that equate actual market shares with predicted shares, using the standard BLP contraction mapping:

$$\delta_{jt}^{H+1} = \delta_{jt}^{H} + \ln S_{jt} - \ln \widehat{S}_{jt}(\theta, \delta)$$

The contraction mapping reduces the dimensionality of the parameter space considerably, but this creates some additional complications when computing the gradient.

Since we estimate θ_2 and δ_{jt} conditional on θ_2, we have to be careful when computing the score of the likelihood function with respect to θ_2. We need to account for the fact that changing θ_d also changes the $\delta'_{jt}s$:

$$\frac{dSLL\big(\theta_2, \delta(\theta_2)\big)}{d\theta_2} = \underbrace{\frac{\partial SLL}{\partial \theta_2}}_{(1)} + \underbrace{\frac{\partial SLL}{\partial \delta}}_{(2)} \cdot \underbrace{\frac{\partial \delta}{\partial \theta_2}}_{(3)}$$

To simplify exposition in the discussion that follows, I'm going to subsume all of the interaction terms in one vector. Let X_{ijt} denote the $(1 \times (K \times D))$ vector of choice characteristics interacted with demographic characteristics (and demeaned) that each individual i faces when choosing location j at time t:

$$X_{ijt} \equiv \left[\left(x_{jt}^1 D_{i1}, ..., x_{jt}^1 D_{iD}\right), \left(x_{jt}^2 D_{i1}, ..., x_{jt}^2 D_{iD}\right), ..., \left(x_{jt}^K D_{i1}, ..., x_{jt}^K D_{iD}\right)\right]$$

This notation lets us write the following:

$$\Pi X_{ijt} = \sum_{k=1}^{K} x_{jt}^k \left(\pi_{k1} D_{i1} + ... + \pi_{kD} D_{iD}\right)$$

We can also do the same thing for the choice characteristics interacted with the simulation draws:

$$V_{ijt}^{s} \equiv \left[x_{jt}^{1} v_{i}^{s}, x_{jt}^{2} v_{i}^{s}, ..., x_{jt}^{K} v_{i}^{s}\right]$$

This is a $(1 \times K)$ vector, unique for each sector s and year t.

Gradient, First Term

The first term of the gradient is straightforward to compute:[52]

$$\frac{\partial SLL}{\partial \theta_2} = \sum_{s=1}^{S} \frac{1}{\widetilde{P}_{\mathbf{j}^*(s)}} \left[\frac{1}{R} \sum_{r=1}^{R} \left\{PROD_{\mathbf{j}^*(s)}^{r} \cdot \frac{\partial \ln PROD_{\mathbf{j}^*(s)}^{r}}{\partial \theta_2}\right\}\right]$$

For the $((K \times D) \times 1)$ vector of demographic parameters, Π, the derivative of the log of the product of the simulated choice probabilities for sector s and simulation r is given by:

$$\frac{\partial \ln PROD_{\mathbf{j}^*(s)}^{r}}{\partial \Pi} = \sum_{i=1}^{N_s} \left(X_{ij(i)^*t} - \sum_{k=1}^{J_t} \left(\frac{\exp\{\cdot_{ikt}^{r}\}}{1 + \sum_l \exp\{\cdot_{ilt}^{r}\}}\right) X_{ikt}\right)$$

Similarly, for the random coefficients portion, we have:

$$\frac{\partial \ln PROD_{\mathbf{j}^*(s)}^{r}}{\partial \Sigma} = \sum_{i=1}^{N_s} \left(V_{ij(i)^*t} - \sum_{k=1}^{J_t} \left(\frac{\exp\{\cdot_{ikt}^{r}\}}{1 + \sum_l \exp\{\cdot_{ilt}^{r}\}}\right) V_{ikt}\right)$$

To compute $\partial SLL/\partial \theta_2$, we first compute the partial of the log product of the simulated choice probabilities for each sector s and simulation r. We then interact this with $PROD_{\mathbf{j}^*(s)}^{r}$, then take averages over the simulations. Finally, we divide by $\widetilde{P}_{\mathbf{j}^*(s)}$ and then sum across sectors.

Gradient, Second Term

The second term in the gradient is similar to the first:

$$\frac{\partial SLL}{\partial \delta} = \sum_{s=1}^{S} \frac{1}{\widetilde{P}_{\mathbf{j}^*(s)}} \left[\frac{1}{R} \sum_{r=1}^{R} \left\{PROD_{\mathbf{j}^*(s)}^{r} \cdot \frac{\partial \ln PROD_{\mathbf{j}^*(s)}^{r}}{\partial \delta}\right\}\right]$$

[52] Note that here, we make use of this fact:

$$\frac{\partial y}{\partial x} = y \frac{\partial \ln y}{\partial x}$$

This helps us to evaluate the derivative of the product of probabilities.

The derivative of the log of the product of the simulated choice probabilities for sector s and simulation r with respect to δ is given by:

$$\frac{\partial \ln PROD^r_{\mathbf{j}^*(s)}}{\partial \delta} = \sum_{i=1}^{Ns} \frac{\partial \ln P^r_{ijt}}{\partial \delta}$$

where P^r_{ijt} is the conditional logit choice probability, given by an expression similar to (1.25). To compute this derivative, it is helpful to introduce some more notation. Let D_{ijt} be an indicator equal to 1 if firm i chose location j at time t and zero otherwise. This derivative is equal to the following:

$$\frac{\partial \ln P^r_{ijt}}{\partial \delta_{jt}} = \begin{cases} D_{ijt} - P^r_{ijt} & \text{if } i \text{ chooses at time } t \\ 0 & \text{if } i \text{ chooses at time } s \neq t \end{cases}$$

It is helpful to compute this derivative year-by-year. For a given year t, we first compute $\partial \ln P^r_{ijt}/\partial \delta$. We then sum this across all individuals in the sector to obtain $\partial \ln PROD^r_{\mathbf{j}^*(s)}/\partial \delta$. We interact this with $PROD^r_{\mathbf{j}^*(s)}$ and average across simulations, then take averages across sectors.

Gradient, Third Term

To obtain an expression for $\partial \delta / \partial \theta_2$, we proceed by remembering that δ is implicitly defined by θ_2 as the solution to:

$$S_{jt} - \widehat{S}_{jt}(\theta_2, \delta) = 0$$

Taking derivatives with respect to θ_2, using the chain-rule, and rearranging, we have:

$$\begin{aligned} 0 = \frac{d\widehat{S}_{jt}}{d\theta_2} &= \frac{\partial \widehat{S}_{jt}}{\partial \theta_2} + \frac{\partial \widehat{S}_{jt}}{\partial \delta} \cdot \frac{\partial \delta}{\partial \theta_2} \\ \implies \frac{\partial \delta}{\partial \theta_2} &= -\underbrace{\left(\frac{\partial \widehat{S}_{jt}}{\partial \delta}\right)^{-1}}_{(A)} \underbrace{\frac{\partial \widehat{S}_{jt}}{\partial \theta_2}}_{(B)} \end{aligned}$$

Note that $\partial \widehat{S}_{jt}/\partial \delta$ is a $(N_a \times N_a)$ matrix, and $\partial \widehat{S}_{jt}/\partial \theta$ is a $(N_a \times K)$ matrix.

Gradient, Third Term, Part (A)

Let $\widehat{\mathbf{S}} = \left(\widehat{S}_{11}, ..., \widehat{S}_{J_T T}\right)'$ denote the $(N_a \times 1)$ vector of market shares. Although $\partial \widehat{\mathbf{S}}/\partial \delta$ is a large $(N_a \times N_a)$ matrix, fortunately many of its elements are zeros, because:

$$\frac{\partial \widehat{S}_{jt}}{\partial \delta_{ks}} = 0 \text{ if } t \neq s$$

Hence, the matrix is block diagonal. To compute this matrix, define $\widetilde{P}_{ijt}$ to be the average simulated choice probability for firm i:

$$\widetilde{P}_{ijt} = \frac{1}{R}\sum_{r=1}^{R} \frac{\exp\{\cdot^r_{ijt}\}}{1 + \sum_l \exp\{\cdot^r_{ilt}\}}$$

Then, a typical element of this matrix of derivatives is given by:

$$\begin{aligned}\frac{\partial \widehat{S}_{jt}}{\partial \delta} &= \frac{1}{N_t}\sum_{i=1}^{N_t} \frac{\partial \widetilde{P}_{ijt}}{\partial \delta_{jt}} \\ &= \frac{1}{N_t}\frac{1}{R}\sum_{i=1}^{N_t}\sum_{r=1}^{R}\left\{ P^s_{ijt} D_{ijt} - P^r_{ijt}\sum_{k=1}^{J_t} P^r_{ikt} D_{ijt}\right\}\end{aligned}$$

The full matrix, $\partial \widehat{\mathbf{S}}/\partial \delta$, is an $(N_a \times N_a)$ matrix of derivatives. The computation of this portion of the gradient is also done year-by-year because of the block-diagonal structure.

Gradient, Third Term, Part (B)

To form $\partial \widehat{S}_{jt}/\partial \theta_2$, the $(N_a \times K)$ matrix of partial derivatives, note that we have:

$$\frac{\partial \widehat{S}_{jt}}{\partial \theta_2} = \frac{1}{N_t}\sum_{i=1}^{N_t} \frac{\partial \widetilde{P}_{ijt}}{\partial \theta_2}$$

For the demographic parameters, we have:

$$\frac{\partial \widetilde{P}_{ijt}}{\partial \Pi} = \frac{1}{R}\sum_{r=1}^{R}\left\{ P^r_{ijt} X_{ijt} - P^r_{ijt}\sum_{k=1}^{J_t} P^r_{ikt} X_{ikt}\right\}$$

For the choice characteristics interacted with the simulation draws, we have:

$$\frac{\partial \widetilde{P}_{ijt}}{\partial \Sigma} = \frac{1}{R}\sum_{r=1}^{R}\left\{ P^r_{ijt} V^r_{ijt} - P^r_{ijt}\sum_{k=1}^{J_t} P^r_{ikt} V^r_{ikt}\right\}$$

To compute this term, we first compute $\partial \widetilde{P}_{ijt}/\partial \Pi$ and $\partial \widetilde{P}_{ijt}/\partial \Sigma$ for each individual. We then average this over all individuals who faced a choice situation with location j at time t.

1.B.3 Standard Errors

To get appropriate standard errors, we characterize the estimation procedure as a two-step estimator and use asymptotic GMM approximations, stacking the moments from each step. For the non-linear parameters, θ_d', the method of moments estimator sets the sum of the scores of the log-likelihood equal to zero. Let W_{ijt} collect all variables used in the first step (i.e. choice and time indicators, interactions of choice characteristics with firm characteristics). The method of moments estimator solves the following sample moment condition:

$$\Psi_1(\theta_2, \delta(\theta_2)) = LL_{\theta_2}(\theta_2, \delta(\theta_2)) = 0$$

To allow for clustered standard errors at the region level (Arellano, 1987), the second sample moment is the following:

$$\Psi_2(\delta(\theta_2), \beta) = \sum_{r=1}^{R} \widetilde{\mathbf{Z}}_r \left(\delta_r(\theta_2) - \widetilde{\mathbf{X}}_r' \beta \right)$$

where $\widetilde{\mathbf{Z}}_r$ denotes a vector that stacks the history of demeaned instruments for region r, and $\widetilde{\mathbf{X}}_r$ and $\delta_r(\theta_2)$ are defined similarly. Define $\theta \equiv (\theta_2', \beta')'$ to be a vector collecting all of the parameters estimated directly in the model. Estimating θ with GMM, we have the usual asymptotic results:

$$\sqrt{N}\left(\widehat{\theta}_{GMM} - \theta_0 \right) \xrightarrow{d} N(0, \mathbf{V}_0)$$

where

$$\mathbf{V}_0 = \left(\mathbf{G}_0' \mathbf{C_0} \mathbf{G}_0 \right)^{-1} \mathbf{G}_0' \mathbf{C}_0' \mathbf{\Lambda}_0 \mathbf{C}_0 \mathbf{G}_0 \left(\mathbf{G}_0' \mathbf{C_0} \mathbf{G}_0 \right)^{-1}$$

and we have:

$$\mathbf{\Lambda}_0 = \mathbb{E}_N \begin{bmatrix} \Psi_1 \Psi_1' & \Psi_1 \Psi_2' \\ \Psi_2 \Psi_1' & \Psi_2 \Psi_2' \end{bmatrix}$$

and

$$\mathbf{G}_0 = \mathbb{E} \begin{bmatrix} \partial \Psi_1 / \partial \theta_2 & \partial \Psi_1 / \partial \beta \\ \partial \Psi_2 / \partial \theta_2 & \partial \Psi_2 / \partial \beta \end{bmatrix}$$

and $\mathbf{C_0}$ is a weighting matrix, set to $\mathbf{I}$ because of the 2-step nature of the computation.[53]

Note that $\mathbf{\Lambda}_0$ is easily computable. As for $\mathbf{G}_0$, note that the upper right term in the matrix, $\partial \Psi_1 / \partial \beta$, is zero. Morever, the upper left term in the matrix, $\partial \Psi_1 / \partial \theta_2$, is just the Hessian of the log likelihood function, $\mathbf{H}(\theta_2)$, which is returned in the estimation procedure. The bottom right term, $\partial \Psi_2 / \partial \beta$, is just $-\sum_{r=1}^{R} \widetilde{\mathbf{Z}}_r' \widetilde{\mathbf{X}}_r$.

The only term that is challenging is the bottom left term:

$$\frac{\partial \Psi_2}{\partial \theta_d} = \sum_{r=1}^{R} \widetilde{\mathbf{Z}}_r' \left(\frac{\partial \delta_r(\theta_2)}{\partial \theta_2} \right)$$

However, we solved for the $\partial \delta(\theta_2) / \partial \theta_2$ matrix above in computing the third term of the gradient (parts A and B). So, we have:

$$\widehat{\mathbf{G}} = \begin{bmatrix} \mathbf{H}(\theta_2) & \mathbf{0} \\ \sum_{r=1}^{R} \widetilde{\mathbf{Z}}_r' \left(\frac{\partial \delta_r(\theta_2)}{\partial \theta_2} \right) & -\sum_{r=1}^{R} \widetilde{\mathbf{Z}}_r' \widetilde{\mathbf{X}}_r \end{bmatrix}$$

[53] Since I am not using the optimal GMM weight matrix, this procedure is inefficient.

1.C Data Appendix

1.C.1 Road Quality Data

Data on the quality of Indonesia's highway networks were produced by DPU as part of Indonesia's Integrated Road Management System (IRMS). This appendix section begins by providing some background on road management in Indonesia, describing the road classification system and discussing IRMS coverage. It then discusses the measures of road quality that are collected in IRMS and how they are measured. I then discuss how the road network data were created.

Background on Road Management

Indonesia's national road network is currently managed and maintained by the Department of Public Works (*Departemen Pekerjaan Umum*, DPU), specifically by the Directorate General of Highways (*Direktorat Jenderal Bina Marga*). According to Law No. 38, 2004, roads are classified into four different types of roads, primarily based on their function for users. Arterial roads (*jalan arteri*) serve as the major transportation linkages between urban areas, and are characterized by longer distances, higher speeds, and limited access. Speeds are meant to be a minimum of 60 km/h, and width should be at least 11 meters to accommodate larger traffic volumes. Collector roads (*jalan kolektor*) serve "collector or distributor transportation" and are characterized by medium distance travel with medium speeds. Collector roads are subdivided into primary collector roads (*jalan kolektor primer*), which should have a minimum speed of 20 km/h and width of 9 meters, and secondary collector roads, which should have a minimum speed of 20 km/h and width of 9 meters. Local roads (*jalan lokal*) and Neighborhood Roads (*jalan lingkungan*) serve local areas at lower speeds, and are characterized by unlimited access.

Roads can also be classified by their management authority, or "status" (*wewenang penyelenggaraan*). Generally, arterial and primary collector roads are managed by the national government (specifically by DPU). Secondary and tertiary collector roads are managed by provincial governments, while local and neighborhood roads are managed by the kabupaten, kecamatan, and desa governments. Table 1.C.1 describes the road classification system, minimum speed and width guidelines, and management authorities.

Table 1.C.2 depicts the coverage of the IRMS dataset by road function and managing authority, as measured by counts of the number of kilometer-post observations that appear in the entire dataset. Most of the observations, and indeed most of the road network, is made up by collector roads (K1-K3), though the category with the next largest coverage is the arterial roads. Local and neighborhood roads are not very well surveyed in this dataset. Although the network of village and kabupaten roads is doubtless extremely dense, I cannot use this dataset to say very much about it. But since the data do cover arterial and collector roads, the major roads connecting regions and cities in Indonesia, this dataset seems particularly well suited for evaluating models of economic geography and regional trade.

Measures of Road Quality

There are a number of different devices that transport engineers have developed to collect measurements of road quality, and there are several different measures of road quality. The most widely used measure of road roughness, and the measure used in this study, is the international roughness index (IRI), developed by the World Bank in the 1980s. IRI is constructed as a filtered ratio of a standard vehicle's accumulated suspension motion (in meters), divided by the distance travelled by the vehicle during measurement (in kilometers). Expressed in units of slope (m/km), IRI is a characteristic of a vehicle's longitudinal profile. Importantly, since it is a measure of a physical quantity, IRI is standardized, as opposed to other subjective measures of ride quality. Figure 1.C.1 shows the relationship between different ranges of IRI and surface type; generally, larger roughness levels correspond to worse surfaces, but the mapping is not one-to-one.

Bennett et al. (2007) distinguish between several different types of devices for measuring road roughness and provide a good overview of their relative strengths and weaknesses. Over the course of its existence, Indonesia's IRMS has largely made use of two different types of measuring devices.[54] Before 1999, roads were surveyed using devices like the ROMDAS, which estimate IRI indirectly. The ROMDAS machine is a calibrated bump integrator, which must first be calibrated and estimates IRI from correlation equations. It is very useful for measuring roughness on bumpy roads and can record high levels of IRI, but the device must be calibrated manually, and measurement error can occur if the device is miscalibrated.

The ROMDAS device is also portable, meaning that it can be used inside different vehicles (each of which would require unique calibrations). The portability contrasts with devices like the high-speed laser profilometer, which is essentially a separate vehicle reserved entirely for the purposes of collecting road quality data. The device uses lasers and optical techniques to scan the road as it is traversed and create measures of surface profiles. These instruments are very accurate, but are much more expensive. Moreover, they might become mis-calibrated on extremely rough roads. Indonesia started using the high speed laser profilometer for collecting its road quality data in 1999, licensing vehicles from the Australian government.

Road width and surface type are more straightforward variables to measure, involving visual inspection and simple measurement. I categorize a kilometer-post interval as being unpaved if it is either an earth, gravel, or sand road, or if it was given a granular base (crushed stone) treatment, a first step in the process of paving.

Creation of Road Network Data

Using GIS shapefiles of the road network provided to me by DPU, I have georeferenced the kilometer post observations of road quality, in order to capture the evolution of Indonesia's transportation network over space and time. This proved to be a challenging exercise, because the identifiers for

[54] I am very grateful for the extensive discussions I've had with Glen Stringer about IRMS; this section of the appendix benefits highly from our conversations.

each road-link-interval observation were not consistent over time, and because the identifiers in the shapefile and in the linearly referenced dataset were often different, even though both did refer to exactly the same link.

Once the IRMS interval data was successfully merged to the regional network shapefiles, I converted the GIS database of road links into a weighted graph of arcs and nodes, as commonly used in the transportation literature. Nodes represent locations (such as ports, cities, or the centroids of kabupatens, my unit of analysis), arcs represent the possibility of traveling between two nodes, and weights represent the cost of moving goods along a given arc. Weights were constructed according to the IRMS data on road quality, and for simplicity, the cost of moving along each road was assumed to be the same, no matter which way you were traveling.[55]

For computational reasons, I have used a simplified representation of Indonesia's road network, where the number of nodes and links was small enough for network algorithms to operate on it using a desktop computer.[56] Table 1.C.3 depicts the number of network arcs, the total distance of the network, and merge statistics for the kilometer-post observations. Merge statistics are pretty good for arterial and collector roads, but the quality of merges falls substantially for local and neighborhood roads, due most likely to poor shapefile coverage for that type of road network.

The interval observations were not matched directly to their exact locations in the network, because I had no knowledge of the exact location of the kilometer posts. To deal with this, I first aggregated the kilometer-post interval observations to the road-link level by constructing distance-weighted averages of the road quality variables. Each network arc-year observation was then assigned the value of this average road quality variable that corresponds to its road link.[57]

Roughness, Speed, and Ride Quality

One effect that rough roads have on vehicles is that they require the driver to travel at lower speeds. When faced with potholes, ragged pavement, or poor surfaces, drivers slow down, and this reduction in speed increases travel time and hence the cost of travel. Of course, there is not a one-to-one relationship between road roughness and speed, because drivers choose the speed at

[55] Another tedious issue involved the construction of junction points where the road links intersected. The shapefiles were originally stored as MapInfo files, an older shapefile format that required conversion for use with Arcview, and in this conversion, information on where the roads crossed was lost, requiring painstaking editing. The shapefiles were also not designed to be used in any network analysis, so much care had to be taken to make them usable.

[56] The road lines were straightened using the "Generalize" command from ET Geotools, which employs the RamerDouglasPeucker algorithm for reducing the number of points that represents a line.

[57] In some cases, when a network arc had no data for a particular year, I assigned the network arc the average value of road quality for arcs with the same function. This was done because constructing the transport cost variables involved a search over the entire network, and if certain network arcs were coded as missing, this could distort the search substantially. Overall, imputation amounted to no more than 5 percent of network arc observations in any given year.

which they travel, and different preferences for smoothness of the ride or the desired arrival time might induce different choices of speed.

Yu et al. (2006) explore the relationship between *jolt*, or the "jerk" experienced by road users, and subjective measures of ride quality and road roughness at different speeds.[58] Using survey data in which users were asked to rate the quality of particular rides, the authors find that people experience greater discomfort while traveling at higher speeds on rough roads, but lowering speed on rough roads can reduce discomfort. The authors provide a mapping between subjective measures of ride quality and roughness at different speeds, and this mapping can be used to infer the maximum speed that one can travel in order to achieve a ride of a certain quality, given pavement roughness. Table 1.C.4 reproduces this mapping. Because travel times were unreasonably long for high quality rides given Indonesia's rough roads, and because the subjective quality measures were chosen by Western drivers, I have focused on the poor ride quality speed thresholds in my empirical work.

Given the maximum speed that one can travel on roads of different roughness levels, it is straightforward to calculate travel times for each network arc, the primary measure of transport costs used in this study. Note that the travel times on road sections were computed using the detailed kilometer-post interval roughness data. These were then aggregated to the network arcs using distance-weighted averages.

Shortest Paths Between Kabupaten Centroids

Given the distance and travel times associated with traversing each network arc, constructing the shortest path between points on the network is straightforward, using Djekstra's shortest path algorithm. Although the network of inter-urban roads is fairly dense, kabupaten centroids were generally not directly connected to the network. When this was the case, a small segment was added to the network connecting the centroid to the closest road junction point, on the assumption that the network of local and neighborhood roads is sufficiently dense for this to be a reasonable approximation. The shortest-path search was then conducted on this augmented network.

Ports and Inter-Island Transportation

The locations of major ports in Indonesia come from DPU. Travel times between ports were calculated using physical distances and assuming a constant speed of 20 knots per hour, typical of major cargo ships. The entire matrix of transportation between regions of Java, Sumatra, and Sulawesi is depicted visually in Figure 1.C.2.

1.C.2 Administrative Boundaries

Administrative boundary shapefiles were constructed by BPS for use during the 2000 Household Census. These shapefiles contain the polygon boundaries of all provinces, kabupatens, kecamatans,

[58] *Jolt* is officially defined as the vector that specifies the time-derivative of acceleration; in other words, the third derivative of the vertical displacement of vehicle to time t.

and desas for the entire extent of the Indonesian archipelago. However, after the fall of Suharto and a massive decentralization program, many new kabupatens were created, splitting existing kabupatens into new ones. For instance, in 1990 there were 290 kabupatens and kotas, but by 2003, there were 416 kabupatens and kotas. The fact that administrative boundaries are not fixed over time create difficulties for the analysis.

Because of the need for a geographic unit of analysis that was consistently defined over time, I used kabupaten borders as they were defined in 1990. BPS provided the administrative boundary shapefile for 2000, as well as a correspondence table between kabupaten codes in 2000 and kabupaten codes from 1990 to the present. This information was processed using ArcView to create the 1990 shapefiles that form the basis of the analysis. Throughout the paper, all survey data were appropriately merged back to the 1990 kabupaten definitions.

Choice Set Aggregation

For many years, there were only a handful (1-2) of new firms that located in certain remote kabupatens, but in other years, those kabupatens were never again reached. In the choice model that I estimate, having a location appear for only one or two years and then disappear completely creates a substantial amount of noise. Moreover, some of the BPS definitions of kabupatens in Indonesia are fairly arbitrary; for instance, the city of Jakarta is split into five separate kabupatens, but there doesn't seem to be any reason to treat North Jakarta and Central Jakarta as separate regions in the analysis. For these reasons, I decided to aggregate some of the kabupatens, as follows:

- **Java**
  - Jakarta is split up into 5 separate kabupatens (3171 Kota Jakarta Selatan, 3172, Kota Jakarta Timur; 3173, Kota Jakarta Pusat; 3174, Kota Jakarta Barat; 3175, Kota Jakarta Utara), and these were combined to form a single urban region (3100).
  - The island of Madura is also split up into 4 separate kabupatens (3326, Kab. Pekalongan; 3327, Kab. Pemalang; 3328, Kab. Tegal; 3329, Kab. Brebes), and these were combined into a single region (3526).
- **Sumatra**
  - Aceh is divided into 8 separate kabupatens. Due to the small amounts of manufacturing activity in these areas, we chose to combine Aceh Utara (1108) and Aceh Tengah (1104) into one region (1104), and three coastal kabupatens of Aceh Barat (1105), Aceh Besar (1106), and Pidie (1107) into another region (1105).
  - Because they include groups of islands on the west coast of Sumatra and contained small levels of manufacturing, we combined Kabupaten Nias (1201) and Kabupaten Padang Pariaman (1305) into one region (1201). This effectively combines Nias with Mentawai Islands.
  - The adjacent kabupatens of Tapanuli Selatan (1202) and Pasaman (1308) were combined due to small amounts of manufacturing activity.

  - The adjacent kabupatens of Dairi (1208) and Karo (1209) were combined due to small amounts of manufacturing activity.
  - Due to their proximity and relatively small levels of manufacturing activity, the cities of Padang (1371), Solok (1372) and Sawah Lunto (1373) into one urban region (1371).
  - The adjacent kabupatens of Kampar (1404) and Lima Puluh Koto (1307) were combined into one region (1404).
  - The adjacent kabupatens of Kerinci (1501) and Bungo Tebo (1502) were combined into one region (1501).
  - The rural kabupaten of Ogan Komering Ulu (1601) and Ogan Komering Ilir (1602) were combined into one region (1601).
  - The adjacent kabupatens of Maura Enim (1603) and Musi Banyu Asin (1606) were combined into one region (1603).
  - The adjacent kabupatens of Rejang Lebong (1702) and Bengkulu Utara (1703) were combined into one region (1702).
  - The coastal kabupatens of Bengkulu Selatan (1701) and Lampung Barat (1804) were combined into one region (1701) due to small levels of manufacturing activity.

- **Sulawesi**
  - Due to their proximity and relatively small levels of manufacturing activity, the cities of Manado (7172) and Bitung (7173) were combined into one urban region (7172).
  - Four kabupatens that make up the entire province of Central Sulawesi (7201, Kab. Banggai; 7202, Kab. Poso; 7203, Kab. Donggala; 7204, Kab. Buol Toli-Toli) were combined into one region (7200).
  - Three kabupatens (7301, Kab. Selayar; 7302, Kab. Bulukumba; 7303, Kab. Bantaeng) were combined into one region (7301).
  - Two kabupatens (7305, Kab. Takalar; 7306, Kab. Gowa) were combined into one region (7305).
  - Two kabupatens (7309, Kab. Pangkajene Kepulauan; 7310, Kab. Barru) were combined into one region (7309).
  - Three kabupatens (7316, Kab. Enrekang; 7317, Kab. Luwu; 7318, Kab. Tana Toraja) were combined into one region (7316).
  - Three kabupatens (7319, Kab. Polewali Mamasa; 7320, Kab. Majene; 7321, Kab. Mamuju) were combined into one region (7319).

While this might seem like a great deal of changes, in the end, this amounted to changing the location identifiers for approximately 13.4 percent (42,561 observations) of firm-year observations for 1990-2005. If we ignore the changes that come from reclassifying firms locating in Jakarta, the aggregation amounted to changing only 2.1 percent (6,660 observations) of the location identifiers for firm-year observations.

1.C.3 Spatial, Topographical, and Agro-climatic Variables

Agricultural and climatic variables were created from a variety of sources and often were calculated with the assistance of GIS software (ArcView). This section describes those data in detail and how each of the variables were constructed.

Map Projection

To compute distances correctly, using linear units of measurement (i.e. meters of kilometers), I made use of the Batavia Transverse Mercator (TM) 109 SE projected coordinate system in all of my GIS work. Specific details on the map projection for use with ArcView or other GIS software are the following:

```
Projected Coordinate System: Batavia_TM_109_SE
Projection: Transverse_Mercator
False_Easting: 500000.00000000
False_Northing: 10000000.00000000
Central_Meridian: 109.00000000
Scale_Factor: 0.99960000
Latitude_Of_Origin: 0.00000000
Linear Unit:  Meter

Geographic Coordinate System: GCS_Batavia
Datum:  D_Batavia
Prime Meridian:  Greenwich
Angular Unit:  Degree
```

Slope, Aspect, and Elevation Data

Topographical variables were created using raster data from the Harmonized World Soil Database (HWSD), Version 2.0.[59] The raster files are compiled from high-resolution source data and aggregated to 30 arc-second grids (approximately 1 km^2 cells).

Elevation data were computed for each administrative boundary polygon as the average elevation over the entire polygon. They were also computed for each centerline GPS coordinate, and

[59] Data from the HWSD project are publicly available and can be downloaded here: `http://www.iiasa.ac.at/Research/LUC/luc07/External-World-soil-database/HTML/index.html?sb=1`. The terrain, slope, and aspect database provided by HWSD researchers was compiled from a high-resolution digital elevation map constructed by the Shuttle Radar Topography Mission (SRTM). SRTM data is also publicly available as 3 arc-second digital elevation maps (DEM) (approximately 90 meters resolution at the equator), available here: `ftp://e0srp01u.ecs.nasa.gov/srtm/`. The proper data citation is: Fischer, G., F. Nachtergaele, S. Prieler, H.T. van Velthuizen, L. Verelst, D. Wiberg, 2008. *Global Agro-ecological Zones Assessment for Agriculture* (GAEZ 2008). IIASA, Laxenburg, Austria and FAO, Rome, Italy.

in the event that the altitude was not properly recorded during the centerline survey, the HWSD elevation data were used.[60]

Slope and aspect data were also recorded for each administrative boundary polygon and calculated similarly. Slope rasters were computed as the percentage of each 30 arc-second grid that has a 0% to 0.5% gradient (`slope1`), a 0.5% to 2% gradient (`slope2`), a 2% to 5% gradient (`slope3`), a 5% to 10% gradient (`slope4`), a 10% to 15% gradient (`slope5`), a 15% to 30% gradient (`slope6`), a 30% to 45% gradient (`slope7`), and a gradient greater than 45% (`slope8`).

Aspect raster data were recorded as the percentage of each 30 arc-second cell sloping North (315°45°), South (135°225°), East (45°135°), and West (225° − 315°). The raster files are only calculated if the gradient is greater than 2%. Variables equal to the average share of each administrative boundary corresponding to each slope class were constructed using ArcView Software.

Data Citation: Fischer, G., F. Nachtergaele, S. Prieler, H.T. van Velthuizen, L. Verelst, D. Wiberg, 2008. *Global Agro-ecological Zones Assessment for Agriculture* (GAEZ 2008). IIASA, Laxenburg, Austria and FAO, Rome, Italy.

Ruggedness

A 30 arc-second ruggedness raster was computed for Indonesia according to the methodology described by Sappington et al. (2007). The authors propose a Vector Ruggedness Measure (VRM), which captures the distance or dispersion between a vector orthogonal to a topographical plane and the orthogonal vectors in a neighborhood of surrounding elevation planes. To calculate the measure, one first calculates the x, y, and z coordinates of vectors that are orthogonal to each 30-arc second grid of the Earth's surface. These coordinates are computed using a digital elevation model and standard trigonometric techniques.

Given this, a resultant vector is computed by adding a given cell's vector to each of the vectors in the surrounding cells; the neighborhood or window is supplied by the researcher. Finally, the magnitude of this resultant vector is divided by the size of the cell window and subtracted from 1. This results in a dimensionless number that ranges from 0 (least rugged) to 1 (most rugged).[61]

For example: on a (3×3) flat surface, all orthogonal vectors point straight up, and each vector can be represented by $(0, 0, 1)$ in the Cartesian coordinate system. The resultant vector obtained from adding all vectors is equal to $(0, 0, 9)$, and the VRM is equal to $1 - (9/9) = 0$. As the (3×3) surface deviates from a perfect plane, the length of the resultant vector gets smaller, and the VRM increases to 1.

[60] The HWSD elevation raster file records the median elevation (in meters) for each 30 arc-second grid of the Earth's surface. The median is computed across space, from the values of all 3 arc-second cells in the SRTM database.

[61] The authors have generously provided a Python script for computing their Vector Ruggedness Measure (VRM) in ArcView. The script and detailed instructions for installation can be found here: `VectorRuggednessMeasure(VRM)ToolforArcGIShttp://arcscripts.esri.com/details.asp?dbid=15423`.

Table 1.C.1: Indonesia's Road Classification System

Function	Code	Minimum Speed	Minimum Width	Management Authority
Arterial	A	60 km/h	11 m	National
Collector-1	K1	40 km/h	9 m	National
Collector-2	K2	20 km/h	9 m	Provincial
Collector-3	K3	20 km/h	9 m	Provincial
Local	L	20 km/h	7.5 m	Kabupaten & Desa
Neighborhood	Z	15 km/h	6.5 m	Kabupaten & Desa

Source: Departemen Pekerjaan Umum, 2008

Table 1.C.2: Road Function and Managing Authority, Kilometer-Post Observations, 1990-2007

	Road Function			Managing Authority		
	Code	Number of Obs.	Share of Total	Code	Number of Obs.	Share of Total
Java	**A**	52,917	0.17	**N**	93,808	0.30
	K1	40,889	0.13	**P**	132,649	0.42
	K2	121,386	0.39	**K**	15,862	0.05
	K3	10,714	0.03	**S**	72,068	0.23
	L	15,862	0.05			
	Z	72,619	0.23			
	Total	314,387	1.00	**Total**	314,387	1.00
Sumatra	**A**	103,160	0.20	**N**	202,915	0.39
	K1	99,782	0.19	**P**	263,409	0.50
	K2	235,750	0.45	**K**	11,391	0.02
	K3	27,632	0.05	**S**	45,680	0.09
	L	11,391	0.02			
	Z	45,680	0.09			
	Total	523,395	1.00	**Total**	523,395	1.00
Sulawesi	**A**	54,496	0.21	**N**	143,147	0.54
	K1	87,728	0.33	**P**	72,198	0.27
	K2	71,234	0.27	**K**	18,232	0.07
	K3	1,887	0.01	**S**	29,371	0.11
	L	18,232	0.07			
	Z	29,371	0.11			
	Total	262,948	1.00	**Total**	262,948	1.00

Source: IRMS and author's calculations. Data come from kilometer-post observations. Standard deviations in parentheses.

Table 1.C.3: Number of Network Arcs, Distances, and Merge Statistics (by road function)

		Road Function						
		A	**K1**	**K2**	**K3**	**L**	**Z**	**Miss**
Java	**# of Arcs**	1168	889	2618	309	315	37	.
	# of Road IDs	220	129	354	43	72	6	.
	Total Distance	2944.91	1970.65	5832.59	750.39	663.44	92.16	.
	Link-Years Merged	16538	13685	38719	3876	4689	14572	3015
	Link-Years Unmerged	1838	735	1842	45	971	21772	157
	% Merged	0.90	0.95	0.95	0.99	0.83	0.40	0.95
	Arc-Years Merged	20,844	16002	46350	5562	5670	666	.
	Arc-Years Unmerged	180	0	774	0	0	0	.
	% Merged	0.99	1.00	0.98	1.00	1.00	1.00	.
Sumatra	**# of Arcs**	1485	1205	2975	453	277	22	41
	# of Road IDs	207	165	412	87	66	6	13
	Total Distance	4964.69	4469.43	11551.28	1492.97	571.67	56.44	147.56
	Link-Years Merged	24755	20035	49171	6808	2603	8730	1406
	Link-Years Unmerged	718	373	537	52	394	9722	12
	% Merged	0.97	0.98	0.99	0.99	0.87	0.47	0.99
	Arc-Years Merged	26730	21690	51876	7830	4986	396	0
	Arc-Years Unmerged	0	0	1674	324	0	0	738
	% Merged	1.00	1.00	0.97	0.96	1.00	1.00	0.00
Sulawesi	**# of Arcs**	1624	2319	2051	15	391	.	45
	# of Road IDs	113	116	150	4	44	.	1
	Total Distance	2836.96	3805.92	4369.33	28.35	732.96	.	70.34
	Link-Years Merged	24006	24006	34711	30911	551	5670	5674
	Link-Years Unmerged	25	356	410	339	9	118	4755
	% Merged	1.00	0.99	0.99	0.99	0.98	0.98	0.54
	Arc-Years Merged	25794	35694	33660	270	7038	.	0
	Arc-Years Unmerged	3438	6048	3258	0	0	.	810
	% Merged	0.88	0.86	0.91	1.00	1.00		0.00

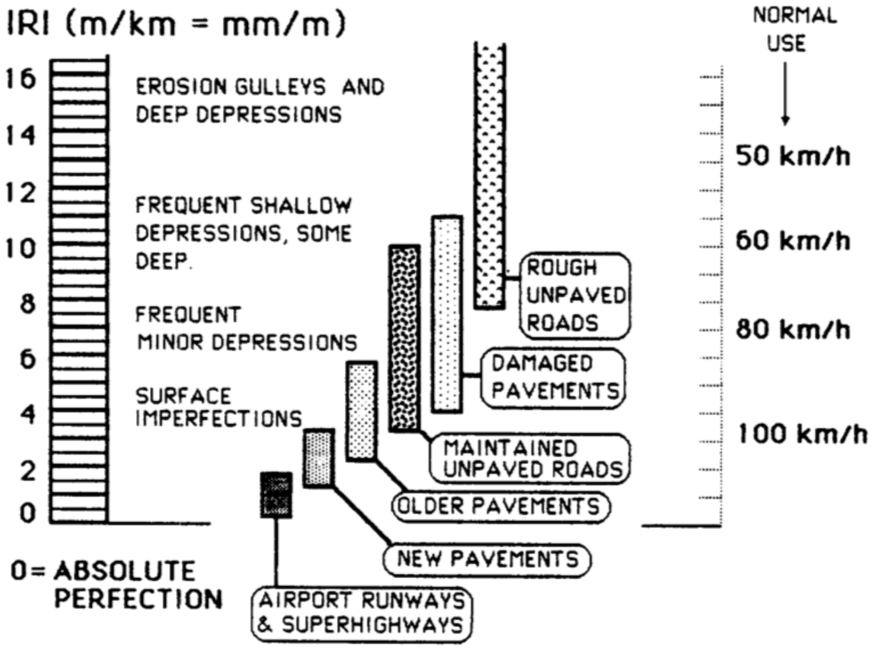
Source: IRMS and author's calculations. Missing function information is attributable to poorly coded shapefiles. Arc-Years could be unmerged potentially because there were no surveys done on that particular link; statistics are computed assuming a balanced panel. Road IDs are defined in the shapefile, while Link IDs are defined from the IRMS data.

Table 1.C.4: Roughness and Ride-Quality Speed Limits

Max Speed	Good	Fair	Mediocre	Poor
120 km/h	$IRI \in [0.00, 1.49]$	$IRI \in [0.00, 1.89]$	$IRI \in [0.00, 2.70]$	$IRI \in [0.00, 3.24]$
100 km/h	$IRI \in [1.49, 1.79]$	$IRI \in [1.89, 2.27]$	$IRI \in [2.70, 3.24]$	$IRI \in [3.24, 4.05]$
80 km/h	$IRI \in [1.79, 2.24]$	$IRI \in [2.27, 2.84]$	$IRI \in [3.24, 4.05]$	$IRI \in [4.05, 4.63]$
70 km/h	$IRI \in [2.24, 2.57]$	$IRI \in [2.84, 3.25]$	$IRI \in [4.05, 4.63]$	$IRI \in [4.63, 5.40]$
60 km/h	$IRI \in [2.57, 2.99]$	$IRI \in [3.25, 3.79]$	$IRI \in [4.63, 5.40]$	$IRI \in [5.40, 6.25]$
50 km/h	$IRI \in [2.99, 3.59]$	$IRI \in [3.79, 4.54]$	$IRI \in [5.40, 6.25]$	$IRI \in [6.25, 8.08]$
40 km/h	$IRI \in [3.59, 4.49]$	$IRI \in [4.54, 5.69]$	$IRI \in [6.25, 8.08]$	$IRI \in [8.08, 10.80]$
30 km/h	$IRI \in [4.49, 5.99]$	$IRI \in [5.69, 7.59]$	$IRI \in [8.08, 10.80]$	$IRI \in [10.80, 16.16]$
20 km/h	$IRI \in [5.99, 8.99]$	$IRI \in [7.59, 11.39]$	$IRI \in [10.80, 16.16]$	$IRI \in [16.16, 32.32]$
10 km/h	$IRI \in [8.99, \infty)$	$IRI \in [11.39, \infty)$	$IRI \in [16.16, \infty)$	$IRI \in [32.32, \infty)$

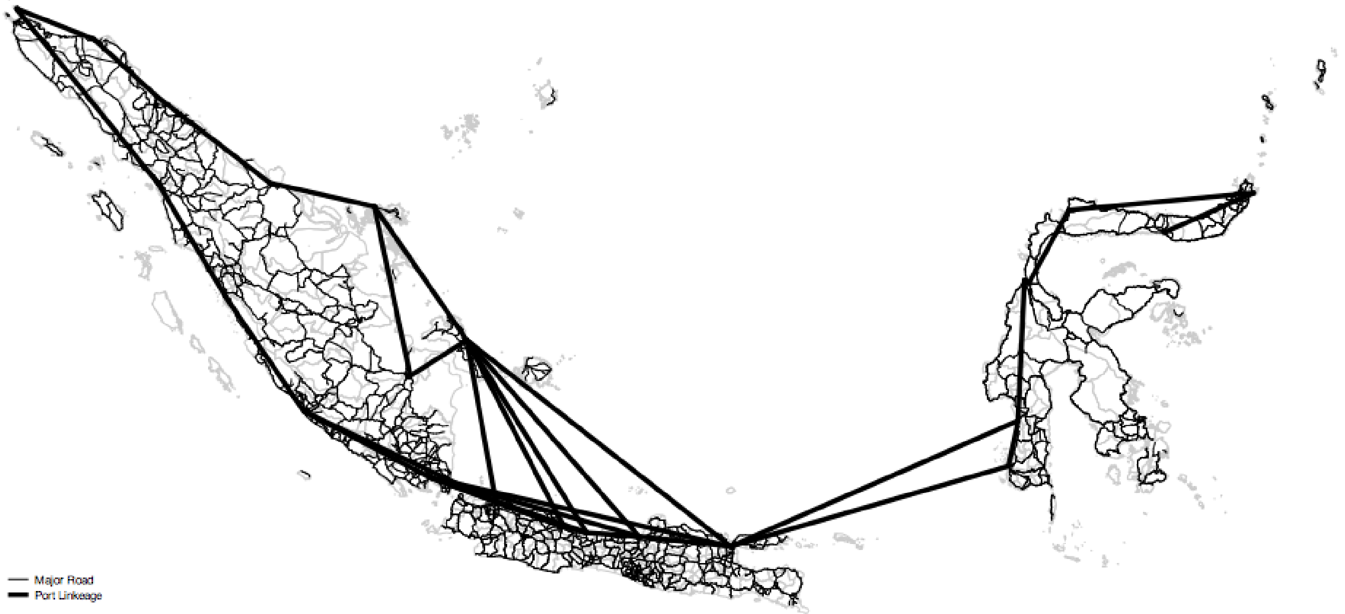
Source: Author's calculations and Yu et al. (2006), Table 2. IRI denotes the international roughness index, measured in m/km. Ride quality levels are subjective and measured on a 5-point scale ("Very Good", "Good", "Fair", "Mediocre", and "Poor").

Figure 1.C.1: Roughness and Surface Type



Source: Sayers et al. (1986).

Figure 1.C.2: Roads and Port Linkeages: Java, Sumatra, and Sulawesi



Source: BPS, DPU and author's calculations. Note that while the links between ports are depicted as straight lines, intersecting large swaths of land, when I compute transport costs using shortest path algorithms, the travel times and distances between these ports are calculated appropriately.

Table 1.C.1: Changes in the Spatial Herfindahl Index, by Industry, 1990-1996

	Description	Mean Δ	Median %Δ	# Decreased / Total
33	Furniture and Wood Products	-0.085	-24.5	9/10
39	Other Manufacturing	-0.047	-25.9	4/5
38	Finished Metal, Machines, and Electronics	-0.039	-24.6	10/13
31	Food and Beverages	-0.030	-16.9	25/29
37	Iron and Steel	-0.030	-21.7	1/1
35	Chemical Products	-0.021	-10.1	12/16
32	Textiles	-0.019	-7.9	11/15
34	Paper Products	-0.017	-0.9	3/5
36	Ceramics, Glass, Cement and Clay Products	0.013	1.6	4/8
99	Total	-0.030	-16.4	79/102

Source: SI and author's calculations. Averages are taken over all 5-digit industries within a given 2-digit industry.

Table 1.C.2: Changes in the Ellison and Glaeser (1997) Index, by Industry, 1990-1996

	Description	**Mean Δ**	**Median %Δ**	**# Decreased / Total**
39	Other Manufacturing	-0.038	-43.9	4/5
33	Furniture and Wood Products	-0.030	-13.0	6/10
37	Iron and Steel	-0.014	-34.6	1/1
38	Finished Metal, Machines, and Electronics	-0.013	-40.7	11/13
36	Ceramics, Glass, Cement and Clay Products	-0.009	-44.7	4/8
34	Paper Products	-0.007	-18.6	5/5
31	Food and Beverages	-0.007	-20.6	17/29
35	Chemical Products	-0.006	-9.6	10/16
32	Textiles	0.005	10.2	6/15
99	Total	-0.010	-20.8	64/102

Source: SI and author's calculations. Averages are taken over all 5-digit industries within a given 2-digit industry.

Table 1.C.3: Trends in Spatial Concentration of Employment, 1985-2005

	Spatial Herfindahl		Ellison and Glaeser (1997) Index	
	(1)	**(2)**	**(3)**	**(4)**
year	-0.002		-0.001	
	(0.000)***		(0.000)***	
1986.year		-0.006		-0.004
		(0.010)		(0.008)
1987.year		-0.008		-0.003
		(0.010)		(0.008)
1988.year		-0.016		-0.004
		(0.010)		(0.007)
1989.year		-0.018		-0.005
		(0.010)*		(0.007)
1990.year		-0.023		-0.009
		(0.009)**		(0.007)
1991.year		-0.031		-0.016
		(0.008)***		(0.006)**
1992.year		-0.041		-0.018
		(0.008)***		(0.006)***
1993.year		-0.045		-0.017
		(0.008)***		(0.006)***
1994.year		-0.045		-0.017
		(0.008)***		(0.006)***
1995.year		-0.049		-0.018
		(0.008)***		(0.006)***
1996.year		-0.053		-0.019
		(0.008)***		(0.006)***
1997.year		-0.052		-0.017
		(0.008)***		(0.006)***
1998.year		-0.053		-0.017
		(0.008)***		(0.006)***
1999.year		-0.055		-0.018
		(0.008)***		(0.006)***
Adj. R^2	0.704	0.714	0.704	0.705
N	2142	2142	2142	2142
5-Digit ISIC FE	Yes	Yes	Yes	Yes

Robust standard errors in parentheses. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level. For columns 2 and 4, the equations included a full set of year indicators, with 1985 being the omitted indicator; estimates for the year effects of 2000-2005 were suppressed for space constraints.

Table 1.C.4: Industrial Concentration Regressions

Panel A: Difference-in-Differences: 1985-2000

	Employment HH	**EG Index**
post	-0.010	0.002
	(0.011)	(0.006)
treatInvXpost	-0.050	-0.037
	(0.019)**	(0.019)*
Adj. R^2	0.605	0.563
N	204	204
5-Digit ISIC FE	Yes	Yes

Panel B: Trend Regressions

	Employment HH	**EG Index**
trend	-0.001	0.000
	(0.000)**	(0.000)*
trendXinvShare	-0.004	-0.003
	(0.001)***	(0.001)***
Adj. R^2	0.786	0.800
N	1632	1632
5-Digit ISIC FE	Yes	Yes

Robust standard errors in parentheses. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.C.5: Reduced Form Regressions (dropping zero observations)

	OLS				**IV**			
Panel A: New Firms	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
MP_{rt}	0.258	0.240	0.149		0.277	0.348	0.336	
	(0.069)***	(0.074)***	(0.064)**		(0.064)***	(0.077)***	(0.080)***	
$\text{MP}_{rt} \times \text{Durable}_s$				0.153				0.148
				(0.054)***				(0.045)***
Adj. R^2	0.282	0.282	0.291	0.140	-0.035	-0.011	-0.016	-0.282
N	5952	5952	5952	5952	5946	5946	5946	5246
Kabupaten FE	Yes	Yes	Yes	.	Yes	Yes	Yes	.
Year FE	Yes	.	.	.	Yes	.	.	.
Sector FE	Yes	.	.	.	Yes	.	.	.
Province Trends	.	.	Yes	.	.	.	Yes	.
Sector-Year FE	.	Yes	Yes	Yes	.	Yes	Yes	Yes
Kabu-Year FE	.	.	.	Yes	.	.	.	Yes

	OLS				**IV**			
Panel B: Employment	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
MP_{rt}	0.590	0.464	0.312		0.640	0.724	0.706	
	(0.205)***	(0.217)**	(0.189)		(0.183)***	(0.154)***	(0.159)***	
$\text{MP}_{rt} \times \text{Durable}_s$				0.284				0.271
				(0.109)***				(0.091)***
Adj. R^2	0.266	0.267	0.273	0.165	-0.035	-0.014	-0.018	-0.285
N	5952	5952	5952	5952	5946	5946	5946	5246
Kabupaten FE	Yes	Yes	Yes	.	Yes	Yes	Yes	.
Year FE	Yes	.	.	.	Yes	.	.	.
Sector FE	Yes	.	.	.	Yes	.	.	.
Province Trends	.	.	Yes	.	.	.	Yes	.
Sector-Year FE	.	Yes	Yes	Yes	.	Yes	Yes	Yes
Kabu-Year FE	.	.	.	Yes	.	.	.	Yes

Unit of observation is a region-industry-year. Robust standard errors in parentheses, clustered at the kabupaten level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.C.6: Reduced Form Regressions (5-digit, using MP_{1985})

Panel A: New Firms	**(1)**	**(2)**	**(3)**
log_MP	0.025 (0.007)***	0.025 (0.007)***	
log_invXMP_DM			0.028 (0.007)***
Adj. R^2	0.123	0.123	0.094
N	301920	301920	301920
Kabupaten FE	Yes	Yes	.
Year FE	Yes	.	.
Sector FE	Yes	.	.
Province Trends	Yes	Yes	.
Sector-Year FE	.	Yes	Yes
Kabu-Year FE	.	.	Yes

Panel B: Employment	**(1)**	**(2)**	**(3)**
log_MP	0.094 (0.023)***	0.094 (0.023)***	
log_invXMP_DM			0.073 (0.017)***
Adj. R^2	0.128	0.129	0.106
N	301920	301920	301920
Kabupaten FE	Yes	Yes	.
Year FE	Yes	.	.
Sector FE	Yes	.	.
Province Trends	Yes	Yes	.
Sector-Year FE	.	Yes	Yes
Kabu-Year FE	.	.	Yes

Unit of observation is a region-industry-year. Robust standard errors in parentheses, clustered at the kabupaten level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.C.7: Reduced Form Regressions (5-digit, dropping zero observations)

Panel A: New Firms	**(1)**	**(2)**	**(3)**
log_MP	0.123 (0.049)**	0.141 (0.058)**	
log_invXMP_DM			0.097 (0.032)***
Adj. R^2	0.229	0.186	0.059
N	8391	8391	8391
Kabupaten FE	Yes	Yes	.
Year FE	Yes	.	.
Sector FE	Yes	.	.
Province Trends	Yes	Yes	.
Sector-Year FE	.	Yes	Yes
Kabu-Year FE	.	.	Yes

Panel B: Employment	**(1)**	**(2)**	**(3)**
log_MP	0.279 (0.163)*	0.136 (0.197)	
log_invXMP_DM			0.215 (0.067)***
Adj. R^2	0.256	0.250	0.074
N	8391	8391	8391
Kabupaten FE	Yes	Yes	.
Year FE	Yes	.	.
Sector FE	Yes	.	.
Province Trends	Yes	Yes	.
Sector-Year FE	.	Yes	Yes
Kabu-Year FE	.	.	Yes

Unit of observation is a region-industry-year. Robust standard errors in parentheses, clustered at the kabupaten level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 1.C.8: Choice Characteristics Summary Statistics

	Mean	**SD**	**N**
Endogenous Cost Shifters			
Wages	6.97	(1.21)	2093
Commercial Land Values	10.91	(1.22)	2093
Log Market Potential	17.36	(0.72)	2093
PLN Share	0.89	(0.29)	2093
Indirect Tax Rate	0.01	(0.01)	2093
Physical and Agroclimatic Chars			
Area	3.05	(4.54)	2093
Ruggedness	7.02	(7.83)	2093
Elevation	0.28	(0.24)	2093
Percentage of Cultivated Land	0.38	(0.16)	2093
Percentage of Forested Land	0.21	(0.15)	2093
Percentage of Grassland	0.14	(0.08)	2093
Distance to Jakarta	6.75	(5.07)	2093
Distance to Major Cities	0.79	(0.49)	2093
Distance to Major Ports	0.87	(0.54)	2093
Distance to Malaysia	7.78	(2.66)	2093
Distance to Singapore	11.17	(4.72)	2093
Year			
1990			155
1991			161
1992			150
1993			144
1994			149
1995			148
1996			141
1997			126
1998			121
1999			118
2000			105
2001			129
2002			123
2003			119
2004			115

Source: SI and author's calculations.

Table 1.C.9: New Firm Summary Statistics

	Mean	**SD**	**N**
Industrial Sector (2-Digit)			
31. Food and Beverage Processing	0.22	(0.41)	17684
32. Textiles and Clothing	0.24	(0.43)	17684
33. Wood Products	0.18	(0.38)	17684
34. Paper Products	0.04	(0.19)	17684
35. Chemical and Oil Products	0.10	(0.30)	17684
36. Ceramics, Glass, and Clay	0.08	(0.26)	17684
37. Iron and Steel Products	0.01	(0.09)	17684
38. Finished Metal Products	0.12	(0.32)	17684
39. Other Manufacturing	0.03	(0.17)	17684
Year			
1990	0.12	(0.33)	17684
1991	0.09	(0.29)	17684
1992	0.09	(0.29)	17684
1993	0.07	(0.26)	17684
1994	0.08	(0.28)	17684
1995	0.09	(0.28)	17684
1996	0.07	(0.26)	17684
1997	0.05	(0.22)	17684
1998	0.04	(0.19)	17684
1999	0.04	(0.20)	17684
2000	0.04	(0.18)	17684
2001	0.06	(0.23)	17684
2002	0.04	(0.21)	17684
2003	0.04	(0.20)	17684
2004	0.05	(0.22)	17684
2005	0.03	(0.16)	17684
Province			
11. Aceh	0.01	(0.07)	17684
12. North Sumatra	0.04	(0.19)	17684
13. West Sumatra	0.01	(0.09)	17684
14. Riau	0.03	(0.18)	17684
15. Jambi	0.00	(0.07)	17684
16. South Sumatra	0.01	(0.11)	17684
17. Bengkulu	0.00	(0.04)	17684
18. Lampung	0.01	(0.09)	17684
31. DKI Jakarta	0.11	(0.31)	17684
32. West Java	0.32	(0.47)	17684
33. Central Java	0.17	(0.37)	17684
34. DI Yogyakarta	0.02	(0.15)	17684
35. East Java	0.23	(0.42)	17684
71. North Sulawesi	0.01	(0.11)	17684
72. Central Sulawesi	0.00	(0.07)	17684
73. South Sulawesi	0.02	(0.14)	17684
74. Southeast Sulawesi	0.01	(0.09)	17684

Source: SI and author's calculations.

Table 1.C.10: Constant Coefficient Logit Results (using MP_{1985})

	OLS (1990)		Fixed Effects LS (1990-2005)					Panel IV (1990-2005)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
wage_rate	-0.005 (0.177)	-0.250 (0.166)	-0.152 (0.051)***	-0.145 (0.050)***	-0.146 (0.050)***	-0.139 (0.050)***	-0.145 (0.050)***	-0.185 (0.087)**	-0.185 (0.087)**
land_value	0.239 (0.101)**	0.161 (0.099)	-0.077 (0.030)***	-0.076 (0.030)**	-0.075 (0.030)**	-0.072 (0.028)**	-0.073 (0.028)**	-0.046 (0.048)	-0.041 (0.048)
MP	1.132 (0.152)***	1.085 (0.223)***	0.977 (0.406)**		0.889 (0.406)**	0.840 (0.406)**	0.916 (0.408)**	2.964 (0.949)***	2.821 (0.945)***
pavedDensity				0.070 (0.075)					
sharePLN					0.228 (0.128)*	0.205 (0.128)			
indTaxRate						-6.301 (2.469)**	-6.633 (2.502)***		-7.204 (2.543)***
Adj. R^2	0.316	0.405	0.089	0.086	0.091	0.094	0.093	0.076	0.082
N	180	180	2093	2093	2093	2093	2093	1937	1937
F Statistic	26.610	16.225	13.066	13.167	12.808	14.115	14.450	13.947	15.156
Fixed Controls	.	Yes	.	.	.	.	.	.	.
Kabupaten FE	.	.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rural-Urban Year FE			Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dynamic Panel IVs	.	.	.	.	.	.	.	Yes	Yes
WTP for MP with wages			6.41**	0.48	6.09*	6.04*	6.33*	16.06	15.22
WTP for MP with rents			12.73*	0.92	11.83	11.70	12.54*	64.92	68.05
WTP for MP with taxes						0.13		0.14	0.39*

Robust standard errors in parentheses, clustered at the kabupaten level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level. In all columns except column 5, the adjusted R-squared reported is the “within R-squared” obtained by estimating the equation in mean-deviation form.

Figure 1.C.1: Average Travel Times: Java

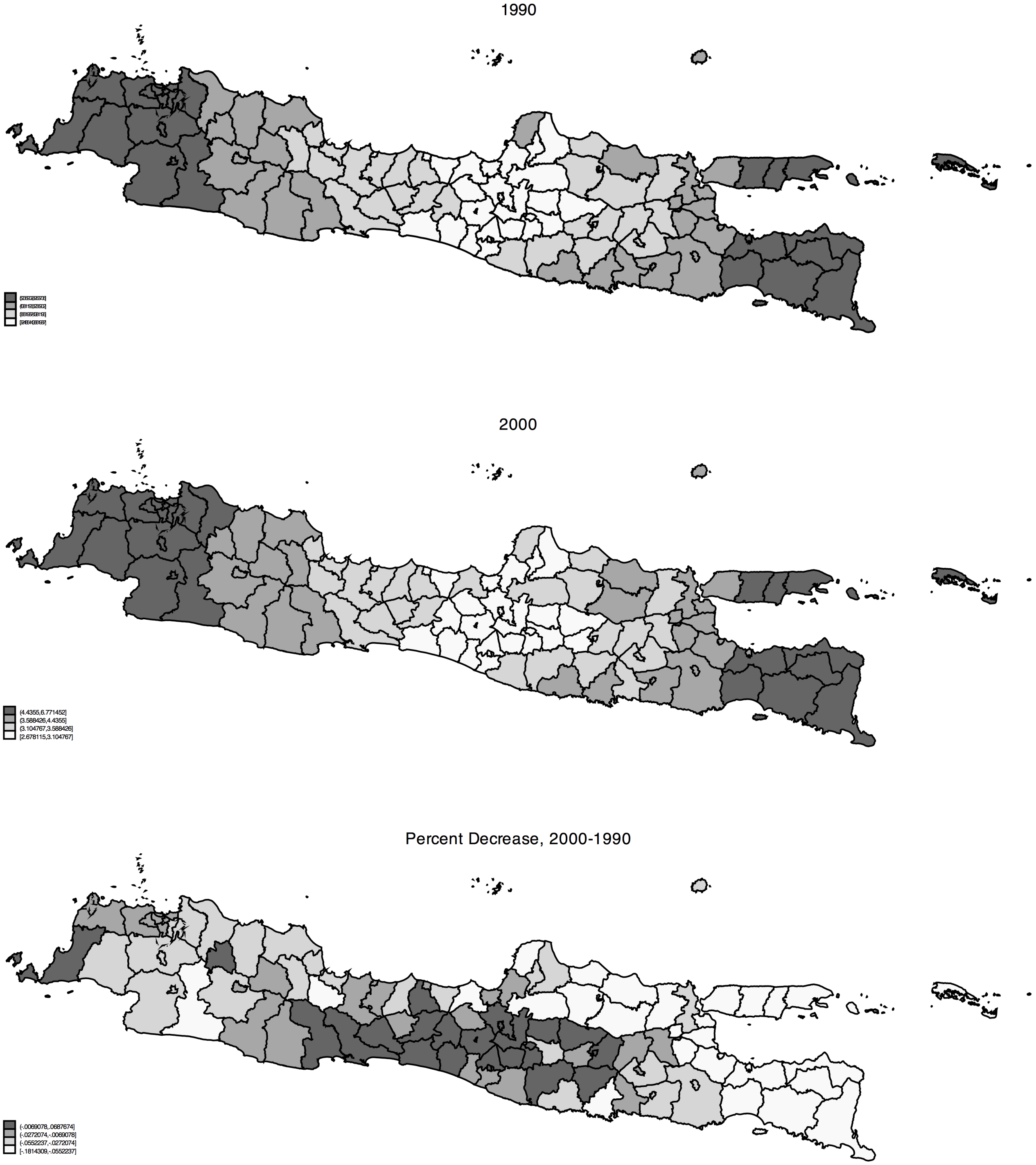


Figure 1.C.2: Average Travel Times: Sumatera

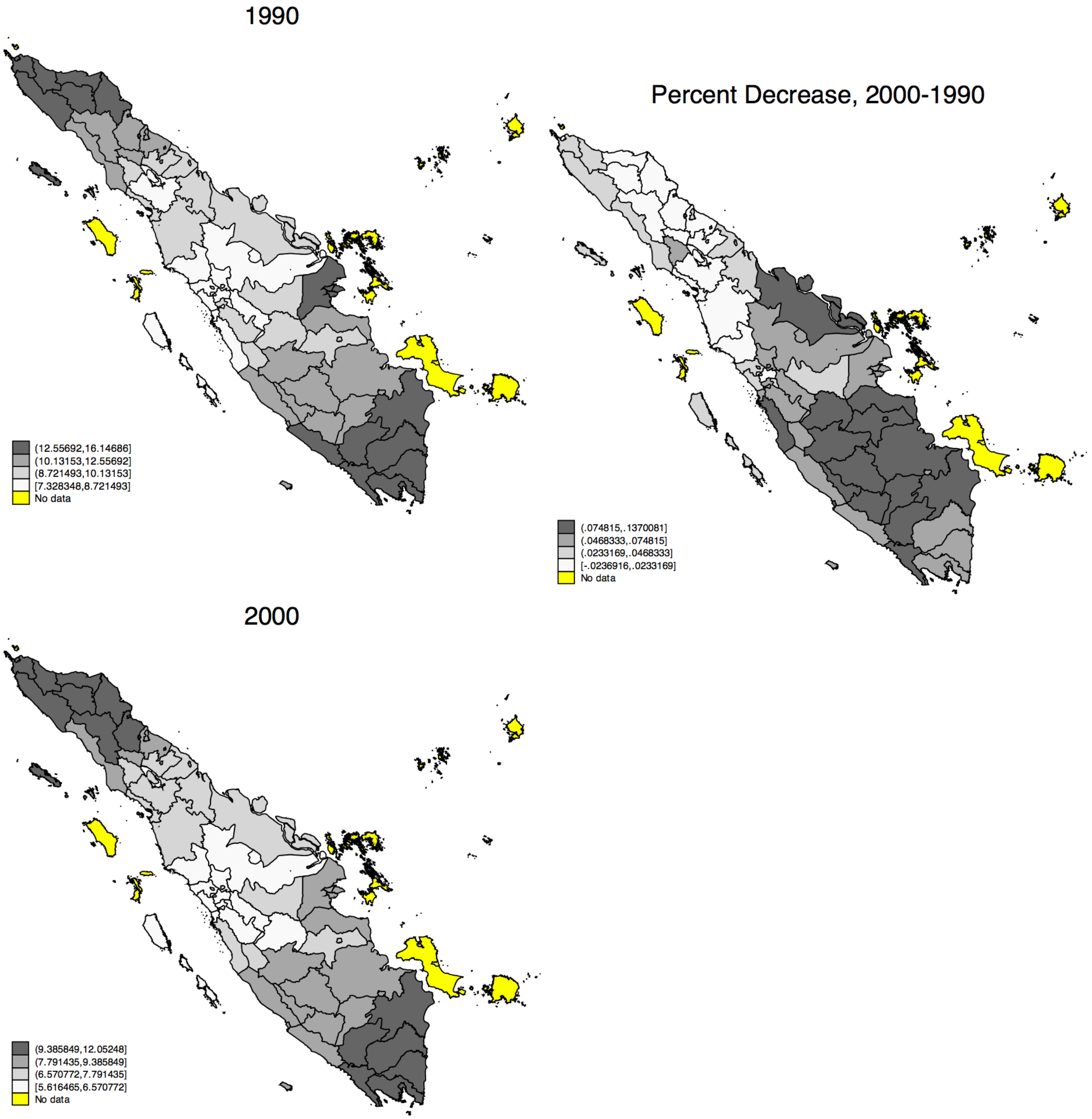


Figure 1.C.3: Average Travel Times: Sulawesi

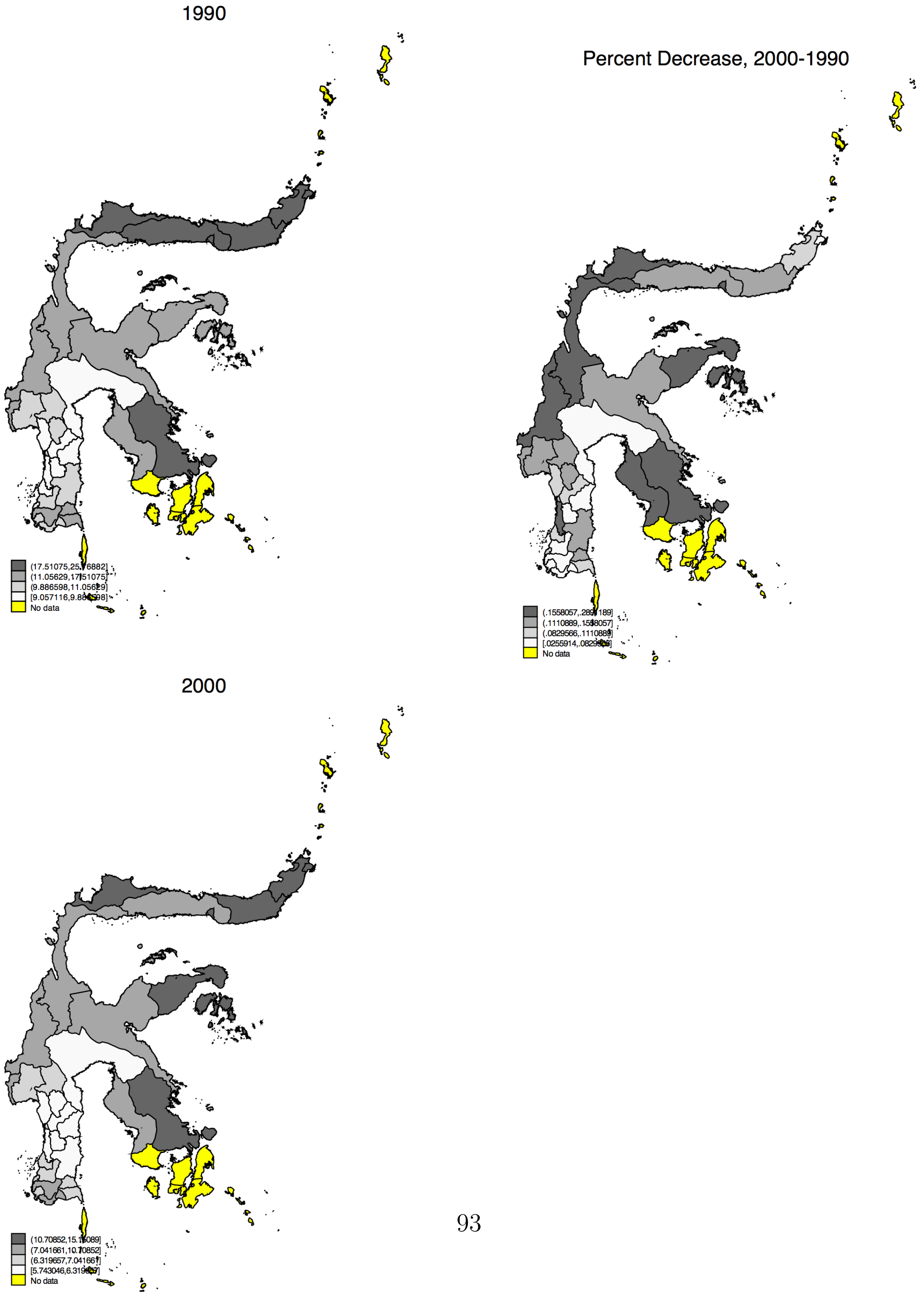
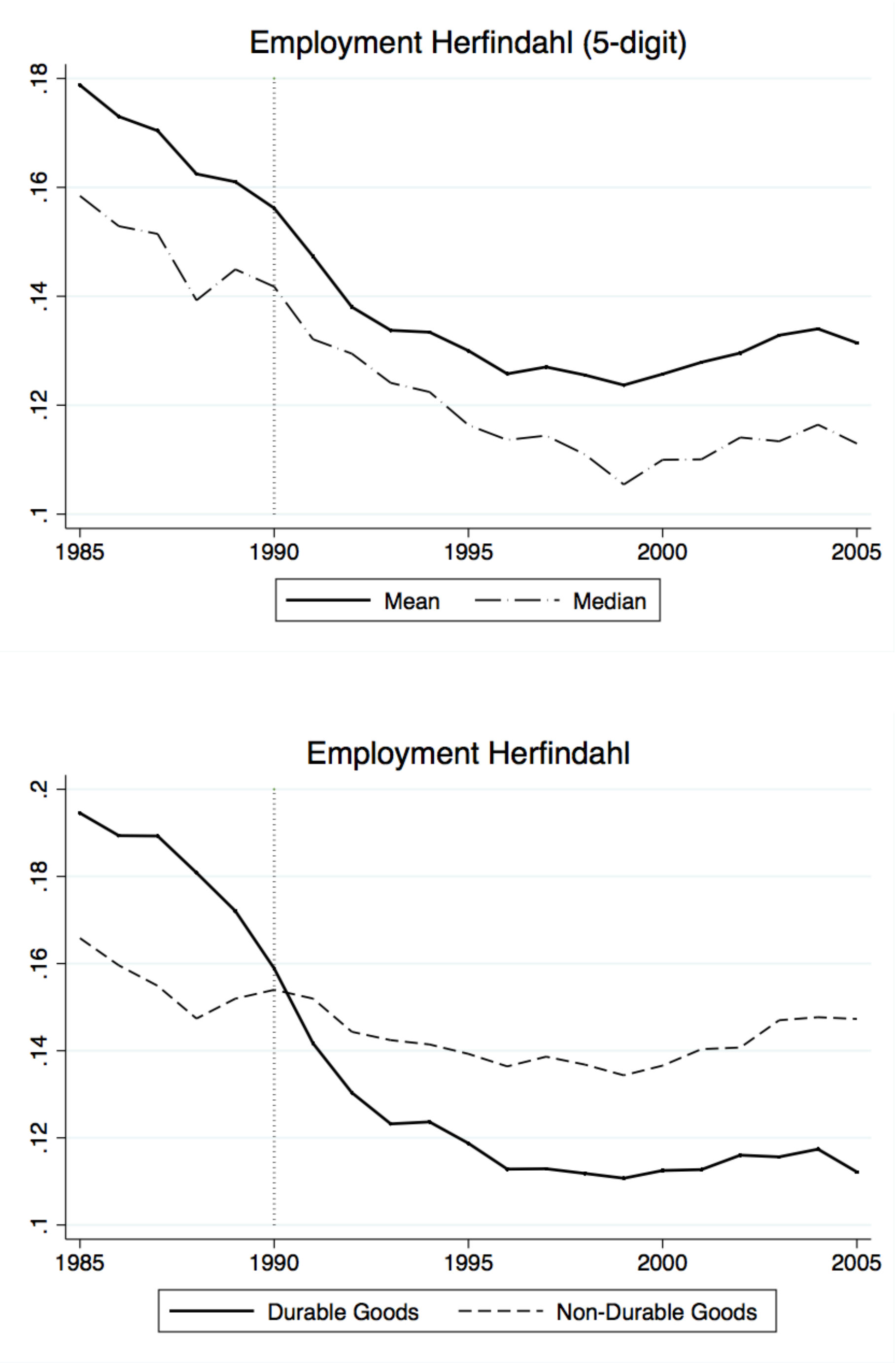
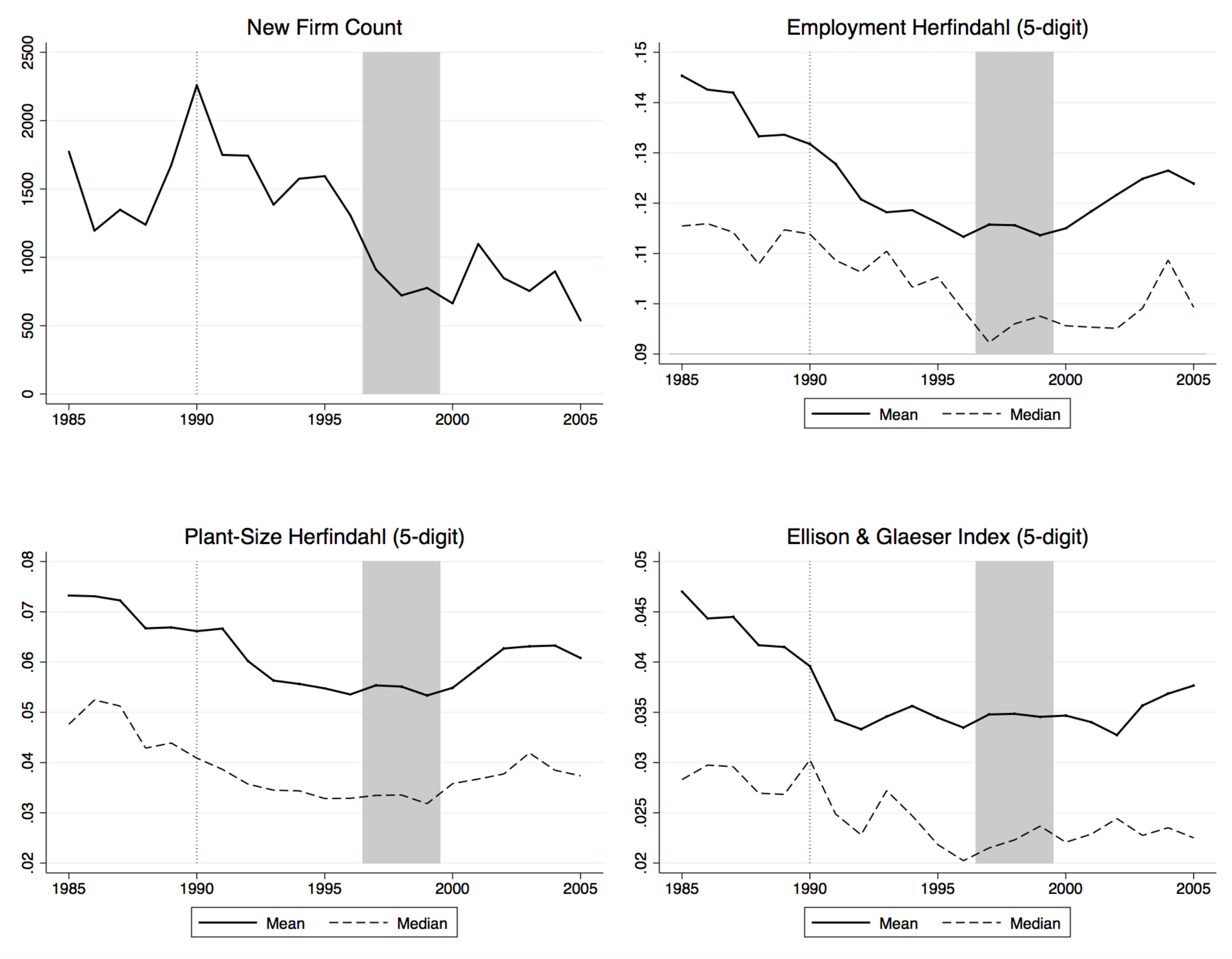


Figure 1.C.4: Trends in the Ellison and Glaeser (1997) Index



Source: SI data and author's calculations. Lines depict annual means or medians of different indices of industrial concentration across 5-digit industries, as well as means by industry type. Grey bar denotes crisis period (1997-1999). Regressions of industrial concentration measures across industry years on a set of year dummies (or a trend) indicate that the observed reductions are statistically significant, beginning in the Ellison and Glaeser (1997) index (see Table 1.C.3). From difference-in-difference regressions (see Appendix Table 1.C.4), the change in the Spatial Herfindahl for durable goods industries, relative to non-durable goods, was -0.05 (s.e. 0.019). Similar magnitudes for difference-in-difference estimates can be found for the Ellison and Glaeser Index, though the estimates are noisier.

Figure 1.C.5: Evolution of New Firm Counts and Industrial Concentration (Using Kabupaten Definitions)



Source: SI data and author's calculations. Lines depict annual means or medians of different indices of industrial concentration across 5-digit industries. Grey bar denotes crisis period (1997-1999).

Figure 1.C.6: Evolution of Industrial Concentration Measures (Inventory Shares)

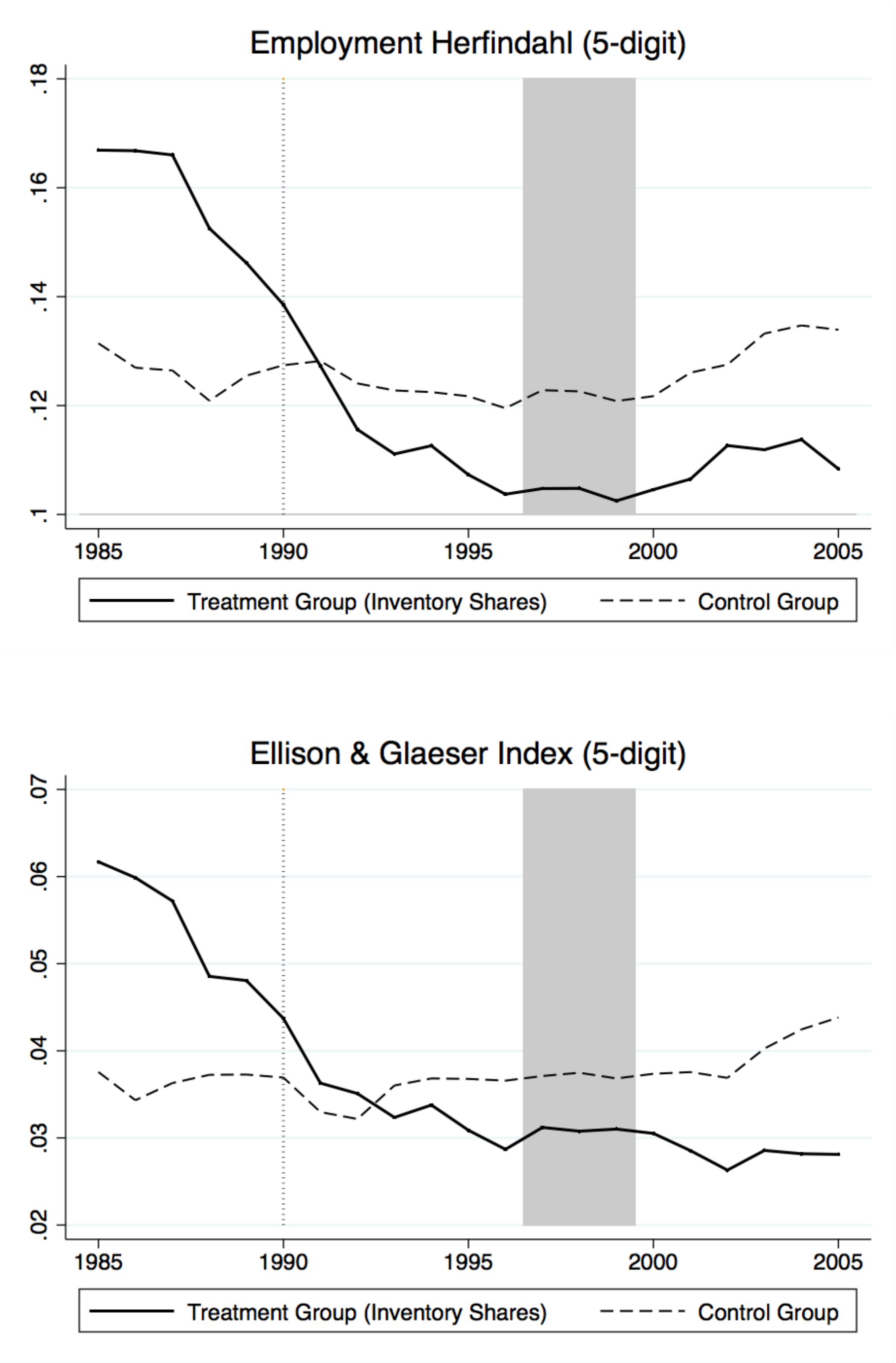
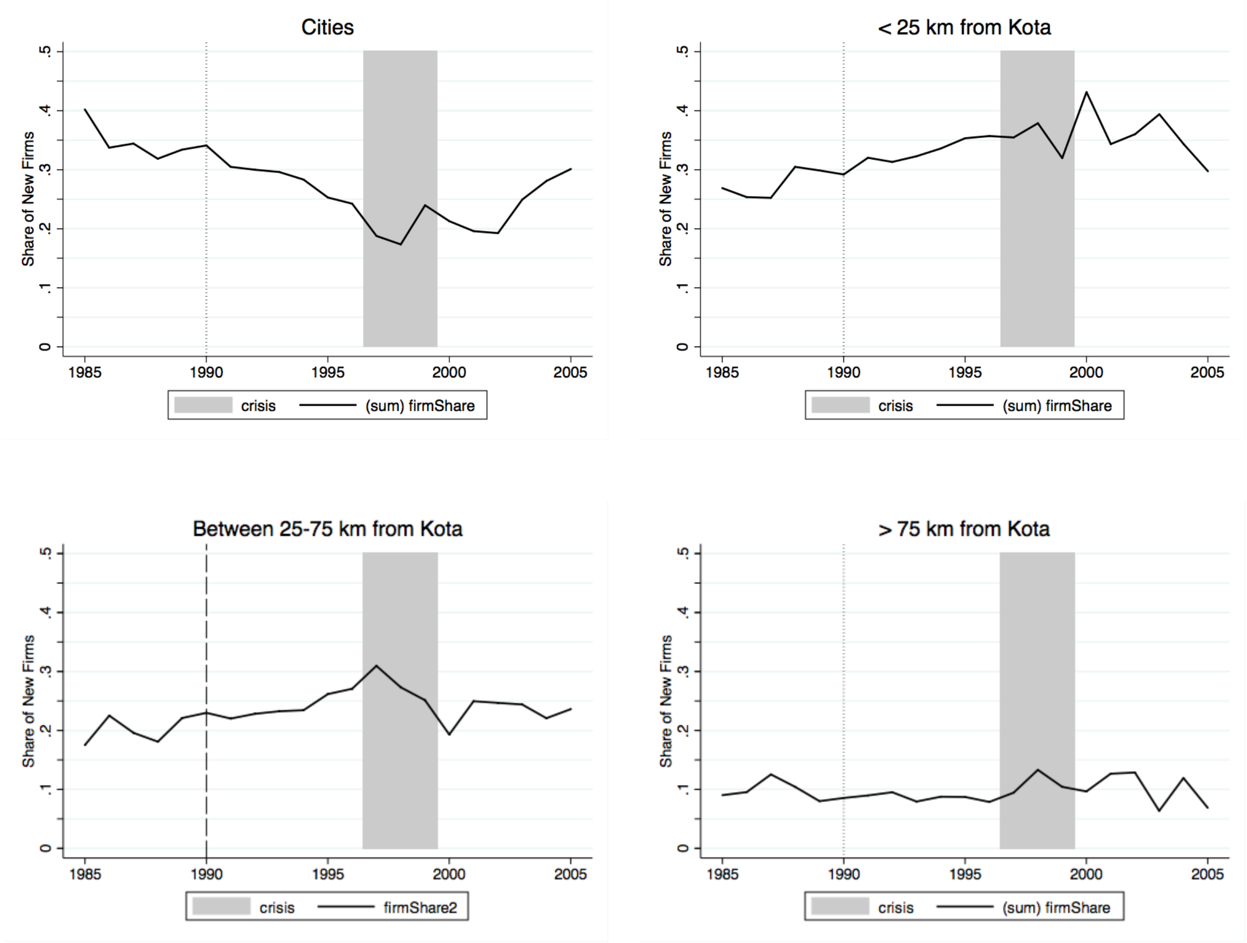


Figure 1.C.7: Share of New Firms Locating in Different Types of Kabupatens (Distance)



Source: SI data and author's calculations. Lines depict shares of new firms locating in different types of kabupatens within Java, Sumatra, and Sulawesi. A total of 51 out of 218 kabupatens were classified as Kota / Kotamadya in 1990. Distance categories are assigned using distance "as the crow flies" between kabupaten and kota centroids. There were 31 kabupatens within 25 km of a 1990 city, 73 kabupatens between 25 and 75 km to a 1990 city, and 63 kabupatens that were greater than 75 km from a 1990 city. In classifying, some kabupatens fit into multiple categories, and when this occurred, the kabupaten was assigned to the closest group possible.

Chapter 2

The Benefits of Transport Infrastructure: Evidence from Banana Company Railroads in Honduras

with Bryan S. Graham

Abstract

Without credible estimates of the welfare effects of infrastructure improvements, it is impossible for policymakers to know whether the benefits of these investments outweigh their substantial costs. A major challenge with trying to understanding the welfare consequences of improved transport infrastructure is an identification problem: transport improvements are never randomly assigned. We overcome this identification program by exploiting variation from a novel natural experiment: Honduras' infestation with Panama disease. The Honduran railroad network was constructed by fruit companies to ship bananas from plantations to port cities, but because of an unpredictable outbreak of Panama disease, major plantations and their associated railway infrastructure were abandoned. We argue that outbreaks of Panama disease were extremely difficult to predict, and because of this, conditional on the railway network that existed in the 1930s and a host of observable characteristics, the areas where railway lines were abandoned were randomly assigned. We use our identification strategy to uncover the implicit prices of access to infrastructure paid by consumers and the implicit production costs paid by firms.

2.1 Introduction

In many low income countries, inadequate transport infrastructure is a major constraint to development. High internal transport costs raise final goods prices by restricting firms' access to consumers and raising input prices. These higher final goods prices make it more difficult for consumers to save and make important investments to human capital. In the extreme, large transport costs could completely segment regional markets, artificially giving power to inefficient, unproductive firms. While the benefits of transport infrastructure improvements are potentially large and significant, the effect sizes remain poorly understood.

Without credible estimates of the welfare effects of infrastructure improvements, it is impossible for policymakers to know whether or not the benefits of these investments outweigh their substantial costs. It is also difficult to assess whether investing in infrastructure has larger returns than alternative investments, such as spending on health or education. Although the World Bank allocated nearly $32 billion in support of transport infrastructure from 1995-2005, more than its combined spending on health, education, and social services (World Bank, 2007), a lack of understanding of the welfare effects of these investments prevents us from being able to evaluate the efficiency of this development portfolio.

A major challenge with trying to understanding the welfare consequences of improved transport infrastructure is an identification problem: transport improvements are never randomly assigned. Estimates of their effects may be confounded with the fact that areas benefitting from improved transport infrastructure were selected by policymakers, creating targeting bias. This targeting bias could be large and substantial, and it is difficult to sign without knowledge of the targeting rules.

In this paper, we overcome this identification program by exploiting variation from a novel natural experiment: Honduras' infestation with Panama disease. The Honduran railroad network was constructed by fruit companies in the early 20th century to ship bananas from plantations on the North Coast to port cities. By the 1930s, banana plantations owned by these companies were all growing one single variety of banana for commercial production (the Gros Michel cultivar). Unfortunately, this cultivar is very susceptible to Panama disease, a root fungus that stunts the growth of the banana plant, blackening its fruit and making it inedible. Because of the ravages of this disease on farms throughout Honduras, many banana plantations on the North Coast were abandoned by the fruit companies.

As plantations were abandoned, the fruit companies often stripped the railway lines that connected them to the ports, literally removing the tracks in many instances. The configuration of the transportation network that resulted from these abandoned lines strongly persists to the present day. We argue that outbreaks of Panama disease were extremely difficult to predict, and because of this, conditional on the railway network that existed in the 1930s and a host of observable characteristics, the abandoned railway lines were randomly assigned. Given the identification problems associated with non-random assignment of major infrastructure improvements (Gramlich, 1994), this unique natural experiment represents a major breakthrough in our understanding of the causal effects of transport infrastructure.

We use our identification strategy to uncover the implicit prices of access to infrastructure paid by consumers and the implicit production costs paid by firms. To do this, we develop a simple long-run spatial equilibrium model, in which workers and firms are perfectly mobile across locations. Differences in local amenities across locations, such as access to transport infrastructure, will be reflected in differences in equilibrium wages and rents (Roback, 1982). These differences yield the implicit prices of those amenities, which tell us how much firms and workers are willing to pay for access to them. This approach for amenity valuation has been used extensively to estimate the quality of life in urban areas (e.g Blomquist et al., 1988; Gyourko et al., 1999), and it has also been proposed as a way to estimate the value of infrastructure access in developing countries (Jacoby, 2000). However, because this approach involves estimating hedonic regressions, it has been widely criticized for delivering imprecise estimates, often with unexpected signs (Chay and Greenstone, 2005). We feel that our identification strategy, combined with detailed census data, will allow us to make progress on these difficult problems.

Our major finding is that abandonment by the fruit companies in the 1930s and 1940s has had a large and lasting impact on the spatial distribution of economic activity in Honduras. Wages, rents, and population density are all substantially lower in the abandoned areas relative to non-abandoned areas, and this holds even after conditioning on a multitude of variables likely to influence abandonment. However, because of imprecise data on rents, the implied implicit prices of infrastructure access are noisy, and the implicit cost savings from access to infrastructure is small. Further work with this research design and better data will allow us to provide more precise estimates of the relationship between land rents and transport infrastructure amenities.

In section 2.2, we will discuss historical background on the banana trade in Honduras and explain how the unpredictable outbreak of Panama disease forced fruit companies' to abandon many plantations and remove railway lines that connected those plantations to major ports. In section 2.3, we present a simple spatial equilibrium model used to guide our empirical analysis. Section 2.4 discusses the data we use, and section 2.5 presents the results. In Section 2.6, we conclude with a discussion of our direction for future research.

2.2 Bananas, Railroads, and Panama Disease

This section provides the historical background for our natural experiment, drawing heavily on environmental and economic history discussed by Soluri (2000, 2005). We first summarize the history of export banana production in Honduras, taking care to highlight the fruit companies' involvement in constructing the railroads. From 1890-1950, the major fruit companies were all exporting a single variety of banana, the Gros Michel cultivar, and we explain why this monocultural production left many banana plantations highly susceptible to Panama disease. Presenting evidence from contemporary authorities, we explain how the incidence of Panama disease was unpredictable, and we explain that when the banana companies responded to panama disease by abandoning plantations, they often removed railway lines.

2.2.1 Fruit Companies and the Banana Trade

Honduras's experience with exporting bananas began in the 1870's, when the first American ships arrived on the north coast to purchase bananas and coconuts to be resold in the United States. Schooners and increasingly steamships set sail from the port of New Orleans, and upon arrival in Honduras, captains would negotiate purchase agreements with local land owners, who cultivated bananas on a small scale in addition to other crops and livestock. At this time, bananas were not a staple of the American diet, and the fruit was marketed as a luxury item.

Although local landowners often grew a variety of crops, including corn, beans, yuca, or plantains, by the early 1880s, many farmers switched to exclusive banana production. However, in this period, the infrastructure necessary to transport bananas from farms to ports was not well developed. Because of this, banana production was typically restricted to areas near ports or along key rivers, where growers could load boats with bananas and send them upstream to be sold. Roads were not well maintained and were often impossible to traverse during the worst months of the rainy season. Because of badly needed infrastructure improvements, the Honduran government began to consult with railroad engineers, attracting them to construct railway lines in precarious environmental conditions by offering them ownership of vast quantities of land.

In fact, many of the major fruit companies that dominate today's markets started out as railway construction enterprises. The leading example is the United Fruit Company (now Chiquita Brands International). Minor C. Keith, one of the company's founders, had originally worked as a railroad engineer for the government of Costa Rica. Keith was hired to establish a railway connection between San Jose and the Carribean sea. As part of his agreement for building the railroad, Keith was given ownership of huge quantities of land, and almost as an afterthought, he converted a number of these lands into banana farms.

In 1889, Keith merged his organization with Andrew W. Preston's Boston Fruit Company, which was primarily a fruit shipping company. The United Fruit Company, created through the merger, was completely vertically integrated, controlling both the means of production and the transport networks necessary for moving fruit from farms in Central America to consumers in the United States. United Fruit would eventually grow and acquire the holdings of several other companies that owned land throughout the north coast of Honduras.

A major competitor to United Fruit, the Standard Fruit Company (now part of the Dole Food Company), was founded in 1899 by three brothers, Joseph, Luca and Felix Vaccaro, who imported bananas from La Ceiba to the port of New Orleans. Between 1910 and 1950, the Vaccaro brothers oversaw the construction of 155 kilometers of railway lines in the northern Honduran department of Atlantida, from La Ceiba to Yoro. Another competitor in the banana trade was the Cuyamel Fruit Company, founded by William Streich, another railroad engineer. In 1902, Streich was awarded a contract to build and operate a railroad in Omoa, a municipality in northwest Honduras, and the terms of this contract included concessions for owning and operating lands alongside the railroad.

2.2.2 The Gros Michel and Panama Disease

By far, the most prominent cultivar of banana grown in Honduras from 1870 to the late 1940's was the Gros Michel (*Musa acuminata*).[1] The plant is large and bears heavy, symmetrical bunches of fruit. Crucially for the banana companies engaged in maritime trade, Gros Michel fruit ripens slowly, and it has a tough peel that protected the fruit while completing a difficult journey from farm to market. The Gros Michel plant also produced relatively more fruit than other cultivars, and consumers enjoyed its flavor, aroma, and texture.

Unfortunately, the Gros Michel is highly susceptible to certain strains of the root fungus that cause Panama Disease (*Fusarium oxysporum*).[2] Initial symptoms include a sudden, unmistakeable yellowing of the plant's lower leaves. As they turn yellow, the leaves rapidly wilt, and within a day or two begin to buckle and hang from their stem, causing the plant to rot and die. Photographs of infected banana plants appear in Figure 2.1. Infected adult plants die slowly and will continue to produce fruit, but because of wilted leaves, the fruit ripens early from overexposure to sunlight. Young plants that are infected display stunted growth and die quickly, producing no fruit.

When Panana Disease first appeared is not certain, but in the 1890s growers in the Bocas del Toro region of Panama spotted the disease and gave it its name (Marquardt, 2001). Early outbreaks are documented throughout the Caribbean Islands, in Surinam (1906), Cuba (1908), Trinidad (1909), Puerto Rico (1910), and Jamaica (1911). Infection on the North Coast of Honduras probably first occurred between 1910-1915, but it wasn't until 1916 that soil surveyors from the United Fruit Company reported outbreaks on farms in Tela. In 1922, the disease was reported in Colon by employees of the Trujillo Railroad Company.

Causes of the disease were not well understood until 1910, when a U.S. researcher in Cuba isolated strains of the soil fungus that was responsible. Even so, there was no academic consensus on the cause until 1919, when researchers in Panama succeeded in reproducing symptoms of the disease from exposure to the soil fungus in laboratory controlled conditions. Despite this academic work, it took decades for farmers to become more careful about spreading the disease from affected soils to unaffected areas. Human activity is almost certainly responsible for the disease's rapid spread throughout Central America, and especially throughout Honduras.[3]

The fruit companies initially employed a variety of techniques to combat the disease, including adding lime to distressed soils, burning infected areas, mulching, and adding manure,

[1] This section draws heavily from discussions in Wardlaw (1972), Marquardt (2001), and Soluri (2005).

[2] Panama Disease is also called Vascular Wilt Disease or Banana Wilt (Wardlaw, 1972).

[3] Discussing agricultural practices today, Koeppel (2008) remarks that farmers who work multiple fields often do not take care to clean their shoes and risk infecting new soils as a result.

but none were successful. Extensive surveys of plantations, beginning with one conducted in 1916, determined that plantations with better drainage and stronger soil composition seemed to be the most resistant to the disease, but these correlations were weak, at best. Importantly, the impact of Panama Disease was very unpredictable; as Marquardt (2001, p. 62) describes, "some infected farms, with apparently good, well-drained soils, succumbed to the wilting syndrome almost immediately, while others, in no clear way superior, continued to produce good harvests for many years despite the presence of the fungus."

2.2.3 Abandonment of Plantations

Because the fungus that causes Panama Disease lives in the soil, it was difficult to eradicate, and banana growers determined that the most cost effective response was to abandon the land when plants showed symptoms of the disease and move to new production sites.[4] For instance, Standard Fruit initially had concessions to work land and build railroads southwest of La Ceiba. However, because of Panama Disease outbreaks, by 1919 the company had renegotiated the terms of its agreement with local governments so that it could move its plantations east of La Ceiba, closer to the department of Colon. By the late 1920s, the company had essentially abandoned its farms west of La Ceiba, and in 1935, it turned over some 25,000 acres of abandoned land to the Honduran government.

The Tela and Trujillo Railroad Companies responded similarly to the Panama Disease epidemic. Tela Railroad Company's initial land concessions were for production in the Lean valley, but by 1932 the company had ceased production in that area because of outbreaks of the disease. In the 1920s, the Trujillo Railroad Company abandoned nearly 10,000 hectares in Colon, and in 1937, only ten years after starting to produce in the Black River valley, the company had completely abandoned the region and returned 17,000 hectares of land to the state. As the fruit companies abandoned lands they had acquired, they renegotiated the terms of their concessions with local governments and moved into untouched areas.[5]

Case Study: Omoa-Cuyamel and United Fruit

The Omoa-Cuyamel region represents a good illustration of the abandonment of banana plantations, the removal of infrastructure, and its impact on the local economy (Soluri, 2005). Initially, the United Fruit Company had many plantations in the region, but by 1931, rumors had circulated that it was going to abandon the area. In May, 1931, United Fruit representative William Turnbull sent a telegram to Omoa's Mayor, Samuel Garcia,

[4] Fruit companies actually tried to flood fields to flush out the fungus, but this procedure was costly and showed very little permanent improvements in fruit production (Marquardt, 2001). This practice of "flood fallowing" was abandoned in the 1950s.

[5] Note that the decision to abandon plantations cannot be entirely attributed to Panama disease. Weakening soil quality was another important cause for relocating. See Wardlaw (1929) for more discussion on this point.

saying "[P]resent business conditions do not allow us to continue absorbing the enormous losses that we have endured for several years in Cuyamel, a situation that we feel has not been appreciated." By 1932, the company began removing branch lines between Cuyamel and Omoa, and growers made public appeals for help with finding a way to transport their produce. In 1933, Alonso Valenzuela, a Honduran official, inspected the region, describing it as follows:

> It's a pity to see the comparison between 1916 and today: then, banana farms covered all of the valleys and the level of commerce was astonishing; today everything is desolate, dead. The valleys are all *guamiles* and it's hard to find a banana plant.

Valenzuela further noted that after United Fruit had left Omoa-Cuyamel, the "greater part" of the inhabitants remained in the area, substituting away from banana production to the cultivation of grains or livestock. Without access to the railroad network to export bananas or other commercial fruit, growers turned to subsistence production. Soluri (2005, p. 87) remarks that "[r]ailroads and export banana production on the North Coast had developed hand-in-hand during the twentieth century; the loss of one tended to spell the end of the other".

2.2.4 Conditional Random Abandonment

This discussion of history and botany has been written entirely for the purpose of supporting one key identifying assumption, namely that conditional on agro-climatic variables (soil quality, slope, elevation) and the railroad network in 1934, the railroad network that was left behind in the 1970s was randomly assigned. The ravages of Panama disease were difficult to predict and confounded plantation owners during the 1940s and 1950s, causing sharp declines in production. Faced with this and a lack of an available cure, owners were forced to abandon their lands, and in doing so they ripped up the railway lines. Once we condition on variables influencing the selection of banana plantations, in addition to the original railway network, the railway network that existed in the 1970s was randomly assigned.

2.3 Model

In this section, we present a simple spatial equilibrium model, based on Roback (1982) and Rosen (1979), which will be used to derive the implicit prices for amenities, such as transport infrastructure, and their effects on the costs of production. Workers and firms are assumed to be perfectly mobile across space, and in this sense, it is meant to capture long run equilibrium in land and labor markets.

Locations, or communities, are indexed by $c = 1, ..., C$, and each community belongs to a municipality, indexed by $m = 1, ..., M$. Conditional on locating in community c, workers choose quantities of housing, H, and a composite commodity, X, to maximize their utility:

$$\max_{X,H} U(X, H; A_c) \quad \text{s.t.} \quad X + r_c H \leq w_c$$

Here, r_c denotes the local price of housing and w_c denotes the local wage. The composite commodity, X, is assumed to be freely traded across locations, and its price is normalized to 1. Communities are endowed with a vector of amenities, denoted by A_c, which also influence utility. In spatial equilibrium, because of perfect mobility, households will have equal realized utility across locations, so we must have the following:

$$V(w_c, r_c; A_c) = \bar{u} \tag{2.1}$$

where $V(\cdot)$ denotes the consumers' indirect utility function.

A firm in location c chooses cost minimizing quantities of land, H and labor, N, to produce a composite commodity, X:

$$\min_{N,H,K} w_c N + r_c H \quad \text{s.t.} \quad X = F(N, H; A_c)$$

Firm production functions depend on local amenities, A_c. Free entry of firms and perfect mobility of firms across locations requires that unit production costs are equal to the price of output everywhere:

$$1 = C(w_c, r_c; A_c) \tag{2.2}$$

A spatial equilibrium denotes the set of wages and rents that ensures that equations (2.1) and (2.2) hold simultaneously.

2.3.1 Implicit Prices

Using this model, we can derive the implicit price that a worker pays for a marginal increase in an amenity, A_c. First, we totally differentiate (2.1) with respect to amenity A:

$$0 = \frac{\partial V}{\partial w_c}\frac{\partial w_c}{\partial A} + \frac{\partial V}{\partial r_c}\frac{\partial r_c}{\partial A} + \frac{\partial V}{\partial A}$$

Next, we solve for the implicit price:

$$p_A \equiv \frac{\partial V/\partial A}{\partial V/\partial w_c} = -\frac{\partial V/\partial r_c}{\partial V/\partial w_c}\frac{\partial r_c}{\partial A} - \frac{\partial w_c}{\partial A}$$

Finally, we invoke Roy's Identity, which states that $H_c = -V_r/V_w$, and rearrange terms to obtain the following:

$$p_A = r_c H_c \frac{\partial \log r_c}{\partial A} - w_c \frac{\partial \log w_c}{\partial A} \tag{2.3}$$

This gives us an expression for the marginal rate of substitution between the amenity, A, and money, or the implicit marginal valuation that consumers place on A. To live in a location with greater amenities, consumers will pay higher land prices, but those land costs are offset by the effect that amenities have on wages. This formula for consumers' marginal willingness to pay depends crucially on the how wages and rents change with changes in amenities, which we will estimate below.

2.3.2 Cost of Production

To understand the impact on firms' production costs and the value of output, we totally differentiate (2.2) with respect to A and rearrange:

$$-\frac{\partial C}{\partial A} = \frac{\partial C}{\partial w_c}\frac{\partial w_c}{\partial A} + \frac{\partial C}{\partial r_c}\frac{\partial r_c}{\partial A}$$

This expression tells us how unit costs decrease as we increase A, which is a measure of the implicit effect of the amenity on firms' productivity. In long run spatial equilibrium, we have free entry and perfect mobility across space. So, if an area experiences a marginal improvement in an amenity that lowers production costs (i.e. $-\partial C/\partial A > 0$), firms must pay more for land in labor in those areas, and wages and rents must increase to exactly compensate for this reduction in production costs. To simplify the expression above, we can make use of Sheppard's Lemma to obtain the following result:

$$-\frac{\partial C}{\partial A} = \left(\frac{w_c N_c}{X}\right)\frac{\partial \log w_c}{\partial A} + \left(\frac{r_c H_c}{X}\right)\frac{\partial \log r_c}{\partial A} \tag{2.4}$$

where the terms in parentheses represent the share of the total value of labor and land in the value of output, respectively.

2.4 Data

To estimate the implicit prices and cost of production effects of infrastructure access, we combine data from a variety of sources. In this section, I describe each of these data sources and summarize variables in turn.

2.4.1 Spatial Unit of Analysis

Throughout the paper, our analysis sample focuses only the north coast of Honduras, since this is the area most affected by the changes in transport infrastructure that took place as a consequence of the ravages of Panama disease. Our north coast sample comprises five of Honduras' 18 departments (Atlantida, Colon, Cortes, Gracias a Dios, and Yoro) and was home to roughly 32 percent of Honduras' 1.8 million people in 1988, according to census data. These five departments, which represent 36 percent of Honduras's total land area, are depicted in Figure 2.2.

Our primary spatial unit of analysis is an aldea, or village. The north coast departments are divided into 1,046 aldeas, with a median area of 13 km^2 and a median population of 481. North coast aldeas are somewhat comparable to U.S. zip codes, but are considerably smaller in terms of both population and area. In 2000, the median population of a U.S. zip code was 2,500, and the median area was 94 km^2.

2.4.2 Infrastructure Maps and Access to Railroads

To understand the changes in railroad infrastructure that took place as a result of banana companies' reactions to Panama disease, we digitally traced the railway lines from several map series produced by cartographic units of the U.S. Military. For our measure of railway access circa 1934, we work with U.S. Military Intelligence maps that were produced in 1934, projecting and tracing them digitally using commercial GIS software. These maps are drawn at a scale of 1:250,000, and three maps at this scale cover the railway lines on the north coast.[6]

For railway access in the 1970s, we used another series of 1:50,000 maps produced by the *Honduran Instituto Geografico Nacional*, often in cooperation with the U.S. Defense Mapping Agency Inter-American Geodetic Survey. Because the more recent maps are considerably more accurate than the 1930s maps, and because some of the more recent maps contain information on abandoned railroads, the 1970s network data were used as a guide to realign the 1930s railroads. Detailed information about all maps used to construct our dataset can be found in Appendix Table 2.12.

Figure 2.3 shows the map of Honduras' railroad network on the north coast and its evolution from 1934 (Panel A) to the 1970s (Panel B). Several changes are readily apparent, including the complete abandonment of lines that used to run east to La Mosquitia. Measures of access to railroad infrastructure were calculated by finding the centroid of each aldea polygon and determining how far it was to the nearest point on the network, using both network maps. On average, a village on the north coast was 15.8 km from the closest point on the network in 1934, but by the 1970s, this average distance had increased to 27.7 km,

[6] The accuracy of the railway network depicted on these U.S. Military maps was confirmed by examining several other maps drawn by banana companies during the same time period. These maps were not used in actually tracing the network, because they often contained insufficient projection and coordinate information.

an average change of 11.94 km.

Table 2.1 shows the percentage of aldea observations that lie within different distance bins. Roughly 73 percent of observations lie along the diagonal, belonging to the same distance bins in 1934 and in the 1970s. These observations can be further divided into the 22 percent of the north coast villages (214 / 983) villages that experienced no changes in railway access and the 51 percent that experienced reductions in access of less than 5 km. The remaining 27 percent of villages lie in the grey shaded upper right triangle, corresponding to observations with increases in distance to the railroads of greater than 5 km. Figure 2.4 presents a scatterplot of the two distance variables, which highlights that most villages either experienced small or no changes in railway access, while others experienced dramatic increases in distances to the railroads. A histogram of the distribution of changes in railway access is presented in Figure 2.5.

While we work with a continuous distance measure, it is also useful to construct a binary measure of railroad abandonment. Let c index aldeas, let Rr_c^{1934} denote the distance in kilometers from the centroid of aldea c to the 1934 railway network, and define Rr_c^{1970} analogously. Restricting our attention to the set of all aldeas with $Rr_c^{1934} \leq \kappa$, where κ is some predetermined constant, we define an indicator of abandonment as follows:

$$A_c^{\kappa} = 1\left\{Rr_c^{1970} - Rr_c^{1934} \geq 5\right\} C\left\{Rr_c^{1934} \geq \kappa\right\} \tag{2.5}$$

where $C\{\cdot\}$ is a censoring indicator. This abandonment indicator is 1 if the village was less than κ kilometers from the railway in 1934, but experienced an increase in distance greater than 5 kilometers after the fruit companies abandoned railway lines. It is undefined (censored) for villages that were greater than κ kilometers away from the railway in 1934.

We primarily work with A_c^{30}, but we investigate the robustness of these results by varying the censoring width; a smaller width approaches a better treatment vs. control comparison, at the cost of losing data. A map of A_c^{30} is depicted in Figure 2.6, and it should provide a good sense of the spatial variation we utilize. Of the 1,046 aldeas on the north coast, 210 were abandoned with this indicator, 669 were non-abandoned, and 167 were censored. Among the abandoned aldeas, the average increase in distance to the railroads was 25 kilometers, with a standard deviation of 29.2.

2.4.3 Census Data

The wage, rent, and population data that are used to estimate hedonic regression parameters are taken from the *Censo Nacional de Población, 1988*. We have access to the entire universe of long form, unit-level data, which includes information from both a housing module and also an individual wage module. The wage module contains nearly 50 variables on individuals' demographic and employment characteristics, while the housing module contains over 40 variables on housing characteristics and costs.

Summary statistics for all of the individual-level variables used in the analysis appear in Table 2.2. Restricting the sample to all individuals over 18 who lived on the north coast municipalities of Honduras, we have 592,494 individual observations. Of those observations, only 184,430 reported any wages, despite the fact that 54 percent of the sample reported that they were employed. The average worker in our sample was 34 years old, with 5 years of schooling and 22 years of experience. The average monthly earnings for workers who reported them was Lp 518 (USD 259) expressed in 1988 prices.

Missing wages were more common for women, and only 16% of females reported any earnings. This was mostly due to the fact that few females were working; 75% did not report that they were working, and only 9% of female observations were working but had missing wages. For males, 48% reported earnings, 14% were not working, and 38% were working but did not report any earnings.

Variables used for the housing analysis are summarized in Table 2.3. There were 289,825 housing observations, 88 percent of which were made of independent houses, 10 percent of apartments or rooms at inns, and the remaining 2 percent of temporary homes. Unfortunately, rental information was only collected from 53,863 observations, less than 19 percent of the total sample. Some of the missingness can be explained by housing type and tenure status; for instance, 76 percent of houses and 60 percent of owned units did not report any rental information. Among houses for which rental data are available, the mean monthly rental expenditure was Lp 132 (USD 66) expressed in 1988 prices. In our opinion, the poor coverage of our rental rates data for various types of housing and the absence of housing price observations in many aldeas exposed to abandonment mean that we should not take the housing results very seriously.

2.4.4 Terrain Variables

In our empirical work, we control for a number of measures of elevation, slope, ruggedness, and soil quality, as these measures were likely to influence the decisions of banana companies to abandon plantations and remove railway lines. Banana companies often developed plantations on low-lying, flat, alluvial soils that straddle the rivers (Stover and Simmonds, 1987), and it is important to control for these aspects of the terrain in our analysis. The elevation and slope measures were constructed from 30 arc-second digital elevation maps, available from the Harmonized World Soil Index, Volume 2.[7] To capture dispersion in the topography within a village, a vector ruggedness measure was constructed using the methodology described by Sappington et al. (2007). These variables are summarized in Table 2.4.

We also constructed several measures of soil quality to determine which areas in Honduras would be suitable for growing bananas. Our measures are based on two digitized soil maps

[7] FAO/IIASA/ISRIC/ISSCAS/JRC, 2012. Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, Laxenburg, Austria, available online at `http://www.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/`.

from the International Center for Tropical Agriculture's (CIAT) *Atlas De Honduras.* The first map, drawn at a 1:1,000,000 scale, is a map of soil typography that uses the soil classification scheme of L. LeForrest Miller, adopted by the Food and Agricultural Organization (FAO). The soils most suitable for growing bananas are fluvisols, which are typically young soils found in alluvial deposits. Most fluvisols have good natural fertility and provide access to sufficient water and nutrients for growing bananas. The variable suel_fao1 measures the percentage of each aldea that is covered by fluvisols from alluvial plains, while the variable suel_fao2 measures the percentage covered by fluvisols from alluvial terraces.[8]

The next set of measures are based on a soil map drawn at a 1:500,000 scale that uses the soil classification taxonomy of Simmons and Castellanos (1968). Three variables measure the percentage of each aldea covered by well drained alluvial soils that are coarse or fine textured (suelsimm_agaf), fine textured (suelsimm_am), or well drained alluvial soils without textural classification (suelsimm_as). A fourth variable measures the share of each village covered by soils of the valleys (suelsimm_sv). Maps of all of the soil quality variables can be found in Figure 2.7.

2.5 Results

2.5.1 Selection and Abandonment

We begin by trying to understand which variables influenced the initial placement of railroads, and which variables were associated with abandonment. The first four columns of Table 2.5 present results from logit regressions that attempt to predict the probability that different villages received access to railroad infrastructure by the end of 1934. The dependent variable is in these first four columns is equal to 1 if a village was less than 5 kilometers from a railway line in 1934 and 0 otherwise.

The first column predicts initial placement using only the soil quality measures, and as expected, many variables capturing fluvisols and well drained alluvial soils are positive and significantly different from zero. The second column predicts placement using only the topographical variables. From this column, it appears that the fruit companies avoided areas that had steep slopes, were more elevated, and had more rugged terrain. The avoidance of rough terrain and high elevation coincides with the historical narrative on the variables banana companies used for selection. For instance, Marquardt (2001) quotes a 1916 letter from a United Fruit plantation manager in Costa Rica that he was "carefully avoiding all

[8] More specifically, the variable suel_fao1 measures the percentage of each village that is covered by either tropical fluvaquents (EAFh-1), tropofluvents (EFTa-1), ustifluvents (ITYe-1), or oxic dystropepts (ITYe-1) and associated soils from alluvial plains. The variable suel_fao2 measures the percentage of each village covered by aquacultural tropofluvents (EFTb-1), aquacultural ustifluvents (EFUb-1), tropaquepts and ustorthents (IATd-1), dystropepts oxicos (ITYe-3), Umbric Tropaquults (UATd-1) and associated soils from alluvial terraces.

hilly land" and using other intuitive standards for site selection. Soil science was not well developed during the early 20th century, and it seems unlikely that the banana companies had considerably better information than we have.[9] The third column controls for latitude and longitude to predict initial selection, as well as distance to historical ports of Puerto Cortes and Trujillo, and many variables are highly significant. In the fourth column, all variables are included in the same specification, and while some variables change signs and become insignificant, the overall story remains the same. Moreover, our final specification has a pseudo R^2 of 0.535, suggesting that we can do a good job of explaining which areas were given infrastructure access by these limited terrain characteristics.

In the next four columns, this analysis is repeated for the abandonment indicator, A_c^{30}, defined in (2.5); the differences in the sample sizes reflect censoring of the abandonment indicator. Here, we also include distance to the railroads in 1934 as an explanatory variable. In column 5, it appears that the fruit companies were more likely to abandon areas covered with soils from alluvial terraces. In column 6, the fruit companies displayed some tendency to abandon areas that are more rugged, and to keep areas with very low gradient, and in column 7, the latitude and longitude polynomial explains roughly 20 percent of the abandonment variation. In column 8, we include all variables in the specification, and again few variables change signs or become insignificant unexpectedly. However, we can explain only 30 percent of the variation in abandonment with these variables. Our identifying assumption is that because of the unpredictable incidence and effects of Panama disease, after conditioning on these variables, abandonment is randomly assigned.

2.5.2 Persistent Changes in Infrastructure

The areas with abandoned infrastructure as a consequence of Panama disease have considerably poorer access to infrastructure today, by several measures. In Table 2.6, we report a comparison of means between abandoned and control areas for a variety of measures of infrastructure access. The first two rows examine electricity and access to piped water from housing observations in the 1988 census; households in abandoned areas were 12 percent less likely to have access to electricity, but there were no significant differences in access to piped water. However, from a nationally representative household survey conducted in 2004 (with substantially smaller spatial coverage), households in abandoned areas were 11 percent less likely to have access to piped water.[10] They were also 18 percent less likely to have access to a toilet connected to a sewer, 16 percent less likely to have public trash removal, 30 percent

[9] One concern is that our soil quality measures were conducted after the banana companies had been around in the region for some time. To the extent that banana growing in the 1930s worsens soil quality metrics collected in the 1960s and 1970s, our estimates may have bias.

[10] These data are from the National Household Survey on Living Conditions (ENCOVI), which was conducted by Instituto Nacional de Estadística (INE) between July 31 and November 30, 2004. This survey incorporated households from both rural and urban areas of the country's 18 departments.

less likely to have public street lighting, and 8 percent less likely to have access to a land phone line.

Distance to paved and dirt roads was also much larger in 2001 for abandoned villages than for the control group.[11] On average, relative to control villages, abandoned villages were 6.5 kilometers away from primary paved roads, 17 kilometers away from secondary paved roads, and 10 kilometers away from secondary dirt roads.

It is very likely that the abandonment by fruit companies involved more than just removal of transport infrastructure. Fruit companies provided housing and schools for employees, food stores, and medical services to their communities (Hord, 1966). Because of the permanent impact of abandonment on several different measures of infrastructure access, it is important to keep a broad interpretation for our estimates.

2.5.3 Effects on Wages, Rents, and Density

Figure 2.8 plots kernel density estimates of the distributions of log wages (Panel A), log rents (Panel B), log population (Panel C) and log density (Panel D) for the abandoned and non-abandoned areas. For all variables except population, mean shifts in the distributions are readily apparent, and Kolmogorov-Smirnov equality-of-distributions tests strongly reject the hypothesis of equal distributions at conventional significance levels.[12] It seems apparent that abandoned villages had lower wages, rents, and population density, but there could be a variety of reasons for these differences. Lower skill distributions in the abandoned population or poorer quality housing characteristics could explain these differences. Moreover, the various factors that determined whether or not a village would be abandoned, such as poorer quality soil, steeper slopes, or higher elevation, could adversely affect rents and wages in those areas today. For these reasons, a regression model will be used to adjust for these differences to isolate the causal impact of abandonment.

We first estimate two sets of linear hedonic regressions, both of which take the following form:

$$\log y_{ci} = \alpha_c + x_{ci}'\beta + \varepsilon_{ci}$$

where y_{ci} denotes either wages (rents), c indexes communities, and i indexes individuals (households). The vector x_{ci} contains a set of explanatory variables, diffent for each dependent variable. For the wage variables, we use years of schooling and experience, while for the rent variables, we use a rich vector of housing characteristics. Each regression includes a full set of aldea-specific intercepts, α_c, which is of primary interest. Notice that the esti-

[11] These variables were constructed from the International Center for Tropical Agriculture's (CIAT) *Atlas De Honduras*, the same dataset used to construct soil quality measures.

[12] The Kolmogorov-Smirnov test is used to test the null hypothesis of equal distributions against the alternative. For log wages, the test statistic was 0.249, with a corrected p-value of 0.000. For log rents, the test statistic was 0.367, with a corrected p-value of 0.000. For log population, the test statistic was 0.079, with a corrected p-value of 0.259. For log density, the test statistic was 0.156, with a p-value of 0.001.

mated village intercept for village c is just the average of the residuals for all $i = 1, ..., N_c$ observations in that village:

$$\widehat{\alpha_c} = \frac{1}{N_c} \sum_{i=1}^{N_c} \left(y_{ic} - x'_{ci} \widehat{\beta} \right)$$

This represents the village-average variation in y that, by construction, is orthogonal to the individual level-characteristics. This is the sense in which we are adjusting for differences in the quality of the housing stock or skill composition of the population.

After estimating these village-specific intercepts, we regress them on our measure of abandonment, A_c, and a vector of aldea-level characteristics, X_c:

$$\widehat{\alpha_c} = X'_c \theta + \delta A_c + \upsilon_c$$

where A_c is either our continuous or discrete abandonment measure, and X_c is a vector of characteristics influencing selection and abandonment by the fruit companies, including railway access in 1934, terrain characteristics such as measures of soil quality, slope, elevation, and ruggedness, and measures of distance to historical ports and the coast. Our identifying assumption is that conditional on this set of characteristics, X_c, abandonment is randomly assigned. For the linear model, our identifying assumption is equivalent to asserting the following conditional moment restriction:

$$\mathbb{E} \left[\upsilon_c \,|\, X_c, A_c \right] = 0$$

These regressions are estimated by weighted least squares, using as weights the number of observations used to construct each village-level intercept.

Results for the individual-level wage regressions are reported in Table 2.7. The first column contains results using the full sample of observations, while the second and third columns restrict the sample to urban or rural areas, and the fourth column excludes the department of Gracias a Dios. Robust standard errors, clustered at the aldea level, are reported. The coefficients on schooling and experience are significant and of the expected signs. In Table 2.8, we report parameter estimates from the hedonic rental regressions, where again we use different models across the columns of the table for our different restricted subsamples. The housing characteristics we use are very detailed and include indicators for the type of housing unit (room, apartment, or temporary unit, with house being the excluded indicator) and indicators for whether or not the walls, roof, and floor are made of quality building material. Variables measuring the number of rooms in the unit, indicators for whether or not the unit has a kitchen, piped water, a functioning toilet, or electricity were also included. Flexible effects for the age of the housing unit were also included. These variables generally significantly different from zero at conventional levels and typically appropriately signed.

In the second step of our estimation procedure, we estimate village-level regressions for the log wage and log rent intercepts, as well as for log population density, and the results are presented in Table 2.9. As shown in the first three columns, log wages are 0.153 points lower, log rents are 0.346 points lower, and log density is 0.522 points lower for abandoned areas, conditional on other observable characteristics. The continuous distance measure is used in columns 4, 5, and 6; converted to an elasticity, a 10 percent increase in distance to the railways in the 1970s is associated with a 0.6 percent reduction in wages, a 0.8 percent reduction in log rents, and a 1.3 percent reduction in density, though the coefficient on density is not precisely estimated.

In any selection on observables design using observational data, there remains the possibility that unobserved variables influenced selection. If fruit companies had better data than we did about the quality of growing bananas in different areas, and if that information influenced abandonment, our estimates might be biased. Using a technique developed by Altonji et al. (2005), we show that the amount of selection on unobserved variables would have to be 1.4 times as large as the selection on observed variables to entirely explain away the wage effects, and 2.5 times as large to entirely explain away the effects on density. However, selection on unobservables need only be 80 percent of the selection on observed variables to explain away the effects on rents.[13]

In Table 2.10, we allow for more flexibility in specifying the relationship between distance to the railroads and our outcome variables. Here, instead of a continuous distance measure, we include separate indicators for villages being of a certain distance away from the railway lines. After fitting these regression functions, we calculate the predicted marginal effects of being in different distance bins by plotting the average value of y observations that belong to these distance bins after averaging over the remaining covariates. Plots of these marginal effects for log wages, log rents, and log density can be found in Figure 2.9. All three figures display a dramatic drop in the dependent variable for distances of greater than 5 kilometers. Increasing distance to the railroads from 0-5 kilometers to 5-10 kilometers is associated with an approximately 12 percent reduction in wages, a 25 percent reduction in rents, and a 67 percent reduction in density. However, beyond that initial increase in distance, further increases do not seem to matter for either of the variables we study.

2.5.4 Implicit Prices and Production Costs for Abandonment

In Table 2.11, we use our estimates of the effects of abandonment on wages and rents to estimate the implicit prices that consumers pay for living in non-abandoned areas, and the implicit costs that firms pay to produce in those areas. Column 1 reports our average marginal willingness to pay to live in a non-abandoned areas; this corresponds to a calculation of equation (2.3). To obtain the rental contribution, we multiply the estimated effect of being in a non-abandoned area on log rents of 0.346 (from Table 2.9, Column 2) by the average

[13] The exact implementation of the test for linear models is discussed in Bellows and Miguel (2009), Appendix A.

annual housing expenditure, which in 1988 was Lp 1743.95, according to the census data. We subtract from this the wage contribution, obtained by multiplying the effect of being in a non-abandoned area on log wages of 0.153 (from Table 2.9, Column 1) by the average annual household wage, of Lp 7500.40. The resulting estimate of the implicit price is negative, but not statistically distinguishable from zero. The 95 percent confidence interval of this implicit price, $(-1324.35, 240.43)$, is quite large and centered in negative territory largely because of the small and imprecise rental estimate. Moreover, the average housing expenditure is estimated to be quite small; on average, a Honduran resident only spends 23 percent of his income on housing, while in the United States, that number is closer to 33 percent.

The implicit costs that firms pay to produce in non-abandoned areas are estimated in column 2. The rental contribution to firms' costs, from equation (2.4), is estimated by multiplying the share of the value of land in output (0.395) by the total rental effect of 0.345.[14] The wage contribution to firms costs is obtained by multiplying the share of the value of labor in output (0.531) by the log wage effect of being in a non-abandoned area of 0.153. We estimate that firms pay roughly 22 percent of unit costs to be able to produce in non-abandoned areas. While this effect is statistically different from zero, it is somewhat small, most likely because of the small estimated effect on rents.

2.6 Conclusion

This paper has tried to estimate firm and workers' marginal willingness to pay for improvements to transport infrastructure by making use of a novel natural experiment: Honduras' infestation with Panama disease. Fruit companies growing bananas for export along the North Coast of Honduras responded to an outbreak of Panama disease by abandoning banana plantations and removing infrastructure, and these changes in infrastructure reflect the current configuration of railroads and roads today. We argue that conditional random abandonment, a form of selection on observables, is likely to hold in this setting, and we use it to identify and estimate hedonic wage, rent, and population regressions. By combining the estimates of the effects of transport infrastructure on wages and rents with a simple spatial equilibrium model, we can provide estimates of workers' and firms' valuations for this productive amenity.

Unfortunately, the estimates we present in this paper are small and imprecise, but this is likely a product of certain data limitations, particularly the poor rent data we use. Further research will allow us to find better sources of data so that we can properly execute our identification strategy, and we feel confident that such data exist.

There are several important limitations of this paper and the approach it takes. The

[14] This figure is from the United Nations (2000) National Account Statistics for Honduras. Another reason our the rental data are suspect is that they provide a very small valuation for the share of total rental expenditures in output (0.02).

model assumes that all workers are identical and perfectly mobile across space, but it is likely that heterogenous mobility costs or idiosyncratic ties to locations are important even in the long run. Kline (2010) and Moretti (2010) have shown that these assumptions can have implications for how to think about willingness to pay, and a future model will incorporate these effects. Another limitation is that the banana companies provided more than just transport infrastructure, and the effect of abandonment needs to be carefully interpreted. Moreover, while this approach enables us to identify willingness to pay for a marginal improvement in transport infrastructure, valuation for non-marginal improvements may differ in important ways. We leave these important issues to further research.

Table 2.1: Percent of Village Observations in Distance Bins

		Distance in 1970s					
		0-5 km	**5-10 km**	**10-15 km**	**15-20 km**	**20-25 km**	**> 25 km**
	0-5 km	36.11	5.60	4.07	2.24	1.42	3.76
	5-10 km	0.00	11.80	2.03	1.02	0.51	1.02
Distance	**10-15 km**	0.00	0.00	6.00	1.73	0.51	0.92
in 1934	**15-20 km**	0.00	0.00	0.00	3.87	0.31	0.71
	20-25 km	0.00	0.00	0.00	0.00	2.54	0.71
	> 25 km	0.00	0.00	0.00	0.00	0.00	13.12

Source: Authors' calculations. There were 983 north coast villages used in these calculations.

Table 2.2: Summary Statistics: 1988 Census Individual Wage Variables

Variable Name	N	N Missing	Mean	SD	Description
Labor Market Variables					
Employed	592494	0	0.54	(0.50)	Individual is currently employed (perhaps w/out pay)
HoursWorkedLastWeek	515334	77160	22.28	(23.91)	Number of hours individual worked *last* week
MonthlyEarnings	184430	408064	518.16	(1434.17)	Monthly earnings in Lps
LogMonthlyEarnings	184430	408064	5.72	(0.90)	(Log) Monthly earnings in Lps
Demographic Characteristics					
age	592494	0	33.92	(12.31)	Age in years, Source: 1988 census
YrsSch	592494	0	4.08	(4.21)	Years of schooling completed; 1988 Census
exp	592494	0	22.83	(14.07)	Experience; 1988 Census
female	592494	0	0.52	(0.50)	Individual is female; 1988 Census
femXSch	592494	0	2.11	(3.62)	Female x Years of Schooling
femXexp	592494	0	11.67	(15.15)	Female x Experience
Information on Missing Wages					
EarningsMissing0M	285550	306944	0.48	(0.50)	0: Earnings Observed, males
EarningsMissing1M	285550	306944	0.38	(0.49)	1: Working but missing, males
EarningsMissing2M	285550	306944	0.14	(0.35)	2: Not working, males
EarningsMissing0F	306944	285550	0.16	(0.36)	0: Earnings Observed, females
EarningsMissing1F	306944	285550	0.09	(0.29)	1: Working but missing, females
EarningsMissing2F	306944	285550	0.75	(0.43)	2: Not working, females
Aldea Means					
Employed	1000	46	0.51	(0.12)	Individual is currently employed (perhaps w/out pay)
MonthlyEarnings	948	98	380.52	(393.81)	Monthly earnings in Lps
LogMonthlyEarnings	948	98	5.41	(0.49)	(Log) Monthly earnings in Lps

SOURCE: 1988 Census, authors' calculations. Sample of workers includes all individuals aged 18 to 65 from North Coast aldeas.

Table 2.3: Summary Statistics: 1988 Census Household Rent Variables

Variable Name	N	Miss	Mean	SD	Description
Housing Characteristics					
rent	53311	235962	118.26	(204.27)	Monthly household rent, in Lps.; 1988 census
log_rent	53311	235962	4.21	(1.04)	Log of monthly household rent, in Lps.; 1988 census
WallsGood	288071	1202	0.73	(0.44)	Brick, stone, cement, adobe, or wood walls; 1988 census
RoofGood	288071	1202	0.14	(0.34)	Tile, cement, or concrete roof; 1988 census
FloorGood	286610	2663	0.68	(0.47)	Cement, wood, or brick floor; 1988 census
NumRooms	250558	38715	2.66	(1.62)	Number of rooms in house; 1988 census
NumBedrooms	247061	42212	1.58	(0.91)	Number of bedrooms in house; 1988 census
kitchen	250558	38715	0.72	(0.45)	House contains a kitchen; 1988 census
pipedWater	250558	38715	0.64	(0.48)	House has access to piped water; 1988 census
toilet	181466	107807	0.50	(0.50)	House has a functioning toilet; 1988 census
electricity	250558	38715	0.52	(0.50)	House has access to electricity; 1988 census
agePre74	227191	62082	0.25	(0.43)	House built before 1974; 1988 census
ageb7480	227191	62082	0.17	(0.38)	House built between 1974 and 1980; 1988 census
ageb8182	227191	62082	0.09	(0.29)	House built between 1981 and 1982; 1988 census
ageb8384	227191	62082	0.12	(0.32)	House built between 1983 and 1984; 1988 census
ageb8586	227191	62082	0.13	(0.33)	House built between 1985 and 1986; 1988 census
ageb8788	227191	62082	0.11	(0.32)	House built between 1987 and 1988; 1988 census
Type of Tenure					
TenureOwned	250558	38715	0.69	(0.46)	House is owned; 1988 census
TenureInstall	250558	38715	0.06	(0.23)	House owned but paid for in installments; 1988 census
TenureLease	250558	38715	0.16	(0.36)	House is leased; 1988 census
Type of Residence					
TypeHouse	289273	0	0.88	(0.32)	Independent house; 1988 census
TypeRoom	289273	0	0.07	(0.25)	Room at inn or boarding house; 1988 census
TypeApt	289273	0	0.03	(0.17)	Apartment; 1988 census
TypeTemp	289273	0	0.02	(0.12)	Makeshift house; 1988 census
Information on Missing Rents					
RentMissingOwned	289273	0	0.60	(0.49)	Unit is owned and rent is missing
RentMissingInstall	289273	0	0.00	(0.00)	Unit is paid in installments and rent is missing
RentMissingLease	289273	0	0.00	(0.01)	Unit is leased and rent is missing
RentMissingHouse	289273	0	0.76	(0.43)	Unit is a house and rent is missing
RentMissingRoom	289273	0	0.02	(0.15)	Unit is a room in a house / inn and rent is missing
RentMissingApt	289273	0	0.01	(0.10)	Unit is an apartment and rent is missing
RentMissingTemp	289273	0	0.01	(0.11)	Unit is temporary and rent is missing

SOURCE: 1988 Census, authors' calculations. Sample includes all residential units from North Coast aldeas.

Table 2.4: Summary Statistics: Aldea-Level Variables

Variable Name	N	Miss	Mean	SD	Description
Demographic Variables					
pop88	1001	45	1362.01	(9366.61)	Total aldea population; Source: 1988 Census
popdens88	945	101	90.28	(200.48)	Population density in 1988 (people per km2)
hh_size88	1001	45	7.19	(3.56)	Average household size; Source: 1988 Census
Distances to Roads and Railroads					
rr34_cent_edist	983	63	15.83	(33.99)	village centroid distance to railroads, 1934 (km)
rr50_cent_edist	983	63	27.78	(58.66)	village centroid distance to railroads, 1950 (km)
rr_diff_cent_edist	983	63	11.94	(29.41)	Change in railroad access (1950s-1934)
prim_paved_cent	983	63	19.92	(48.73)	centroid dist. to primary paved roads, 2001 (km)
sec_paved_cent	983	63	47.11	(68.84)	centroid dist. to sec paved roads, 2001 (km)
Access to Other Forms of Infrastructure					
piped_water_public	122	924	0.17	(0.31)	% of HH w/ piped water, public, 2004
piped_water_private	122	924	0.61	(0.38)	% of HH w/ piped water, private, 2004
toilet_sewer	122	924	0.14	(0.28)	% of HH w/ toilet connected to sewer, 2004
trash_removal	122	924	0.13	(0.28)	% of HH w/ public trash removal, 2004
street_lighting_public	122	924	0.65	(0.43)	% of HH w/ public street lights, 2004
street_lighting_private	122	924	0.49	(0.38)	% of HH w/ private street lights, 2004
street_lighting_gen	122	924	0.49	(0.38)	% of HH w/ private generator, 2004
land_phone	122	924	0.07	(0.16)	% of HH w/ land phone lines, 2004
Agroclimatic Variables					
area	984	62	40.73	(144.80)	village area (square km)
suel_fao1	983	63	0.21	(0.35)	% with fluvisols from alluvial plains
suel_fao2	983	63	0.04	(0.15)	% with fluvisols from alluvial terraces
suelsimm_agaf	983	63	0.10	(0.25)	% with well drained alluvial soils, coarse or fine
suelsimm_am	983	63	0.15	(0.30)	% with well drained alluvial soils, fine
suelsimm_as	983	63	0.00	(0.05)	% with well drained alluvial soils, other
suelsimm_sv	983	63	0.08	(0.22)	% with soils of the valley
Elev30as	978	68	275.95	(296.88)	Average elevation of aldea (meters), (HWSD)
slope02	928	118	0.52	(0.42)	% of village w. gradient between 0 and 2 percent
slope24	928	118	0.06	(0.13)	% of village w. gradient between 2 and 4 percent
slope46	928	118	0.07	(0.13)	% of village w. gradient between 4 and 6 percent
slope68	928	118	0.08	(0.13)	% of village w. gradient between 6 and 8 percent
slopeG8	928	118	0.28	(0.32)	% of village w. gradient greater than 8 percent
rugged3	969	77	0.23	(0.14)	Vector ruggedness measure, (3x3 window)

Note: Demographic variables were taken from the 1988 and 2001 Household Censuses. Distances to Roads and Railroads variables were computed from digitally traced railroad maps and from maps provided to us by INE. Access to Other Forms of Infrastructure variables were computed from the ENCOVI 2004. Agroclimatic variables were computed from a variety of sources, including soil maps and digital elevation maps provided to us by INE.

Table 2.5: Selection Regressions

	1934 Selection				**Abandonment**			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
main								
rr34_cent_edist					-0.060 (0.017)***	-0.013 (0.018)	0.015 (0.020)	-0.034 (0.028)
suel_fao1	0.899 (0.317)***			0.292 (0.534)	-1.247 (0.323)***			-2.256 (0.385)***
suel_fao2	0.321 (0.484)			-0.258 (0.535)	2.074 (0.533)***			2.456 (0.670)***
suelsimm_agaf	6.606 (1.158)***			4.928 (1.463)***	-0.569 (0.385)			-0.202 (0.461)
suelsimm_am	2.871 (0.341)***			1.308 (0.569)**	0.262 (0.314)			0.218 (0.393)
suelsimm_sv	1.642 (0.345)***			0.881 (0.537)	-0.553 (0.366)			-1.220 (0.604)**
Elev30as		-0.008 (0.001)***		-0.006 (0.001)***		-0.003 (0.001)***		-0.002 (0.001)**
slope02		-0.133 (0.426)		-0.032 (0.519)		-1.970 (0.392)***		-1.484 (0.508)***
slope24		-2.793 (0.801)***		-2.912 (0.826)***		-4.538 (0.876)***		-6.144 (1.482)***
slope46		1.803 (0.713)**		2.458 (0.750)***		-2.004 (0.761)***		-2.357 (0.946)**
slope68		-2.494 (0.749)***		-1.290 (0.832)		-0.894 (0.701)		-0.222 (0.906)
rugged3		-1.479 (0.666)**		-1.022 (0.822)		1.093 (0.562)*		0.349 (0.772)
x			0.001 (0.005)	-0.003 (0.007)			0.055 (0.012)***	0.104 (0.020)***
y			0.030 (0.004)***	0.012 (0.007)*			0.059 (0.007)***	0.068 (0.009)***
d_PuertoCortes			-0.013 (0.005)**	-0.008 (0.008)			0.005 (0.007)	0.009 (0.009)
d_Trujillo			-0.010 (0.002)***	-0.011 (0.003)***			0.064 (0.011)***	0.119 (0.019)***
Pseudo R^2	0.248	0.435	0.225	0.535	0.064	0.052	0.177	0.284
N	983	919	983	918	879	859	879	859

Note: This table reports parameter estimates from logit regressions where the unit of observation is a village. Parameters are estimated using maximum likelihood. In columns 1-4, the dependent variable equal to 1 if the village was less than 5 km from railway access in 1934, while in Columns 5-8, the dependent variable is equal to the abandonment indicator, A_c^{30}. Robust standard errors in parentheses (unclustered). * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level.

Table 2.6: Current Access to Infrastructure: Abandoned and Control

	Abandoned		Control		$H_0 : \mu_1 = \mu_2$	
	mean (sd)	N	mean (sd)	N	**t-stat**	**(p-value)**
% of housing obs. with electricity, 1988 census	0.27 (0.4)	95	0.38 (0.4)	317	-2.706	(0.01)
% of housing obs. with piped water, 1988 census	0.61 (0.4)	95	0.60 (0.4)	317	0.119	(0.91)
% of HH w/ piped water, public, 2004	0.09 (0.2)	28	0.20 (0.3)	84	-2.065	(0.04)
% of HH w/ toilet connected to sewer, 2004	0.01 (0.0)	28	0.19 (0.3)	84	-5.158	(0.00)
% of HH w/ public trash removal, 2004	0.02 (0.1)	28	0.18 (0.3)	84	-4.174	(0.00)
% of HH w/ public street lights, 2004	0.45 (0.4)	28	0.75 (0.4)	84	-3.394	(0.00)
% of HH w/ land phone lines, 2004	0.02 (0.1)	28	0.09 (0.2)	84	-3.326	(0.00)
centroid dist. to primary paved roads, 2001 (km)	13.17 (22.2)	210	6.67 (6.2)	669	4.194	(0.00)
centroid dist. to sec paved roads, 2001 (km)	45.30 (45.9)	210	27.90 (28.7)	669	5.183	(0.00)
centroid dist. to primary dirt roads, 2001 (km)	225.75 (70.0)	210	186.01 (47.8)	669	7.688	(0.00)
centroid dist. to sec dirt roads, 2001 (km)	23.32 (23.4)	210	13.32 (11.7)	669	5.975	(0.00)

Note: This table reports the sample means for various measures of access to infrastructure between abandoned ($A_c^{30} = 1$) and non-abandoned North Coast villages. The t-statistic reported is for a two-sided equality of means tests with unequal variances.

Table 2.7: Hedonic Wage Regressions, 1988

	(1)	**(2)**	**(3)**	**(4)**
YrsSch	0.054 (0.002)***	0.059 (0.002)***	0.043 (0.003)***	0.055 (0.002)***
YrsSch2	0.003 (0.000)***	0.003 (0.000)***	0.003 (0.000)***	0.003 (0.000)***
exp	0.048 (0.004)***	0.055 (0.003)***	0.030 (0.001)***	0.048 (0.004)***
exp2	-0.001 (0.000)***	-0.001 (0.000)***	-0.000 (0.000)***	-0.001 (0.000)***
Adj. R^2	0.367	0.292	0.344	0.369
N	136300	84960	51340	135150
Aldea-Specific Intercepts	X	X	X	X
Urban NC Only	.	X	.	.
Rural NC Only	.	.	X	.
Excluding Gracias a Dios	.	.	.	X

Note: Robust standard errors in parentheses, clustered at the aldea-level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level. Unit of observation is an individual. The estimation sample includes all Honduran male workers reporting wages, aged 18-64, who lived in the North Coast in 1988.

Table 2.8: Hedonic Rent Regressions, 1988

	(1)	(2)	(3)	(4)
TypeRoom	-0.210	-0.206	-0.139	-0.210
	(0.029)***	(0.031)***	(0.060)**	(0.030)***
TypeApt	0.052	0.062	-0.135	0.053
	(0.026)**	(0.027)**	(0.078)*	(0.026)**
TypeTemp	0.632	0.741	0.091	0.637
	(0.119)***	(0.111)***	(0.184)	(0.119)***
WallsGood	-0.064	-0.110	0.137	-0.067
	(0.102)	(0.097)	(0.068)**	(0.101)
RoofGood	0.131	0.124	0.215	0.131
	(0.045)***	(0.045)***	(0.064)***	(0.045)***
FloorGood	0.241	0.204	0.421	0.241
	(0.069)***	(0.067)***	(0.039)***	(0.069)***
NumRooms	0.179	0.183	0.118	0.180
	(0.011)***	(0.009)***	(0.027)***	(0.011)***
kitchen	0.069	0.065	0.075	0.069
	(0.019)***	(0.020)***	(0.044)*	(0.019)***
pipedWater	0.195	0.204	0.154	0.196
	(0.047)***	(0.051)***	(0.053)***	(0.047)***
toilet	0.380	0.386	0.382	0.379
	(0.088)***	(0.096)***	(0.093)***	(0.089)***
electricity	0.306	0.299	0.342	0.307
	(0.034)***	(0.038)***	(0.070)***	(0.034)***
agePre74	-0.100	-0.098	-0.136	-0.099
	(0.019)***	(0.019)***	(0.047)***	(0.018)***
ageb7480	-0.090	-0.097	-0.034	-0.090
	(0.058)	(0.063)	(0.063)	(0.058)
ageb8182	-0.009	-0.008	-0.064	-0.008
	(0.029)	(0.035)	(0.069)	(0.029)
ageb8384	-0.004	-0.006	-0.004	-0.004
	(0.038)	(0.041)	(0.088)	(0.038)
ageb8586	0.047	0.038	0.085	0.047
	(0.050)	(0.052)	(0.067)	(0.050)
Adj. R^2	0.471	0.426	0.559	0.472
N	49530	44837	4693	49376
Aldea-Specific Intercepts	X	X	X	X
Urban NC Only	.	X	.	.
Rural NC Only	.	.	X	.
Excluding Gracias a Dios	.	.	.	X

Robust standard errors in parentheses, clustered at the aldea-level. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level. Unit of observation is a residency. Sample includes all north coast Honduran residencies reporting rents in 1988.

Table 2.9: Village-Level Regressions

	Binary Treatment			**Continuous Treatment**		
	Wages **(1)**	**Rents** **(2)**	**Density** **(3)**	**Wages** **(4)**	**Rents** **(5)**	**Density** **(6)**
treated	-0.153 (0.035)***	-0.346 (0.174)**	-0.522 (0.200)***			
rr50_cent_edist				-0.006 (0.002)***	-0.022 (0.010)**	-0.012 (0.010)
rr34_cent_edist	-0.009 (0.003)***	-0.004 (0.012)	-0.040 (0.015)***	-0.001 (0.002)	0.038 (0.013)***	0.012 (0.013)
suel_fao1	0.149 (0.044)***	0.387 (0.169)**	0.368 (0.368)	0.158 (0.046)***	0.431 (0.172)**	0.480 (0.378)
suel_fao2	-0.006 (0.133)	0.236 (0.504)	-0.295 (0.599)	-0.031 (0.125)	0.030 (0.436)	-1.012 (0.566)*
suelsimm_agaf	0.170 (0.076)**	0.007 (0.221)	1.171 (0.563)**	0.194 (0.076)**	0.049 (0.233)	1.329 (0.572)**
suelsimm_am	-0.046 (0.054)	-0.104 (0.264)	-0.986 (0.336)***	-0.028 (0.055)	-0.007 (0.277)	-0.886 (0.345)**
suelsimm_sv	0.138 (0.067)**	0.243 (0.271)	1.427 (0.470)***	0.184 (0.061)***	0.378 (0.274)	1.721 (0.443)***
Elev30as	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.001 (0.000)	-0.001 (0.000)
slope02	0.114 (0.074)	-0.200 (0.313)	1.443 (0.357)***	0.132 (0.073)*	-0.340 (0.367)	1.506 (0.373)***
slope24	0.025 (0.105)	-1.067 (0.542)**	1.101 (0.520)**	0.119 (0.099)	-1.141 (0.493)**	1.022 (0.441)**
slope46	-0.054 (0.107)	-0.136 (0.385)	1.309 (0.523)**	0.147 (0.136)	0.080 (0.310)	2.405 (0.738)***
slope68	0.150 (0.133)	-0.183 (0.475)	0.250 (0.778)	0.074 (0.133)	-1.017 (0.627)	-0.261 (0.713)
rugged3	-0.203 (0.121)*	-0.813 (0.712)	-3.742 (0.633)***	-0.211 (0.125)*	-0.846 (0.705)	-3.793 (0.673)***
x	0.001 (0.002)	-0.033 (0.010)***	-0.018 (0.007)***	0.003 (0.002)**	-0.014 (0.007)*	-0.009 (0.007)
y	0.001 (0.001)	0.011 (0.004)***	0.012 (0.008)	0.002 (0.001)*	0.017 (0.005)***	0.010 (0.009)
d_PuertoCortes	0.001 (0.001)	0.015 (0.005)***	0.006 (0.006)	0.001 (0.001)	0.015 (0.006)***	0.000 (0.006)
d_Trujillo	0.002 (0.001)**	-0.022 (0.009)**	-0.002 (0.004)	0.005 (0.001)***	-0.003 (0.005)	0.001 (0.005)
Adj. R^2	0.37	0.43	0.68	0.43	0.40	0.68
N	803	401	806	853	426	856
Altonji Stat	1.44	0.89	2.47	.	.	.

Note: Regressions are estimated using weighted least squares, with the weights inversely proportional to the number of observations used to construct village means in columns 1, 2, 4, and 5. Regressions are weighted by total population in column 3 and 6. Robust standard errors, unclustered, are reported in parentheses. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes

Table 2.10: Village-Level Regressions: Flexible Distances

	Wages		**Rents**		**Density**	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
rr50_cent_edist	-0.006		-0.022		-0.012	
	(0.002)***		(0.010)**		(0.010)	
rr34_cent_edist	-0.001	-0.004	0.038	0.034	0.012	0.010
	(0.002)	(0.002)**	(0.013)***	(0.010)***	(0.013)	(0.009)
rr50_0510		-0.207		-0.108		-1.374
		(0.041)***		(0.244)		(0.183)***
rr50_1015		-0.213		-0.468		-1.039
		(0.045)***		(0.166)***		(0.225)***
rr50_1520		-0.221		-0.643		-0.529
		(0.049)***		(0.189)***		(0.312)*
rr50_2025		-0.188		-0.717		-1.048
		(0.058)***		(0.268)***		(0.260)***
rr50_gt25		-0.176		-0.929		-0.667
		(0.062)***		(0.331)***		(0.301)**
Adj. R^2	0.43	0.46	0.40	0.44	0.68	0.72
N	853	853	426	423	856	853

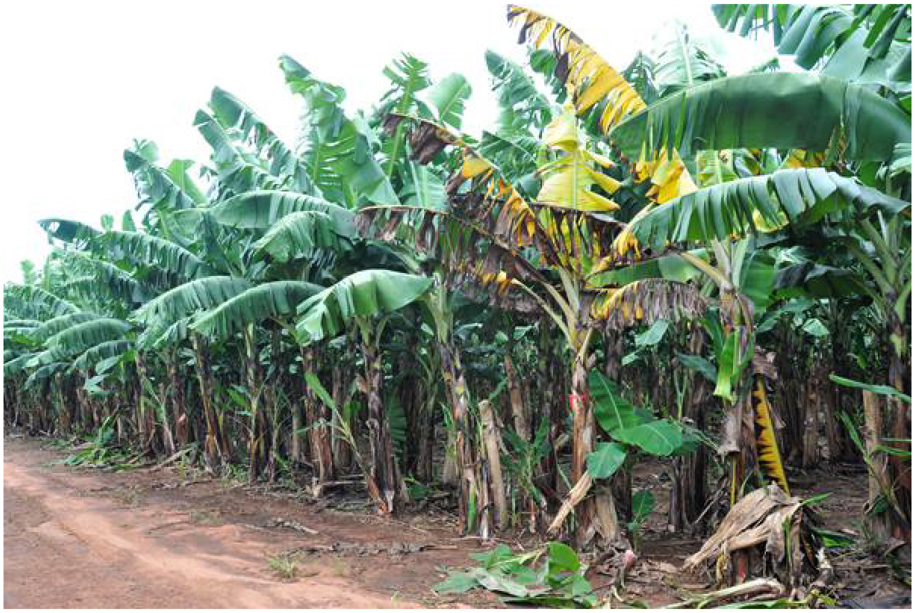
Note: Regressions are estimated using weighted least squares, with the weights inversely proportional to the number of observations used to construct village means in columns 1, 2, 4, and 5. Regressions are weighted by total population in column 3 and 6. Robust standard errors, unclustered, are reported in parentheses. * denotes significant at the 10% level, ** denotes significant at the 5% level, and *** denotes significant at the 1% level. Unit of observation is a village.

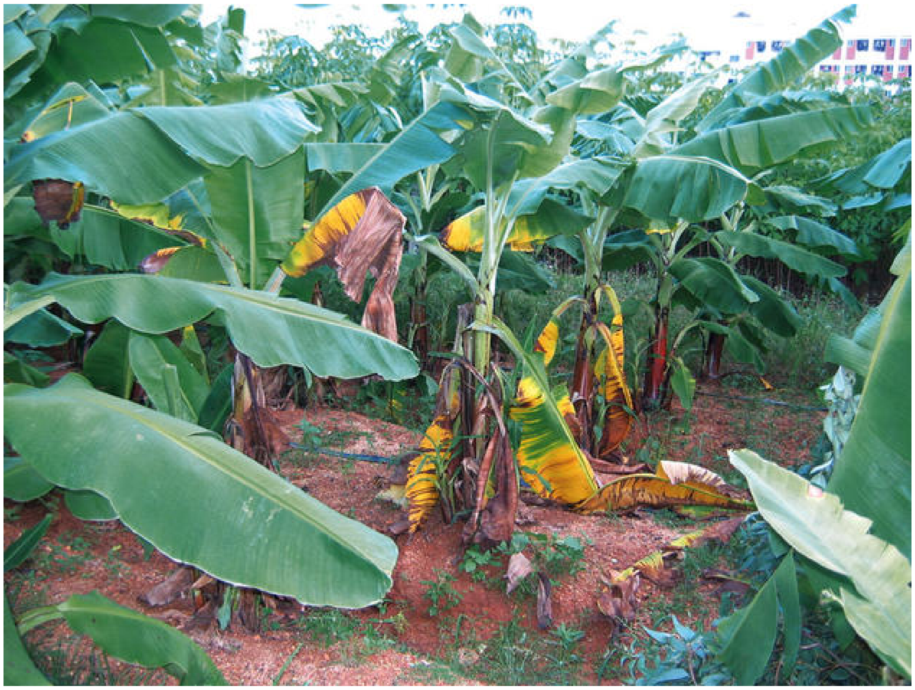
Table 2.11: Implicit Prices and Production Costs of Non-Abandonment

	Implicit Price	**Implicit Cost Savings**
	(1)	**(2)**
1. Rent Contribution	**602.58**	**0.14**
2. Estimate of $\partial \log r_c / \partial A$	0.35	0.35
3. Avg. Housing Expenditure ($r_c H_c$)	1743.95	.
4. Output Share of Total Housing Exp. ($\sum r_c H_c / X$)	.	0.39
	.	.
5. Wage Contribution	**-1144.54**	**0.08**
6. Estimate of $\partial \log w_c / \partial A$	-0.15	0.15
7. Avg. Household Wage (w_c)	7500.40	.
8. Output Share of Total Wages ($\sum w_c N_c / X$)	.	0.53
	.	.
9. Total	**-541.96**	**0.22**
	(399.18)	(0.07)***

Note: Authors' calculations. The parameter estimate reported in row 2 is taken from Table 2.9, column 2, while the parameter estimate reported in row 6 is taken from Table 2.9, column 1. The figures for rows 4 and 8 come from the United Nations (2000) *National Account Statistics*.

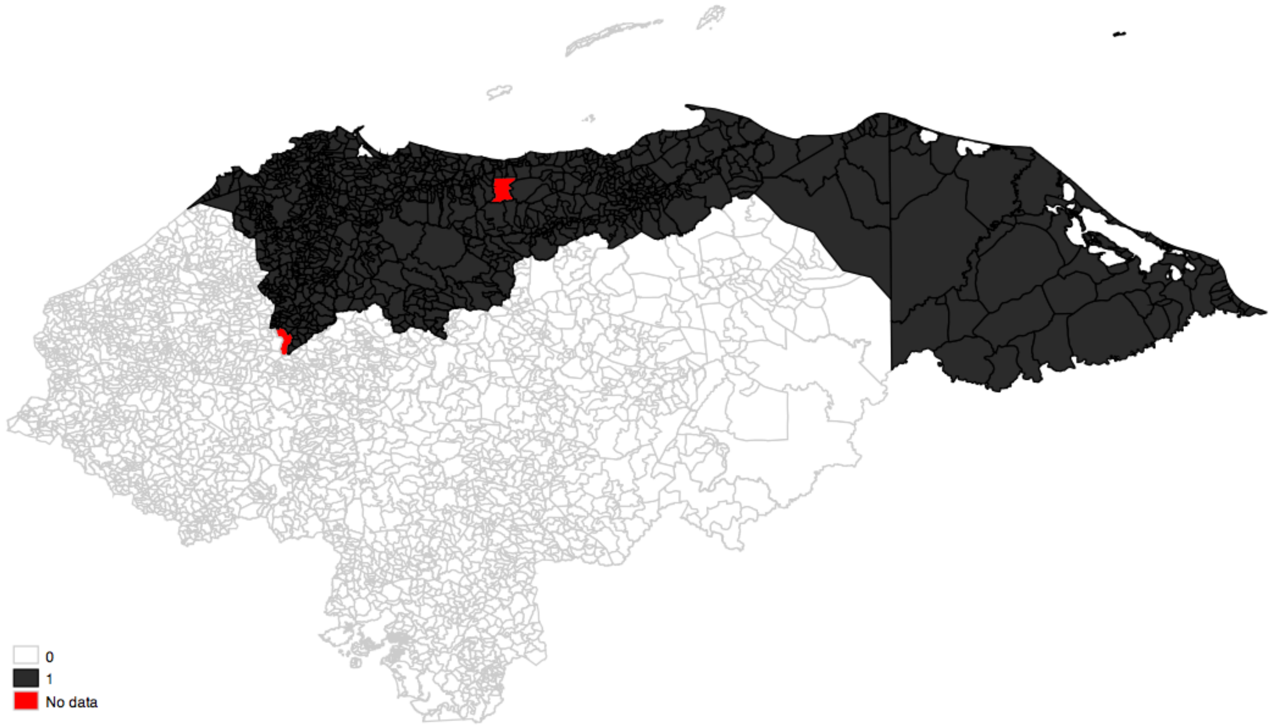
Figure 2.1: Panama Disease





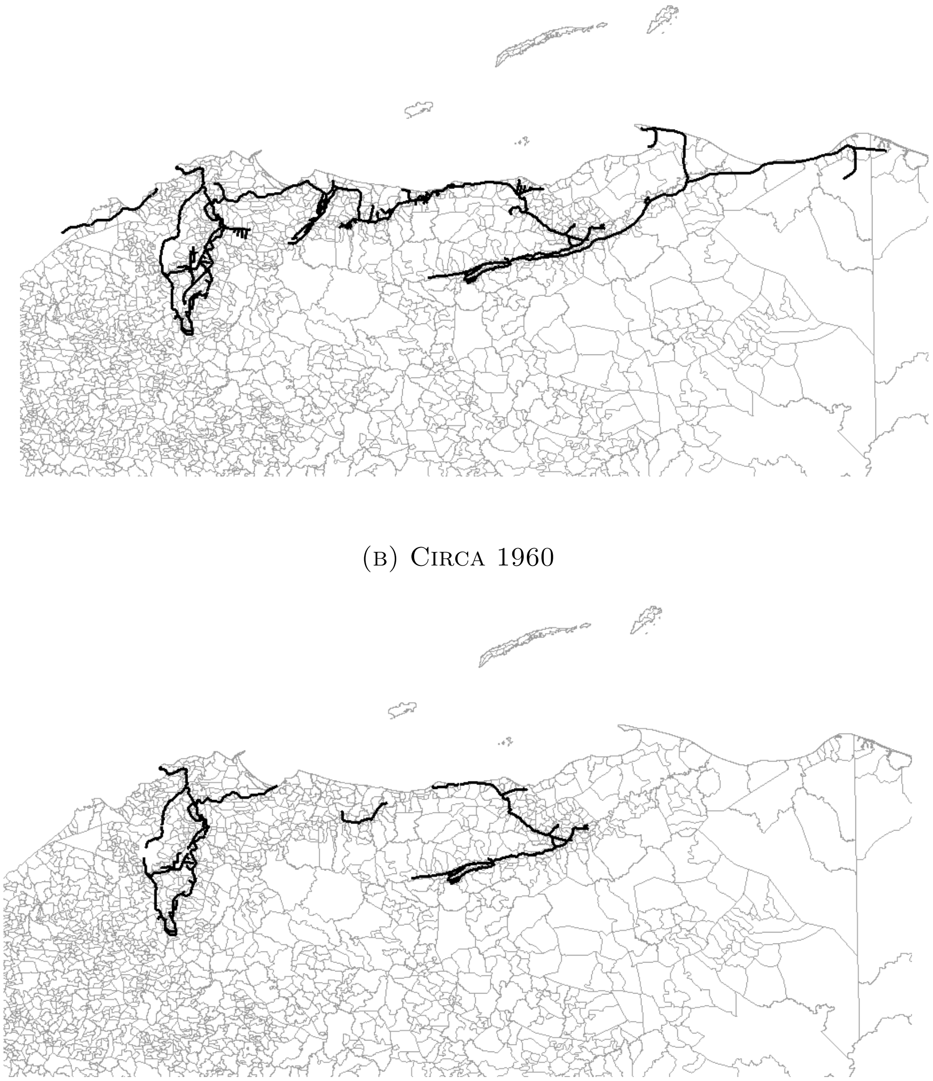
Note: Banana plants with panama disease stand among healthy plants. The yellowing and wilting of the leaves is a characteristic sign of infection. Sources: http://www.abc.net.au/rural/news/content/201109/s3327963.htm and http://uwire.com/wp-content/uploads/2011/09/Banana-Plants.jpg.

Figure 2.2: North Coast of Honduras



Note: North coast aldeas are depicted in black.

Figure 2.3: Honduras's Railroad Network

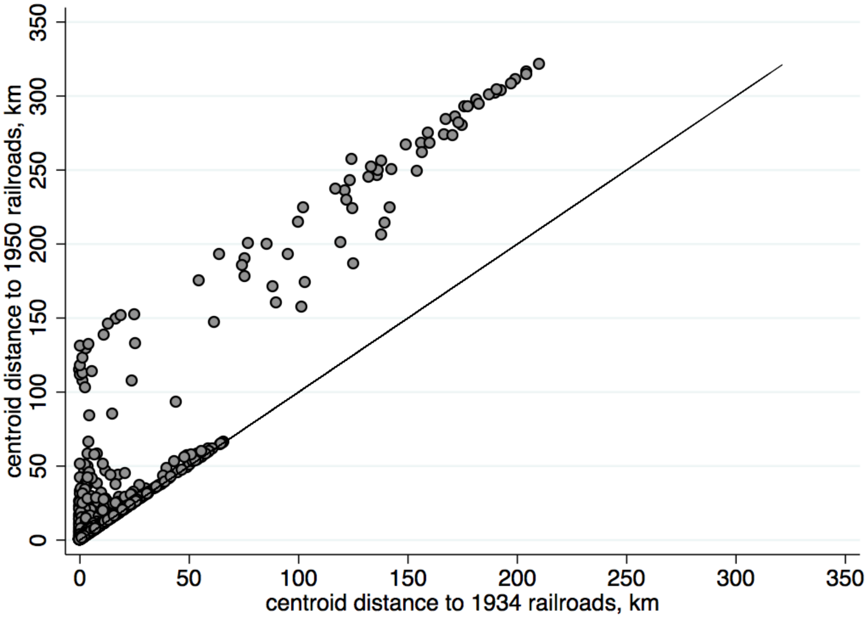


(a) Circa 1934

(b) Circa 1960

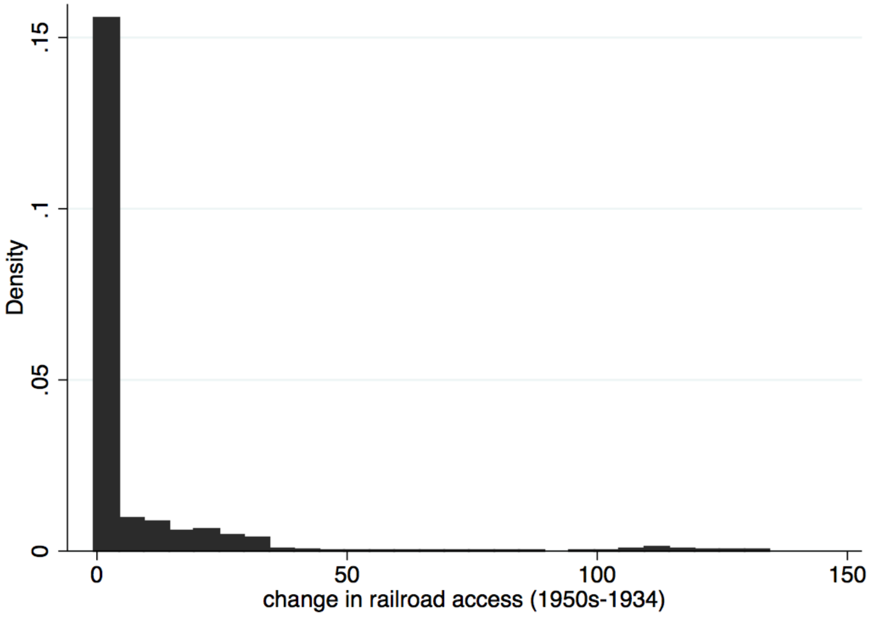
Note: Digitized network traced from fruit company maps and from U.S. Army maps. See Appendix 2.A for more details.

Figure 2.4: Scatterplot of Distance to Railroads



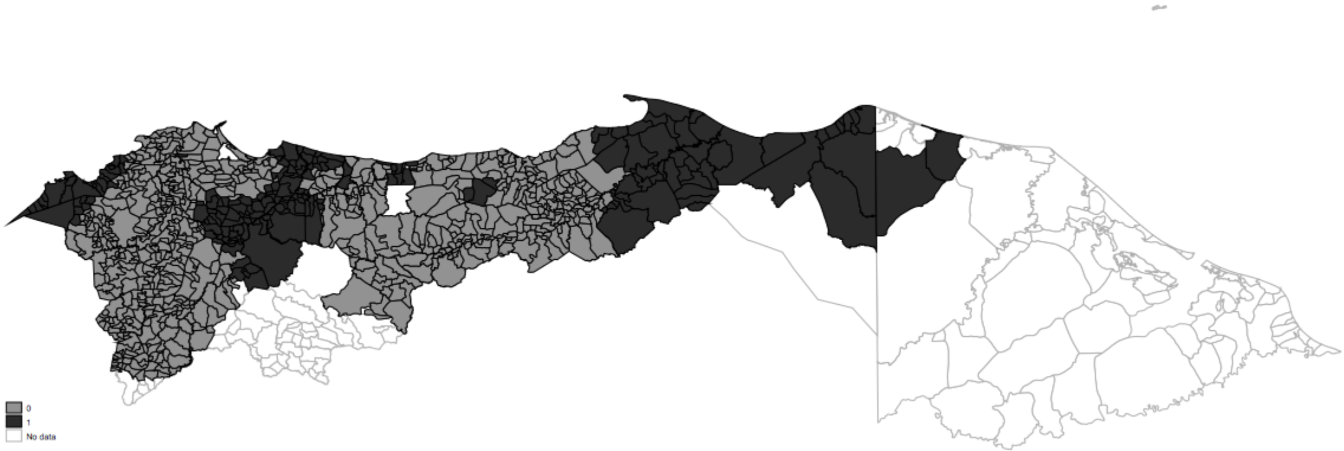
Note: black line is a 45 degree line.

Figure 2.5: Histogram Change in Distance to Railroads



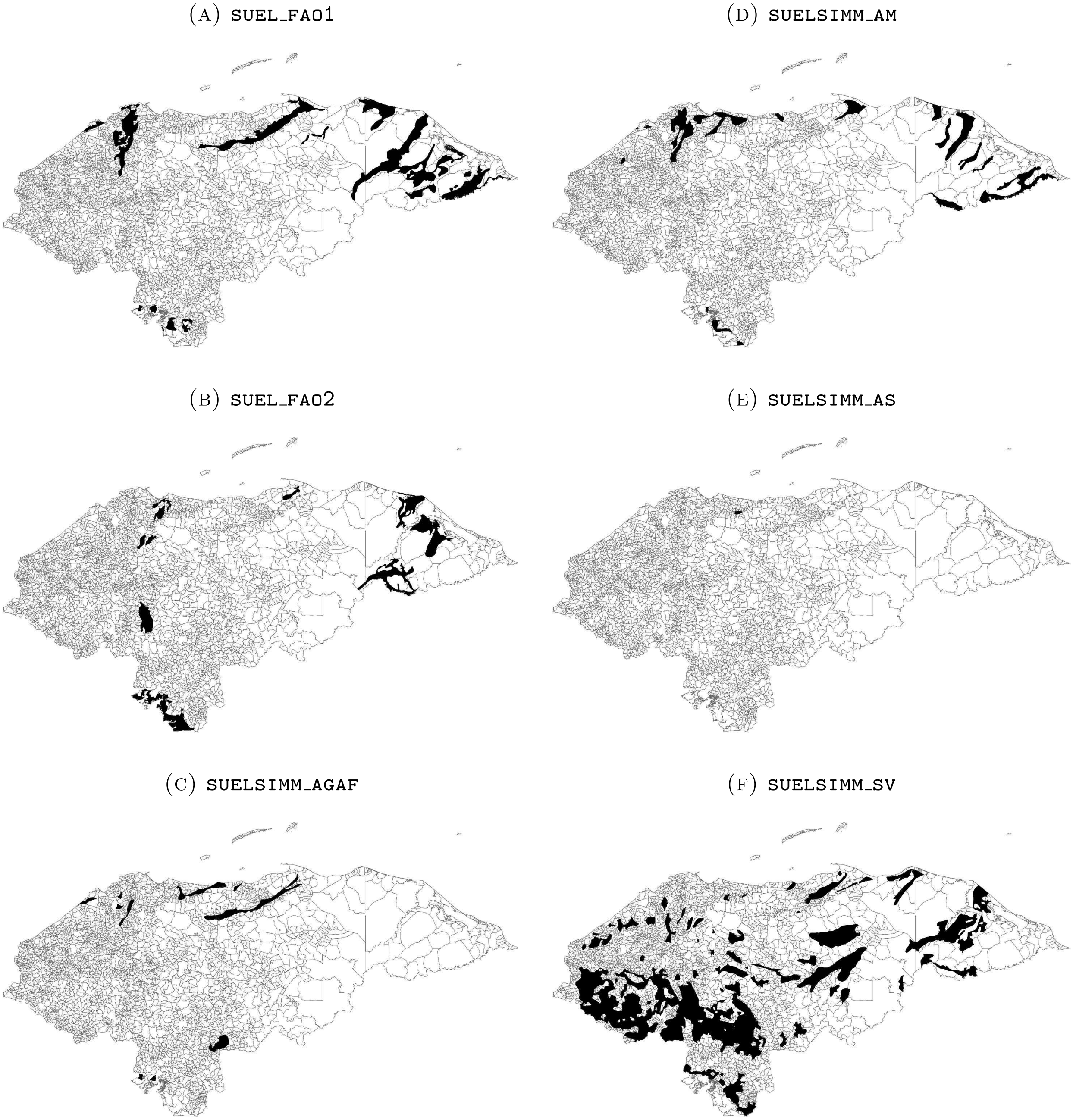
Note: Each bin is set to a width of 5 kilometers.

Figure 2.6: Abandoned and Control Villages



Note: Areas in black correspond to villages with the abandonment indicator, A_c^{30}, equal to 1, while grey areas correspond to villages where the indicator is equal to 0. Areas in white are where the indicator is censored, or undefined.

Figure 2.7: Maps of Soil Quality Variables



(a) SUEL_FAO1

(b) SUEL_FAO2

(c) SUELSIMM_AGAF

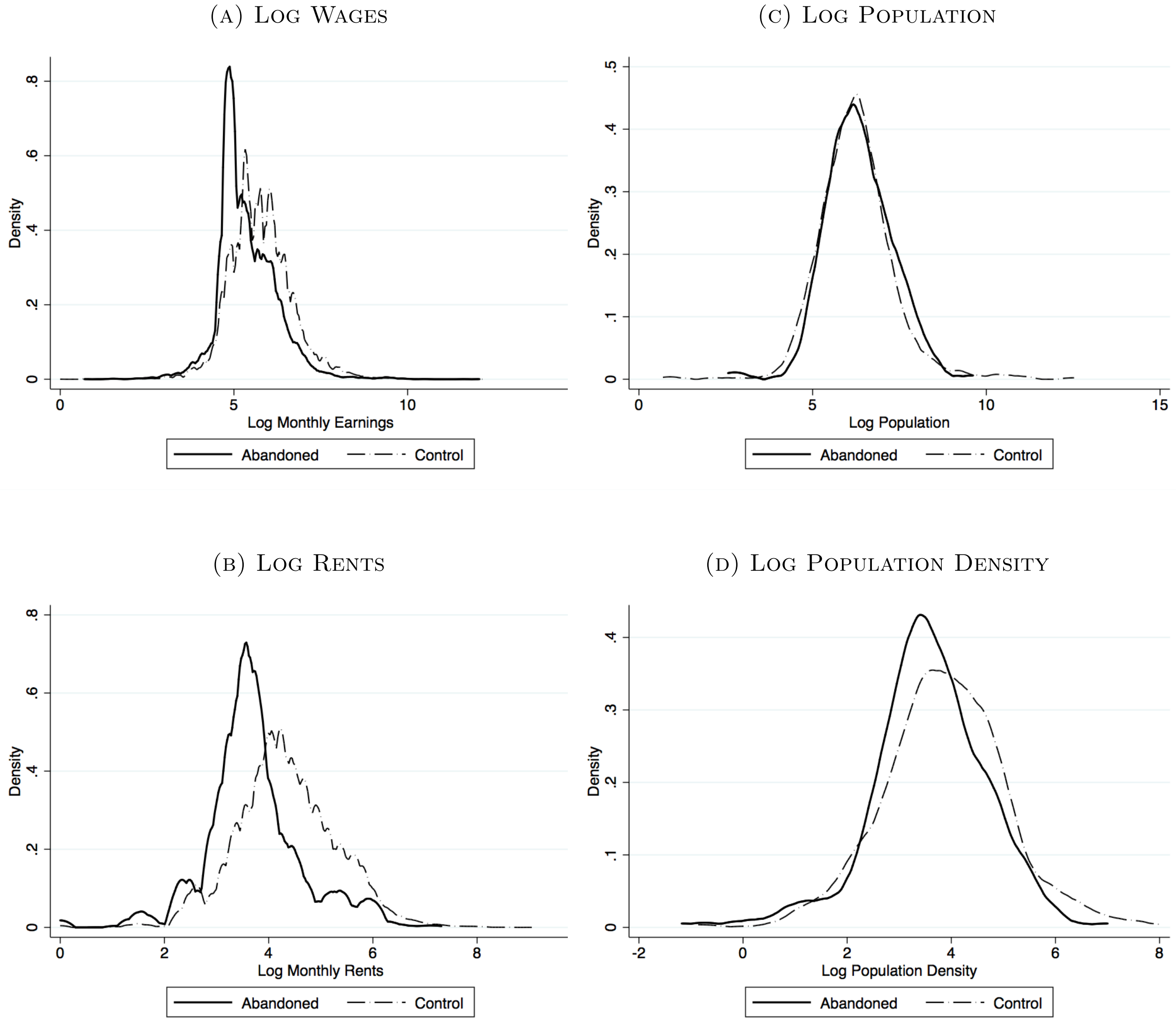
(d) SUELSIMM_AM

(e) SUELSIMM_AS

(f) SUELSIMM_SV

Spatial indicator variable: black areas correspond to areas where the raster is equal to 1, while white areas correspond to 0.

Figure 2.8: Kernel Density Plots of Wage and Rent Distribution: Abandoned and Control Aldeas



(a) Log Wages

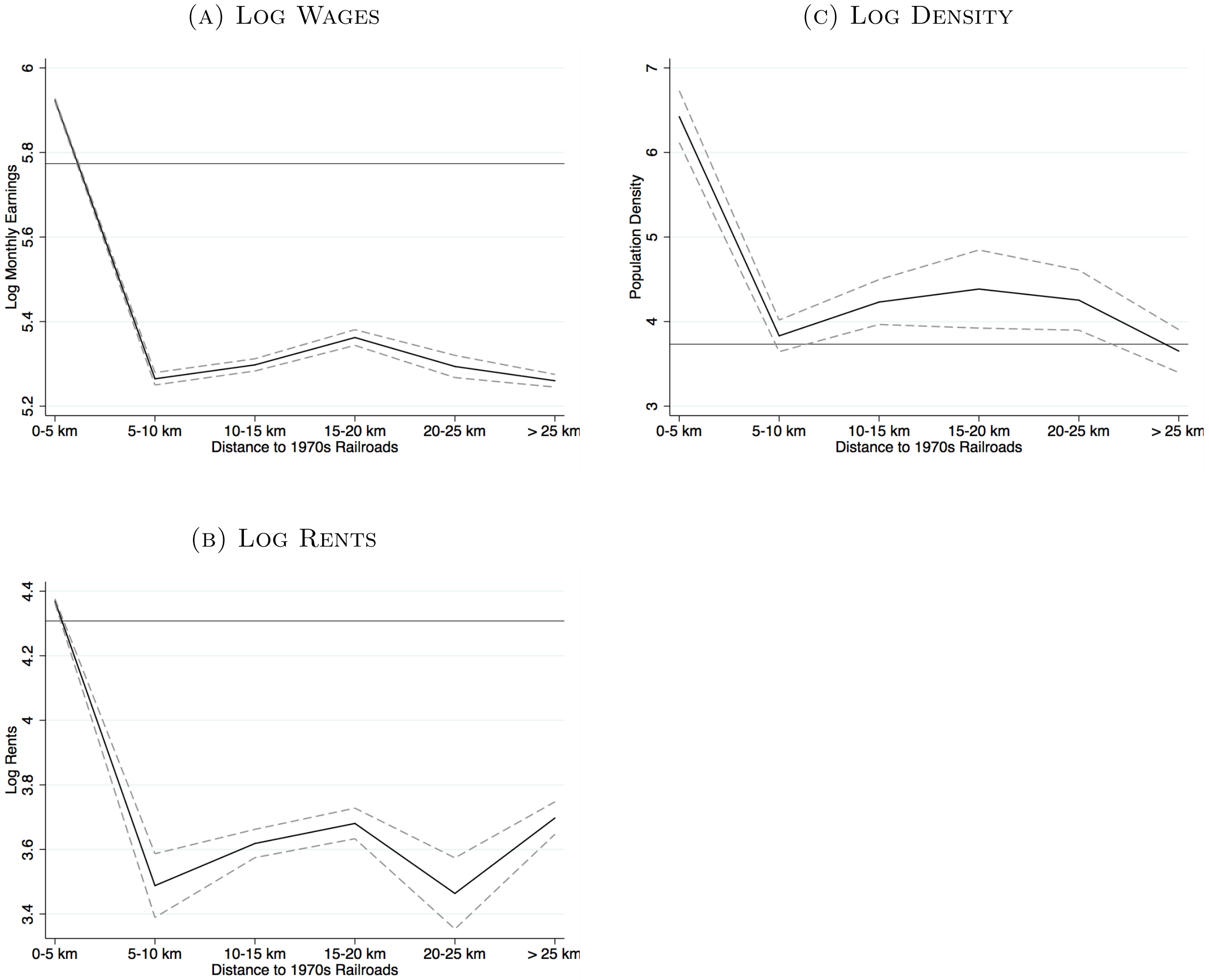
(c) Log Population

(b) Log Rents

(d) Log Population Density

Note: Kernel density estimates use a Gaussian kernel and bandwidths chosen to minimize an integrated mean square error criterion.

Figure 2.9: Predicted Marginal Effects of Distance to Railroads


(a) Log Wages

(c) Log Density

(b) Log Rents

Note: These figures report the marginal effects of being different distances away from railroads, after averaging over the other covariates in the regressions. These marginal effects are taken from regressions reported in Table 2.10. Solid horizontal lines show the mean of the variables on the y-axis.

2.A Data Appendix

2.A.1 1887 Census

Data on the population of Honduras before the intervention of the banana companies was compiled and digitized from the *Censo General de la Republica de Honduras, Levantado el 15 de Junio de 1887.* These data contain information at the municipality level on population totals and the number of people working in different occupations, and they also contain information at the aldea level on the total number of people and households. The data were taken from scanned from physical hard-copies of the census reports, and a data entry firm assisted with digitization. The data entry work was double checked by comparing printed totals with actual totals printed in the documents.

We still haven't completely folded these data into our analysis, because of the difficulty of merging in these data at the aldea level. A few notes:

- Some documentation on merging 1887 village names to 2001 village names would probably be useful, once this has been completed (Bryan ?).
- The following pages of the census hard-copy need to be revisited:
  - Page 37 - cut off at the end.
  - Page 38 - completely missing.
  - Page 116 - cut off at the end.
- The aldea-level dataset in STATA needs a bit of work. Various rows and columns for aggregating totals need to be eliminated; they were checked originally as a way of verifying the accuracy of the data entry, but this is no longer necessary.

2.A.2 1988 Household Census Data

Unit-level census data were taken from the *Censo Nacional de Población, 1988.* We have access to the universe of long form data, which includes both a housing module as well as an individual wage module. There are 891,394 household observations, each with information on housing type and construction materials, ownership type, water supply, assets, among other variables. There are 4,263,912 individual observations, and the long form data contains questions related to individual education levels, occupations, family composition, age, mortality, and migration information. In total, 42 variables were recorded for each household in the data, and 49 variables were recorded for each individual.

- It looks like we can merge the household data to the individual data. This might enable us to estimate hedonic relationships using SUR, and possibly enable us to exploit cross-equation covariance for efficiency gains.
- There are several problems with merging the 1988 aldea codes to the 2001 aldea codes. We're currently trying to merge these data using the codes from 1988 and the 2001 shapefile codes, but this is likely problematic.
  - There were 172 codes appearing in the 1988 census that could not be matched to the 2001 census.
  - There were 121 aldea codes appearing in the 2001 census that did not appear in the 1988 census.
  - This mismatch results in a loss of about 63,000 observations from the individual-level wage data, or 1.5 percent of the sample.

2.A.3 1952 Agricultural Census Data

We have scanned the 1952 Agricultural Census books into PDF format and had a data entry firm hard code the data. Have not started working with these files.

2.A.4 2001 Household Census Data

Although we do not have access to unit-level data from the *Censo Nacional de Población, 2001*, we were able to construct several variables at the aldea level. These include population totals, household counts, average household sizes, ethnicity information, and the total number of workers in each industry and occupation. The aldea-level variables were extracted using the `Redatam+` program, and the data files we use were provided by CIAT and *Instituto Nacional de Estadística*.

2.A.5 Administrative Boundary Shapefile

There are a few (known) problems with the `aldea.shp` file:

1. Two distinct village polygons are assigned the same code and name. This should not be much of an issue for our analysis, especially because this aldea code is not referenced in the ENCOVI datasets, but it would be nice to look into better shapefiles at some point. The details are below:
   - Name = "La Caoba"
   - Geocodigo = "050620"

   Images of the problem shapes appear in Figure 2.10 and Figure 2.11.

2. There were also some issues regarding the correspondence between aldea codes stored in the shapefile and those stored in the 2001 census data. A total of 79 aldea codes were in the census but not in the shapefile, while 83 codes were in the shapefile but not the census.

   Using the names of the villages, stored in both the shapefile and the census, I could resolve by hand a good chunk of these problems. The code for fixing these aldea codes is stored in `$do_files/gis_processing_YYMMDD.do`. After fixing the codes, we still have 39 aldea codes in the census but not in the shapefile, and 43 codes in the shapefile but not the census. This needs to be resolved at some point, although in terms of the ENCOVI sample, we only lose 2 villages of the 379. These villages are:

```
. list aldea_code aldea_name if ac_in_encovi_not_shape == 1;
      +------------------------+
      | aldea~de    aldea_name |
      |------------------------|
2586. |   132313   Las Lajitas |
2587. |   132314     San Ramn  |
      +------------------------+
```

   Note that these villages aren't on the North Coast, so will not cause problems for that subsample of the data.

3. I assume that this shapefile was prepared for the 2001 census? It would be good to have the proper citation on some of these files.

2.A.6 Infrastructure Maps: Projection Information

In order to create variables coding the distance from each aldea to the nearest railroad, we first projected and trace maps digitally, using ArcView. For a guide to make sure the maps were overlaid correctly, I used the `ALDEAS.shp` file, which was originally stored using the `GCS_WGS_1984` ellipsoid and `D_WGS_1984` datum. Because all of the maps I worked with used the 1000 Meter Universal Transverse Mercator (UTM) Zone 16 projection, Clarke 1866 Ellipsoid, and 1927 N. American Datum, I needed to project the original shapefile twice before I could start working with it.[24] The projection information was as follows:

`GCS_WGS_1984`	to	`GCS_North_America_1927`
`GCS_North_America_1927`	to	`NAD_1927_UTM_Zone_16N`

The first spatial transformation changes the datum of the original projection from `D_WGS_1984` to `D_North_America_1927`, without changing the projection. I used the standard transformation here. The second projection changes the ellipsoid, keeping the datum the same.

Documentation on the 21 maps I worked with can be found in Table 1. The maps were all added to the map in ArcView and spatial references were defined using the coordinates listed in Table 1 (all coordinates appear on the corners of these maps). When I succeeded in projecting my first map and creating spatial references for it, it seemed like there was a problem with the vertical alignment; a good chunk of La Cieba's coast seemed to be in the ocean. I resolved this issue with an "eyeball fix" which amounted to reducing all of the longitude coordinates (on each map) by 10 seconds (approx 300 meters). This seems to have worked pretty well.

After I finished projecting the maps and checked their spatial references, I looked them over to make sure they were aligned properly. The `caser2004shp.shp` file had information on the names of small settlements and villages, and I cross-referenced these names with the names appearing on the maps.

After I was satisfied with the alignment of the maps, I traced the railroads that appeared on the maps using ArcView's Edit toolbox. The resulting shapefile is called `Railroad_1950_UTM.shp`. Some maps contained information on "abandoned" railroads, and these railroads were stored in `Railroad_1950_Abandon_UTM.shp`. I used these files as a guide when merging the railroad shapefile from the 1930s given to me by Bryan.

After aligning the 1934 railroads with the abandoned 1950s railroads and the general contours of the railroads that remained in use, I cross-checked the placement of the 1934 railroads with some other maps from the time period. These other maps are available from the Library of Congress.[25]

To create distance measures, I used ArcView's "Straight-line Distance" command in the Spatial Analyst toolbar, which constructs a raster file (a grid) of pixels and assigns values to each of those pixels based on the straight-line distance from that pixel to the railroad polyline. The distance rasters for both the 1930s and 1950s railroad shapefiles were constructed at a resolution of 100m × 100m =

[24] The maps all had vertical datum information (Mean Sea Level) but since I wasn't doing anything with the topography, I ignored this.

[25] Bryan, some documentation on these?

1 pixel[26]. I ran `Zonal Statistics as Table` on each of these rasters, which resulted in the mean, median, min, and max distance variables you see in Table 2.

I also constructed centroids for each of the Aldeas polygons, following the instructions here: `www-laep.ced.berkeley.edu/classes/tool_time/addXYcentroid/addXYcentroid.html`. Using the same distance rasters, constructed a centroid distance measure (this will be less accurate than a vector-distance measure using ArcInfo's "Near" command, because it assigns the distance to whichever pixel the centroid point falls on), but I think it's a reasonable first pass. I also made no distinction between branch lines and main lines in any of these distance calculations; I figured this would be the kind of thing to do at a later point.

2.A.7 Infrastructure Maps, 2001

To create distance measures, I used ArcView's "Straight-line Distance" command in the Spatial Analyst toolbar, which constructs a raster file (a grid) of pixels and assigns values to each of those pixels based on the straight-line distance from that pixel to the railroad polyline. The distance rasters for both the 1930s and 1950s railroad shapefiles were constructed at a resolution of 100m × 100m = 1 pixel[27]. I ran `Zonal Statistics as Table` on each of these rasters, which resulted in the mean, median, min, and max distance variables you see in Table 2.

I also constructed centroids for each of the Aldeas polygons, following the instructions here: `www-laep.ced.berkeley.edu/classes/tool_time/addXYcentroid/addXYcentroid.html`. Using the same distance rasters, constructed a centroid distance measure (this will be less accurate than a vector-distance measure using ArcInfo's "Near" command, because it assigns the distance to whichever pixel the centroid point falls on), but I think it's a reasonable first pass. I also made no distinction between branch lines and main lines in any of these distance calculations; I figured this would be the kind of thing to do at a later point.

2.B ENCOVI Data

The National Household Survey on Living Conditions (ENCOVI) was conducted by Instituto Nacional de Estadística (INE) between July 31 and November 30, 2004. It incorporated both rural and urban areas of the country's 18 departments.

The ENCOVI is a multipurpose research to know the different aspects and dimensions of household welfare. Includes, in addition to income and expenditure of households, a set of variables describing the living standards of households. In this sense this publication includes information on housing characteristics, demographics, migration, education, health, anthropometry, the labor market (gender, people with work, child labor and juvenile), income and household expenditure, poverty and other important issues.

– Multiple Purpose Permanent Household Survey (EPHPM): Bryan, would this be useful?

[26] I tried to go down to 50mx50m = 1 pixel and less, but the program kept crashing on me.

[27] I tried to go down to 50mx50m = 1 pixel and less, but the program kept crashing on me.

2.C Ruggedness

A 30 arc-second ruggedness raster was computed for Honduras according to the methodology described by Sappington et al. (2007). The authors propose a Vector Ruggedness Measure (VRM), which captures the distance or dispersion between a vector orthogonal to a topographical plane and the orthogonal vectors in a neighborhood of surrounding elevation planes. To calculate the measure, one first calculates the x, y, and z coordinates of vectors that are orthogonal to each 30-arc second grid of the Earths surface. These coordinates are computed using a digital elevation model and standard trigonometric techniques. Given this, a resultant vector is computed by adding a given cells vector to each of the vectors in the surrounding cells; the neighborhood or window is supplied by the researcher. Finally, the magnitude of this resultant vector is divided by the size of the cell window and subtracted from 1. This results in a dimensionless number that ranges from 0 (least rugged) to 1 (most rugged).[28] For example: on a (3×3) flat surface, all orthogonal vectors point straight up, and each vector can be represented by (0,0,1) in the Cartesian coordinate system. The resultant vector obtained from adding all vectors is equal to $(0, 0, 9)$, and the VRM is equal to $1(9/9) = 0$. As the (3×3) surface deviates from a perfect plane, the length of the resultant vector gets smaller, and the VRM increases to 1.

We computed ruggedness measures using multiple neighborhoods: a (3×3) window (corresponding to a 3 square kilometer window), a (5×5) window (corresponding to a 5 square kilometer window), and a (9×9) window (corresponding to a 9 square kilometer window).

– Calculation needs to be redone because of issues with the border areas in the shapefile. Right now, the elevation of those areas is undefined, and this is screwing up the GIS variable construction for many villages. It should be set to zero.

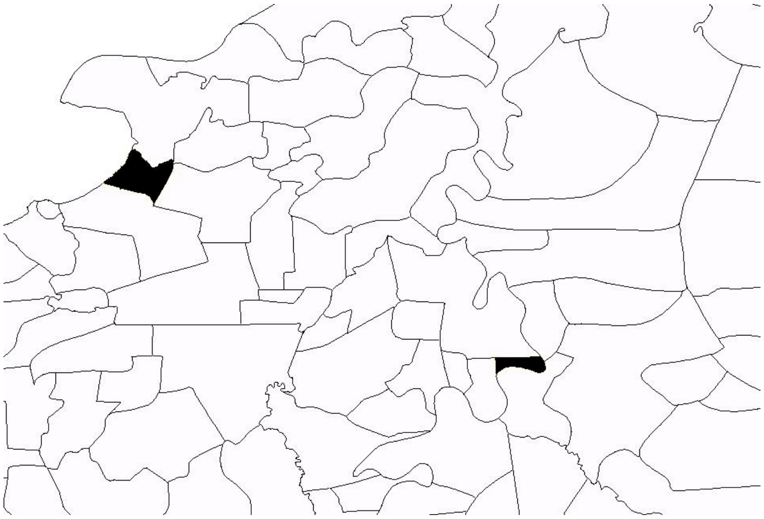
[28] The authors have generously provided a Python script for computing their Vector Ruggedness Measure (VRM) in ArcView. The script and detailed instructions for installation can be found here: `http://arcscripts.esri.com/details.asp?dbid=15423`.

Table 2.12: Detailed Documentation for Railroad Maps

			Upper Left Corner		Lower Right Corner			
	Serial No.	**Date**	**Lat**	**Lon**	**Lat**	**Lon**	**RMSE**	**Prepared By**
Livingston	D-16-N-I	1940	16° N	90° W	15° N	88° W	.	GBMID
La Ceiba	D-16-N-II	1934	16° N	88° W	15° N	86° W	.	GBMID
Trujillo	D-16-N-III	1934	16° N	86° W	15° N	84° W	.	GBMID
Conception de Norte	2561-I	1977	15°20′ N	88°15′ W	15°10′ N	88°00′ W	3.463	IGN, DMA, IAGS
Cuyamel - SPS	2562-I	1992	15°40′ N	88°15′ W	15°30′ N	88°00′ W	2.757	IGN, NIMA
Valle de Naco	2562-II	1988	15°30′ N	88°15′ W	15°20′ N	88°00′ W	30.059	IGN
Rio Lindo	2661-III	1980	15°10′ N	88°00′ W	15°00′ N	87°45′ W	5.822	IGN
Villanueva	2661-IV	1990	15°20′ N	88°00′ W	15°10′ N	87°45′ W	5.191	IGN
El Progresso	2662-III	1987	15°30′ N	88°00′ W	15°20′ N	87°45′ W	2.402	IGN
Choloma	2662-IV	1994	15°40′ N	88°00′ W	15°30′ N	87°45′ W	2.101	IGN, DMA
Laguna de Los Micos	2663-II	1992	15°50′ N	87°45′ W	15°40′ N	87°30′ W	3.347	IGN, DMA
Baracoa	2663-III	1993	15°50′ N	88°00′ W	15°40′ N	87°45′ W	8.017	IGN, DMA
Puerto Cortes	2663-IV	1992	15°60′ N	88°00′ W	15°50′ N	87°45′ W	9.010	IGN, DMA
La Masica	2672-I	1989	15°40′ N	87°15′ W	15°30′ N	87°00′ W	6.138	IGN
San Francisco	2673-II	1992	15°50′ N	87°15′ W	15°40′ N	87°00′ W	4.188	IGN, DMA
Tela	2763-III	1992	15°50′ N	87°30′ W	15°40′ N	87°15′ W	9.402	IGN, DMA
Olanchito	2862-II	1970	15°30′ N	86°45′ W	15°20′ N	86°30′ W	4.891	IGN, USTC
Arenal	2862-III	1970	15°30′ N	87°00′ W	15°20′ N	86°45′ W	6.734	IGN
Jutiapa	2863-II	1994	15°50′ N	86°45′ W	15°40′ N	86°30′ W	5.843	IGN, DMA
La Ceiba	2863-III	1993	15°50′ N	87°00′ W	15°40′ N	86°45′ W	8.051	IGN, DMA
Saba Tocoa	2962-I	1991	15°40′ N	86°15′ W	15°30′ N	86°00′ W	2.001	IGN
Confl. Rios Aguan-Mame	2962-III	1989	15°30′ N	86°30′ W	15°20′ N	86°15′ W	16.011	IGN
Sonaguera	2962-IV	1989	15°40′ N	86°30′ W	15°30′ N	86°15′ W	3.966	IGN
Balfate	2963-III	1992	15°50′ N	86°30′ W	15°40′ N	86°15′ W	2.634	IGN, DMA

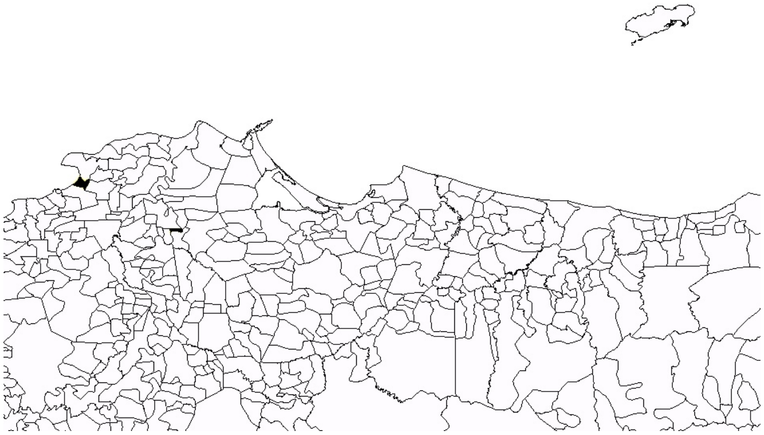
1934 Maps have scale 1:250,000; 1960s Maps have scale 1:50,000. Date column lists the date of the most recently incorporated data. RMSE column lists the root mean square error of the map projection, calculations taken from ArcView (units are squared meters). GBMID stands for Geographic Branch, Military Intelligence Division. IGN stands for the Honduran Instituto Geografico Nacional. IAGS stands for the Inter-American Geodetic Survey. DMA stands for the U.S. Defense Mapping Agency. USTC stands for the U.S. Army Topographic Command.

Figure 2.10: Duplicate Aldea: "La Caoba - 050620"



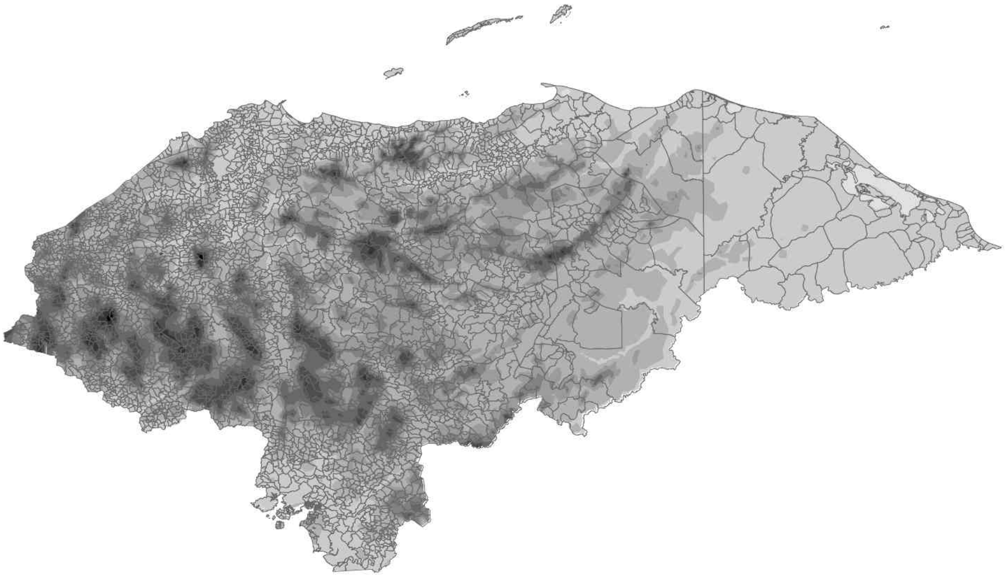
Source: `aldea.shp` file

Figure 2.11: Duplicate Aldea: "La Caoba - 050620", Map 2



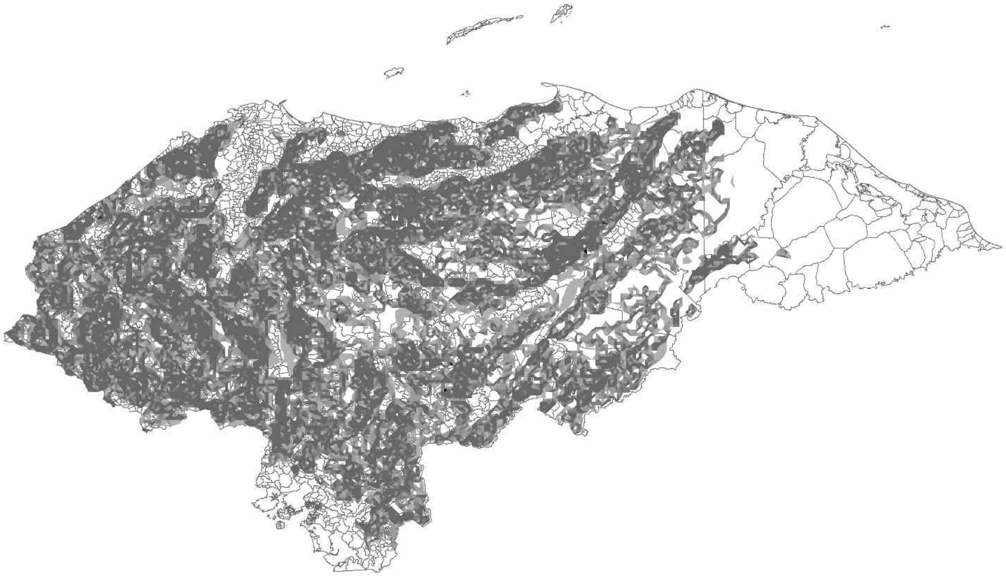
Source: `aldea.shp` file

Figure 2.12: Map of Elevation



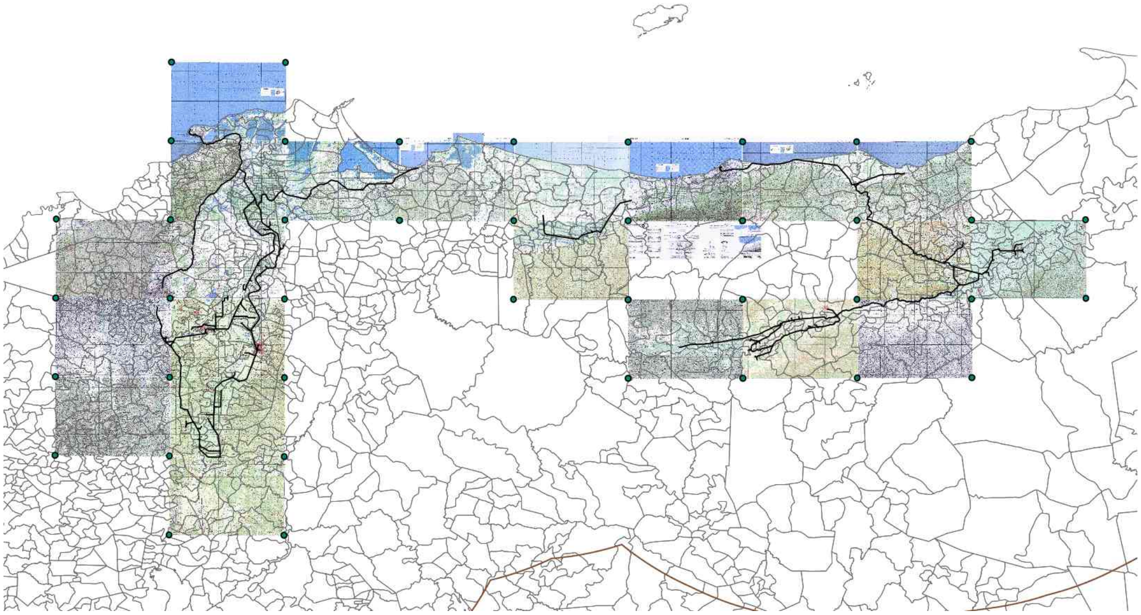
Raster: darker areas correspond to higher elevation.

Figure 2.13: Map of Slope



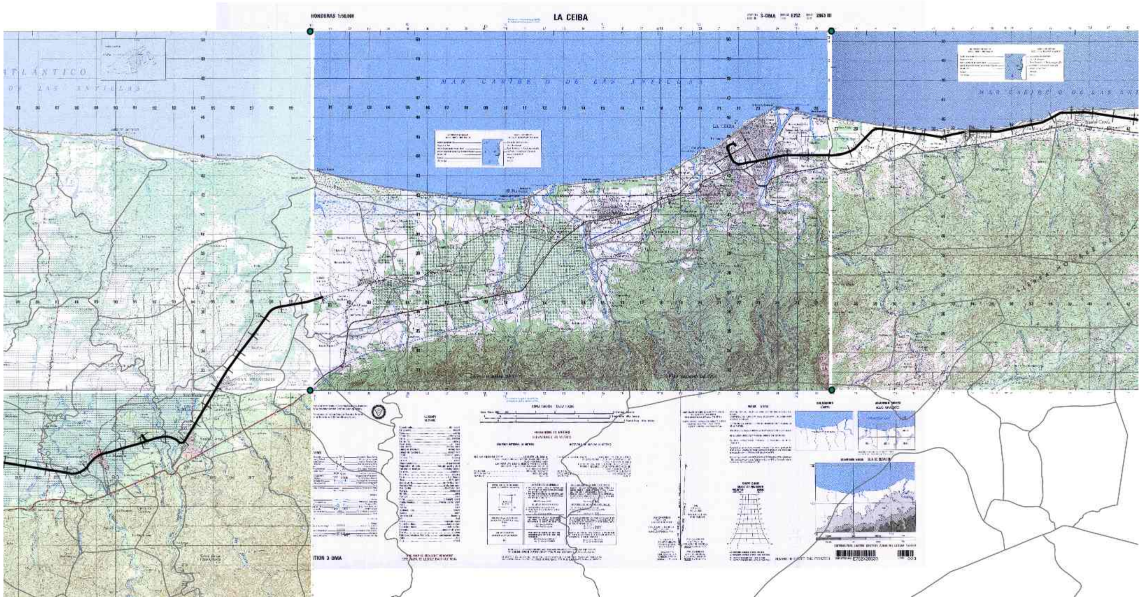
Raster: darker areas correspond to larger gradient.

Figure 2.14: Projection of 1950s Maps



See Table 1 for more information.

Figure 2.15: Projection of 1950s Maps - Detail of La Ceiba



See Table 1 for more information.

Chapter 3

Collective Action in Diverse Sierra Leone Communities

with Edward Miguel and Rachel Glennerster

Abstract

Scholars have pointed to ethnic divisions as a leading cause of underdevelopment, due in part to their adverse effects on public goods. We investigate this issue in post-war Sierra Leone, one of the world's poorest countries. To address concerns over endogenous local ethnic composition, we use an instrumental variables strategy relying on historical census data on ethnic composition. We find that local diversity is not associated with worse public goods provision across a variety of outcomes, specifications, and diversity measures, with precisely estimated zeros. We investigate the role that leading mechanisms proposed in the literature play in generating the findings.

3.A Introduction

Many scholars have argued that ethnic diversity is an important impediment to economic and political development. Economic growth rates are slower in ethnically diverse societies, and local public goods provision often suffers (Easterly and Levine, 1997; Alesina et al., 1999, 2003). The leading explanation for why diversity affects outcomes in less developed countries is the inability to overcome the public good free-rider problem, due to monitoring and enforcement limitations (Miguel and Gugerty, 2005; Habyarimana et al., 2007, 2009). These issues are particularly salient in sub-Saharan Africa, the world's most ethno- linguistically diverse region.

This paper examines the relationship between ethnic diversity and local collective action, public goods, and social capital outcomes in post-war Sierra Leone, using new datasets designed for this purpose. Sierra Leone is among the world's poorest and most ethnically diverse countries, and is recovering from a decade of civil war that displaced millions and caused untold human suffering. Ethnic appeals and divides are salient in national politics in Sierra Leone, making it a reasonable setting to test the thesis that ethnic divisions stifle local public service delivery and economic development.

Both here and in other studies, the endogenous residential sorting of individuals complicates the reliable estimation of ethnic diversity impacts, and a main contribution of this paper is the progress we make in addressing this issue. Recent sorting is likely to be particularly problematic in Sierra Leone, where many fled civil war violence. We first document that during and after the war, there was systematic movement of individuals towards areas where their own ethnic group was historically more numerous. These preferences vary strongly as a function of individual characteristics, with, for example, education being associated with more residential movement to diverse areas. This finding underlines the possibility that correlations between ethnic diversity and local public goods outcomes might be biased.

In a methodological advance over most of the empirical ethnic diversity literature, we then use historical ethnic composition measures from the 1963 Sierra Leone Population Census as instrumental variables (IV) for current ethnic diversity to address the endogeneity problem created by migration. We find that in rural areas the historical ethnic diversity measures strongly predict current diversity, with a coefficient estimate of 0.8 in the first stage regression.

Using this IV approach, the paper's main finding is that local ethnic diversity is not associated with worse local public goods or collective action outcomes in Sierra Leone. This holds across a variety of regression specifications, measures of diversity, levels of aggregation, and outcomes that capture local collective action, including road maintenance, community group membership, trust, and school funding and staffing. Far more than in many developed countries, basic public goods are organized and produced locally in Sierra Leone, and many of these outcomes are very important for local economic development. For instance, road maintenance – the clearing of tropical brush that quickly engulfs dirt paths, as well as the construction of road drainage ditches and bridges – is a critical infrastructure investment in rural areas. Without it, trade and contact with the outside world becomes more expensive and less frequent, and in the extreme some villages would become isolated from their neighbors. We use a mean effects analysis to jointly consider the effect of diversity on groups of related outcomes (e.g., trust measures, school quality measures). We measure these zero impacts precisely, and thus with high levels of confidence can rule out that diversity has even moderate adverse impacts.

The IV approach would not be valid if there had already been systematic residential sorting correlated with local public goods quality by 1963, the year of the historical data. However, we document the absence of any historical correlation between ethnic diversity and socioeconomic measures (including literacy and formal employment), suggesting that little such sorting had taken place. Our results are also robust to excluding both urban areas and also areas in the country's east that benefited from the diamond boom of the 1950s, where pre-1963 sorting might have been an issue. We also employ several different measures of diversity, such as those based on language families and historical conflict, and we find no impacts with any of these measures.

These results quantify and reinforce claims by several scholars that, despite the leading role of ethnic appeals in national politics, ethnic divisions have been much less damaging in Sierra Leone than in many of its African neighbors, and in particular were not a leading factor in the recent 1991-2002 civil war. The Revolutionary United Front (RUF) rebels targeted people from all ethnic groups, and statistical analysis of documented human rights violations shows that no ethnic group was disproportionately victimized. There is also no evidence that civilian abuse was worse when armed factions and communities belonged to different ethnic groups (Humphries and Weinstein 2006). Ethnic grievances were not rallying cries during the war and all major fighting sides were explicitly multi-ethnic (Keen, 2005).

Beyond documenting the lack of a relationship between ethnic diversity and local public goods, we also discuss the institutional and historical factors that foster inter-ethnic cooperation. A leading explanation for Sierra Leone's relatively good inter-ethnic cooperation is the presence of strong traditional local authorities that help overcome the classic free-rider problem in local public goods provision. One persistent consequence of Britain's colonial system of decentralized despotism (Mamdani, 1996) in Sierra Leone was the empowerment of Paramount Chiefs, elected from and by tribal ruling families. Chiefs collect local taxes, royalties from diamond mining and logging, market fees, and they serve as the final arbiter in local courts. These Chiefs, who effectively have lifetime tenure, together with an entire hierarchy of village chiefs and village elders that they head, continue to dominate local politics, and have the authority to punish free-riders through fines, public embarrassment, and corporal punishment.[1]

Another important explanation for why ethnic diversity does not appear to undermine local collective action in Sierra Leone involves the historical interactions between different ethnic groups. At the time of the founding of the Sierra Leone colony in the late 18th century and through much of the 19th century, Krio (Creoles), former slaves who returned to Africa to settle Freetown, enjoyed a relatively privileged political and economic position due to their facility with English and special links with the British even though they were numerically small. Before independence, the key political division in Sierra Leone was Krio vs. non-Krio, but because of growing tensions between the Krio and up country ethnic groups, the British progressively limited their political power. After independence, the fact that the country's long-serving dictator Siaka Stevens belonged to a small ethnic group (Limba), rather than one of the country's two dominant groups (Mende and Temne), may have helped to further limit the politicization of ethnicity between the largest groups.

The Krio people gave Sierra Leone their language, also called Krio, which is a dialect of English that has been influenced by Portuguese, Arabic, Yoruba and many African languages

[1] Ostrom (1990) is seminal work on how communities overcome free-riding to achieve collective action.

as a legacy of the slave trade. Serving as a national lingua franca for decades, Krio is currently spoken (usually as a second language) by nearly all Sierra Leoneans, and is increasingly taught in schools. In many other African countries the lingua franca is the former colonial language, usually English or French. While Krio has a base in English, it is unique to Sierra Leone and widely spoken even by those with no schooling. While the existence of a common national language is clearly insufficient to guarantee social stability – as the African cases of Rwanda and Somalia poignantly illustrate – Krio's ubiquity in Sierra Leone may (through historical accident) help promote the consolidation of a common national identity that transcends tribe (wa Thiongo, 2009), as with Swahili in post-independence Tanzania.[2]

While ethnic diversity does not impede local collective action in Sierra Leone, and ethnic divisions did not feature prominently in the civil war, it would be wrong to conclude that ethnic identity is unimportant in contemporary Sierra Leonean society. Our migration findings show that Sierra Leoneans strongly prefer to move to areas where their own ethnic group is numerous, perhaps to benefit from ethnic job networks, informal insurance, or patronage from co-ethnic chiefs. Casey (2009) also finds that ethnicity remains salient in national politics. The two major political parties, SLPP and the APC (discussed below) have strong ethnic ties, the SLPP being connected to the Mende and other ethnic groups in the South and the APC to the Temne and other northern groups. To illustrate, in the 2007 Parliamentary elections, the APC won 36 of 39 seats in North while the SLPP (and a splinter party, PMDC) swept 24 of 25 seats in the South.

But there are limits to ethnic voting in Sierra Leone: while voters strongly prefer the party linked to their own group, Casey (2009) uses exit poll data to show that they are much more willing to cross ethnic-party lines in local elections, where they have better information about candidates. Moreover, the APC was able to win the 2007 national elections in part because the Mende splinter PMDC party aligned itself with the APC in the presidential runoff rather than their SLPP co-ethnics. Unlike in Tanzania, where nation-building reforms were accompanied by a dismantling of the entire system of chiefs, in Sierra Leone chieftancy institutions remain powerful. The continued prominence of tribal chiefs in Sierra Leone also arguably makes it more likely that ethnic divisions will at some point re-emerge.

[2] The sensitivity of the effect of ethnic diversity to local history, formal institutions, and social norms has been emphasized by other research on other African societies. For instance, Miguel (2004) finds no diversity impacts on local outcomes in Tanzania, a country whose leadership has consistently sought to bridge ethnic divisions by promoting a common language (Swahili) and abolishing traditional tribal chiefs, but does find adverse diversity impacts in neighboring Kenya, where post-independence leaders have exacerbated ethnic divisions for political gain. Posner (2004) examines two ethnic groups that straddle the Zambia-Malawi border, and finds that national political rivalry between them translates into worse local relations in Malawi, in contrast to Zambia, where they are not on opposing political sides. In a recent contribution, Dunning and Harrison (2010) argue that cross-cutting "joking cousinage" institutions limit ethnic salience in Mali. These cousinage institutions are not found among Sierra Leone's main ethnic groups, although they do exist in Kuranko areas in the north (Jackson, 1974). Baldwin and Huber (2010) argue that between-group economic inequality is the key driver of adverse ethnic diversity impacts observed across countries.

The rest of the paper is organized as follows. Section 2 provides background on economic development and ethnicity in Sierra Leone. Section 3 presents results on ethnic-based migration patterns, and discusses our historical instrumental variable approach. Section 4 describes the estimation strategy and the data, and Section 5 presents the main results. Section 6 weighs the contrasting mechanisms that might explain our results, and the final section concludes.

3.B Background on Economic Development and Ethnicity Sierra Leone

Viewed from multiple perspectives, Sierra Leone is among the world's poorest countries. According to the United Nations Development Program's 2007-2008 Human Development Report, Sierra Leone's human development index in 2005 was 0.336, the lowest score in the world at 177th out of 177 countries with data. Per capita GDP (adjusted for purchasing power parity) is US$806. Life expectancy at birth is a tragic 41.8 years, ranking Sierra Leone 173rd out of 177 countries. Adult literacy is just 34.8%, and while there has been progress in school enrollment after the civil war, gross secondary school enrollment was only 32% in 2007. Nearly half of the population lacked access to an improved water source (such as a borehole well, protected spring, or piping) in 2004. While the recent 1991-2002 civil war is undoubtedly a contributing factor, Sierra Leone already had the second lowest human development index in the world before the war began (UNDP 1993). In fact, the country's disappointing economic performance, together with ubiquitous government corruption, arguably contributed to the outbreak and duration of the war.

Sierra Leone is also one of the world's most diverse countries. The household module of the 2004 Population Census identifies eighteen major ethnic groups. The Mende and Temne are numerically dominant, occupying shares of 32.2% and 31.8%, respectively, while the Limba, Kono, and Kuranko are the next largest groups, at 8.3%, 4.4%, and 4.1%, respectively. Other groups occupy a substantially smaller share, including the Krio, whose population share fell to only 1.4% by 2004. Data from the 1963 Census demonstrates the stability of national ethnic composition over time (Appendix Table 3.A.1).

These groups are characterized by distinct customs, rituals, and history, and, most importantly, language. With the exception of Krio, an English dialect, the other languages are members of the Niger-Congo language family. Within this family, the most salient distinction is between the Mande languages – including Mende, Kono, Kuranko, Susu, Loko, Madingo, Yalunka, and Vai – and the Atlantic-Congo languages, including Temne, Limba, Sherbro, Fullah, Kissi, and Krim. These groups are mutually unintelligible, and much further apart linguistically, for example, than English and German.[3] It is not a coincidence that the main

[3] See for example the World Language Tree of Lexical Similarity (2009).

political fault line lies between the two language groups.

The 2004 Census contained an ethnicity question, allowing us to compute ethnicity shares at the chiefdom level. Chiefdom boundaries have been relatively unchanged since independence, and the chiefdom is still the geographic unit by which most Sierra Leoneans self-identify their origins, as well as the administrative level at which traditional authorities are organized. There are 149 chiefdoms in the country, and the median chiefdom population is roughly 22,000. Denote ethnicity shares by $\pi_{ik} = N_{ik}/N_i$, where N_{ik} is the number of individuals of ethnicity k living in chiefdom (or EA) i and $N_i = \sum_k N_{ik}$ is the total chiefdom population. Using these shares, the standard ethnolinguistic fractionalization measure (which is closely related to a Herfindahl index) is $ELF_i = 1 - -\sum_{k=1}^{K} \pi_{ik}^2$. ELF_i captures the probability that any two individuals randomly chosen from the population belong to different ethnic groups.[4] We also create ethnicity shares at the enumeration area (EA) level. In rural areas, an EA is equivalent to a medium sized village or a small village and surrounding hamlets.[5] There are approximately 9,600 EAs in Sierra Leone, with an average population of 483.

The mean of chiefdom ELF in our sample is 0.264 (standard deviation 0.196). Figure 3.1 presents non-parametric estimates of the distribution of ELF_i across chiefdoms (panel A) and EAs (panel B). It should be clear from these figures that ethnic diversity is, on average, greater at the chiefdom than at the village level. Across EAs, most of the mass of ELF_i is in the left tail, while the distribution across chiefdoms is more diffuse. This is consistent with the view that much of the rural population in Sierra Leone is settled in remote and relatively homogeneous communities (the average share of the dominant ethnic group in rural EAs is 88%). Nevertheless, there is considerable variation in ELF even within rural communities, with the average share of the dominant group falling to 63% in the most diverse quartile of EAs.[6] Figure 3.2 panels A and B map chiefdom ethnic diversity currently and historically, respectively. Visual inspection indicates that diverse areas were likely to remain diverse between 1963 and 2004, a result we confirm in a regression below. Moreover, diverse chiefdoms are found throughout the country.

Questions on religious identification were unfortunately not collected in either the 1963 or 2004 censuses, so we use nationally representative household survey data from the 2005 and 2007 National Public Services (NPS) surveys to construct religious diversity measures. We consider the proportion of respondents in each chiefdom who practice the country's two major religions, Islam and Christianity, ignoring their internal subdivisions. Sierra Leone is predominantly Muslim, at 76.8%, but Christianity is also widely practiced (22.4%), with other religions making up the remaining 1%. The mean of chiefdom religious fractionalization

[4] Using Montalvo and Reynal-Querol (2005) preferred ethnic polarization measure in place of ELF_i does not change the main result of no ethnic impacts below (not shown).

[5] Sixty-three percent of rural EAs contain only one locality (village), and 90 percent contain three or less.

[6] Note that diversity across EAs is not driven by differences between EAs with one vs. multiple localities. The EA level ELF measures do not change appreciably for EAs that contain a single locality (not shown).

is 0.229 (standard deviation 0.179, Appendix Figure 3.A.1).

3.C Migration and the Persistence of Local Ethnic Composition

In this section, we use data from the nationally representative 2007 National Public Services (NPS) household survey to study individual internal migration decisions during and following the war. Many Sierra Leoneans place a high value on living in chiefdoms that were historically settled by members of their own ethnic group, and this preference varies across population sub- groups, as discussed below. This systematic sorting as a function of local ethnic composition necessitates the use of the instrumental variables strategy presented in section 3.2.

3.C.1 Revealed Preferences for Ethnic Sorting

The 2007 NPS survey collected information on respondents' current and 1990 chiefdom of residence. To understand why individuals moved, we estimate a conditional logit model, which can be derived from the following random utility model. Let $i = 1, ..., N$ index individuals and $j = 1, ..., J$ index chiefdoms. We model the indirect utility of individual i living in chiefdom j as:

$$V_{ij} = X'_{ij}\beta - -\alpha D_{ij} + \varepsilon_{ij} \tag{3.1}$$

Here, X_{ij} denotes a $(K \times 1)$ vector of characteristics for chiefdom j, including certain characteristics of individual i interacted with chiefdom values. For example, one component of this vector is the ethnolinguistic fractionalization in chiefdom j, and another is this value interacted with individual i's educational attainment. Other specifications focus on the coethnic residential share and its interaction with education. It is through these interactions that the discrete choice model captures preference heterogeneity. The variable D_{ij} denotes the distance between the centroids of individual i's home chiefdom and chiefdom j. If D_{ij} is thought of as the "price" individual i pays to move to chiefdom j, we can interpret the ratio $-\beta_k/\alpha$ as the willingness to pay for a one unit increase in characteristic X_{kij} in terms of kilometers moved. Individual i chooses to live in chiefdom j if $V_{ij} > V_{ij'}$ for all other chiefdoms j'. Given these standard assumptions, the probability that individual i chooses chiefdom j, denoted P_{ij}, is:

$$P_{ij} = \frac{\exp\{X'_{ij}\beta - -\alpha D_{ij}\}}{\sum_{k=1}^{J} \exp\{X'_{ik}\beta - -\alpha D_{ik}\}} \tag{3.2}$$

We use weighted maximum likelihood estimation to address the choice-based sampling issue.[7]

Of the 5,488 individuals in the sample, 26.5% had moved to a different chiefdom since 1990, and among those who had moved, nearly two-thirds (62.2%) moved to a different district (there are 19 districts in all); Appendix Table 3.A.2 presents descriptive statistics. The average distance between the centroids of the 1990 and 2007 chiefdoms of residence for movers was 74.3 kilometers. Information was not collected on migration patterns during the war; we only observe retrospective data on the chiefdom of residence before the war started and the post-war chiefdom of residence in 2007. However, we do know whether anyone from the respondent's 1990 household was made a refugee: 23.2% of our sample had 1990 household members who temporarily fled Sierra Leone, often to refugee camps in Guinea.

We do not include 2004 chiefdom ethnicity shares when estimating equation 2 because they are endogenous to war and post-war migration choices. Instead, we include chiefdom level ethnicity data from the 1963 Population Census for a predetermined measure (and use this data again below in the construction of historical ethnicity instrumental variables). Table 3.1 shows the main conditional logit results. All columns include distance D_{ij} and either the co-ethnic population share in 1963 (columns 1-2) or the 1963 chiefdom ELF score (columns 3-4) as the key explanatory variable. Greater distance between chiefdoms is associated with a lower propensity to move, as expected, and there is a significant positive preference for living in areas traditionally dominated by one's own ethnic group. In column 1, the ratio of these two coefficient estimates implies that individuals are on average willing to travel an additional 10.1 kilometers to live in a chiefdom with a 10 percentage point greater share of her/his own ethnic group. The coefficient estimate on chiefdom ELF is also statistically significant (column 3) conditional on other factors (including remoteness from cities as well as population size and density), suggesting a positive preference for diversity, though this is smaller than the preference for a higher co-ethnic share.[8] Sierra Leoneans on average also show a strong preference for moving to chiefdoms with historically larger populations, and a dislike for moving to remote areas or to areas not well connected by roads.[9] Controls for the number of attacks and battles experienced in chiefdom j during the war, and the presence of mining operations do not change the estimated willingness to pay for residence with co-ethnics.

We next explore differential willingness to pay for ethnic homogeneity for people who have "some education" and those who have none (column 2 and 4); recall that the median Sierra Leonean adult has zero years of schooling. Educated individuals are less responsive

[7] Because the survey was designed as a stratified random sample (based on current location), the sample is choice- based. Under the assumption that migration between 2004 and 2007 was negligible, which is plausible since most postwar resettlement occurred by 2004, weighted maximum likelihood resolves the issue (see Manski and Lerman (1977) and Appendix 3.B).

[8] Note that this diversity result holds whether or not the local co-ethnic population share is controlled for.

[9] Beyond Freetown, other towns include the other large cities in Sierra Leone, namely, Makeni, Bo, Kenema, and Koidu, as well as smaller towns such as the district capitals, Kabala and Kailahun Town.

to moving distance and care much less about living in chiefdoms with greater shares of their own ethnic group. The ratio of these two coefficient estimates implies that educated individuals are only willing to travel an additional 8.6 kilometers to live in a chiefdom with 10 percentage point greater share of her/his own ethnic group. This finding suggests that education dampens co-ethnic residential preferences. More educated people are more likely to move to ethnically diverse areas, and this finding underlines the potential for bias in simple OLS estimates. For example, if those with higher education are more likely to move to diverse areas and also exhibit greater participation in collective action, then the OLS coefficient estimate on ethnic diversity could be biased.

Individuals who directly experienced violence during the war find moving greater distances more costly, prefer living with co-ethnics and dislike ethnic diversity compared to the average Sierra Leonean (columns 2 and 4).[10] Individuals from chiefly "ruling" families have a somewhat greater aversion to moving further distances away from their home area, which is sensible since their influence rarely extends beyond chiefdom borders, and appear to have stronger co-ethnic residential preferences than others.

3.C.2 Using Historical Data to Identify the Impact of Ethnic Diversity

In the absence of random assignment of people to locations, the systematic sorting of individuals from particular ethnic groups, or with certain (unobserved) tastes for public goods, into more or less diverse areas could potentially introduce omitted variables bias into cross-sectional estimates of the impact of diversity on local collective action. Recent sorting, during and after Sierra Leone's 1991-2002 civil war, is a particular concern for our empirical work. Hundreds of thousands abandoned their homes, fleeing violence, and some spent years in refugee camps, while others sought out regions of the country protected from RUF attacks. As discussed above, while 73.5% returned to their 1990 home chiefdom by 2007, those that did not were different on both observable and unobservable characteristics than those that did. This could bias simple OLS estimates of the effect of diversity in a direction that is difficult to sign.

In the ideal thought experiment, the impact of ethnic diversity on local outcomes would be credibly estimated if individuals were first randomly allocated to jurisdictions and then worked together to provide local public goods. In this subsection, we argue that a close historical parallel occurs in areas with stable ethnic land settlement, where the causes of the

[10] A number of different interpretations of this result are possible. For example, those who found it more costly to move in the face of approaching violence may have been more likely to experience it directly, or the effects of experiencing violence (e.g., maiming) may have made it harder for them to move and more reliant on local (including ethnically-based) networks. As discussed above, there is no evidence that civil war violence was ethnically targeted, nor do we see civil war violence leading to less local collective action in higher ELF communities in the next section.

current residential patterns – in rural west Africa, the slave raids, wars droughts, famines, and epidemics that took place in the 18th century if not earlier – are largely uncorrelated with modern-day socioeconomic factors that might affect public goods provision. In particular, we focus on specifications where current local (chiefdom or enumeration area) ethnic diversity is instrumented using historical local diversity measures from the 1963 Population Census. The IV exclusion restriction is that historical ethnic diversity affects only current residential diversity and is not correlated with any unobserved local factors that might change the costs of, or preferences for, providing local public goods. While even longer historical lags, i.e., census data before 1963, would have made the case even stronger, there is unfortunately no comprehensive national population data for earlier periods.

In Sierra Leone, most historical ethnic boundaries were shaped during the period of the Atlantic slave trade, as raiding tribes settled in conquered areas and drove weaker groups deeper into the forest. The Mane, progenitors of the Mende ethnic group, arrived after the collapse of the Mali empire and first settled in today's Sierra Leone in 1545 (Oliver and Atmore, 2001). Throughout the 16th and 17th centuries, Mane tribes invaded and conquered the ethnic groups that already lived there, reshaping ethnic boundaries and taking prisoners, either to be kept as domestic slaves or for sale to European slave traders.

In a separate historical episode, the Fulbe of Futa Jallon formed a powerful Muslim state in what is now eastern Guinea (which borders Sierra Leone) in 1726, and declared jihad against the neighboring tribes. Their state conducted regular slave raids throughout the rest of the 18th century, putting pressure on groups to move and resettle, especially into Sierra Leone's northern districts. By the time the first British and freed slaves arrived in Freetown in 1787, most of the current ethnic borders had already been drawn. The decline of the external slave trade during the late 19th century, combined with an increased British military and administrative presence in the Protectorate by century's end, partially restrained wars between ethnic groups and helped to preserve largely stable ethnic borders.

The fact that historical ethnic settlement patterns were driven by slave raiding and warfare centuries ago makes it far less likely that local diversity is correlated with omitted factors that would affect current public goods, relative to more recent migration. However, there remain at least five plausible violations of the exclusion restriction – i.e., ways in which historical ethnic diversity might still influence current local public goods provision other than through current ethnic diversity – that merit consideration. For one, individuals may have different preferences for diversity than society as a whole, relocate based on these preferences, and then pass down to their descendents a higher than average preference for cross-ethnic cooperation. For this mechanism to undermine the validity of our instrument, however, there would need to have been considerable relocation based on ethnic cooperation preferences prior to 1963 and very high persistence in these preferences across generations, which seems implausible.[11]

[11] For example, if one's grandfather had a 1 s.d. higher preference for cooperation than the average and

Similarly, if more educated individuals have greater taste for diversity and for providing public goods, and if these characteristics are passed down through the generations, this could also undermine the validity of our IV strategy. Yet this is not a major concern because only 2.8% of individuals in rural Sierra Leone were literate in 1963, and thus ancestors' education is not a strong predictor of current education. Moreover, there are no significant correlations between literacy and ethnic diversity in 1963.[12]

A third potential concern with the IV strategy would be if current levels of public goods were directly determined by historical investments, as would be the case, for instance, in the United States, where many present-day libraries and schools were built in the early 1900s. If public goods were persistent and preferences for public goods were persistent then the distribution of ethnic settlement and public goods now would reflect Tiebout (1956) style sorting in the 1960s. In rural Sierra Leone, however, this is unlikely to matter. The vast majority of public goods investments were made after 1963 – there were virtually no rural schools in 1963, for example, as is illustrated in the abysmally low literacy rate – and many of our key public goods measures have very high depreciation rates; road clearing and maintenance, for instance, typically lasts only a few months in Sierra Leone's dense tropical rainforests.

Fourth, historically strong chiefs may have been more successful at encouraging (or forcing) assimilation of slaves and other "strangers" into adopting the ethnic identity of the dominant local group, as Posner (2005) argues occurred in Zambia in the early 20th century. However, to the extent that strong rulers did promote ethnic assimilation, this would bias us towards finding a negative relationship between local diversity and public goods, but despite any such bias we do not find negative impacts below.

Finally, if certain economic activities (such as trading or mining) require greater inter-ethnic cooperation and also produce higher levels of income, and the geographic distribution of these activities persists over time, this could undermine the validity of our instrumental variable. However, there is no correlation between formal sector employment and chiefdom ethnic diversity in 1963 (see Appendix Table 3.A.3), indicating little sorting along these lines in colonial times, as well as arguing against the view that richer areas saw more ethnic assimilation. The census indicates that the vast majority of households in rural Sierra Leone were engaged in the same economic activities in 1963, namely subsistence farming of rice and cassava. Yet because of this concern, we exclude all urban areas throughout the analysis, and as a robustness check also exclude the diamond mining areas in the country's east (Kono district), which experienced an economic boom in 1940s and 1950s, attracting migrants from throughout Sierra Leone.

moved to a more diverse area as a result, and there was partial mean reversion such that each generation's preferences were halfway between their parents and the national average, then preferences for inter-ethnic cooperation in the grandchild's generation would be only be 0.25 s.d. higher than average.

[12] See Appendix Table 3.A.3. Note that although the coefficient estimate is nearly significant at the 90% level, the magnitude remains small and falls closer to zero if a handful of outliers are omitted (not shown).

The lack of statistically significant relationships between observable socioeconomic characteristics, namely literacy and formal employment, with local ethnic diversity in the 1963 census, together with the historical evidence on the determination of ethnic boundaries during invasions and slave raids during the 16th to 19th centuries, both help alleviate concerns about bias caused by endogenous historical sorting.

Table 3.2 presents the first stage regressions of 2004 ethnic diversity on the historical measures, and finds remarkably strong correlations both at the chiefdom level (panel A) and the enumeration area level (for the NPS sample in panel B, although note that the historical measures can only be disaggregated to the chiefdom level). In the key result, the coefficient estimate on 1963 chiefdom ethnolinguistic fractionalization is 0.797 (standard error 0.089, column 1, Panel A), for a t-statistic of 9. Historical ethnic shares (and squared shares) for the two largest ethnic groups are also included as instruments for current ethnic shares to capture possible differences in average public goods preferences across groups (columns 2-5). Judging by the R^2 values, 1963 ethnic diversity variables explain the lion's share of the chiefdom-level variation in current ethnicity measures. Chiefdom level historical diversity measures also predict EA diversity, as ethnic groups are not perfectly segregated within villages, and chiefdom diversity is partly reflected at the village-level. The coefficient estimate is 0.429 (s.e. 0.110 – Table 3.2, Panel B, Column 1). This allows us to employ historical chiefdom measures to IV for current EA diversity below. Graphical representations are depicted in Figure 3.3, plotting the residuals from the regression of ELF in 2004 on the 1963 ethnic share controls (on the y-axis) versus the residuals from regressing 1963 ELF on the same ethnic share variables (x-axis). The slope of the line corresponds to the coefficient on 1963 ELF in Table 3.2, column 1.

3.D Estimation and Data

We next describe our regression specifications (section 4.1) and the data (section 4.2).

3.D.1 Regression Specifications

Let $k = 1, ..., K$ index the outcome variables Y_k, and let j index observations (usually at the chiefdom or enumeration-area level). For each outcome, we first estimate the OLS regression:

$$Y_{jk} = \alpha_k + \beta_k ELF_j + X'_j\delta_k + S'_j\gamma_k + \varepsilon_{jk} \tag{3.3}$$

where ELF_j is the chiefdom ethnolinguistic fractionalization measure and X_j is a vector of average socioeconomic and demographic controls for households in locality j. S_j is a vector denoting the ethnicity shares (and squared shares) of Mendes and Temnes in chiefdom j, and ε_{jk} is the error term. Vigdor (2002) argues that including ethnicity group shares is essential for the correct interpretation of the diversity coefficient estimate β_k. We also interact ELF_j with some characteristics X_j to explore heterogeneous impacts. When the outcome is measured by EA, disturbance terms are clustered by chiefdom.

In the IV specifications, current ELF and ethnic shares (and squared shares) are instrumented with their historical 1963 values. We interpret the resulting IV-2SLS estimates as capturing the local average treatment effect (LATE) of ethnic diversity on outcomes among the chiefdoms that had stable ethnicity patterns over 1963-2004. Because we have a strong first stage relationship (Table 3.2), we argue that this sub-group of ethnically stable chiefdoms is large and important. However, it is worth emphasizing that the IV strategy does not allow us to estimate diversity impacts in areas that experienced large changes in diversity over the period; examining the impact of diversity in these areas is also potentially of interest but is not a topic we can study with this identification strategy.

The specifications below report results with both the chiefdom and the enumeration area as the unit of analysis. One reason to focus on chiefdoms is that the 1963 census data are not available at a more disaggregated geographic level. Moreover, the chiefdom is also a relevant political unit of analysis given the continued power of Paramount Chiefs in rural Sierra Leone. Paramount Chiefs, and the section and village chiefs below them, have a particularly prominent role in organizing local collective activities, and are well known and respected among citizens. For some quantitative evidence of this, in 2007 NPS data, 82% of household respondents could correctly name their local Paramount Chief while only 44% were able to identify their Local Council representative or representative in the national parliament. Individuals were also much more likely to have visited the chiefdom headquarters than they were to have visited the local council headquarters; self-expressed trust for chiefs (at 43%) is much higher than trust for elected local councilors (29%); and respondents are much more likely to think that chiefs are responsive to local needs (62%) than local councilors. Yet we also examine diversity impacts at the EA level because many of our outcomes, such as road maintenance, are organized primarily on a village by village basis. Different aggregation choices do not affect the main results.

We investigate ethnic diversity impacts on a number of closely related outcomes, and create summary impact measures using a mean effects analysis, following Katz et al. (2007). The groupings of related outcome variables are denoted by Y_k, $k = 1, ..., K$. We then standardize each outcome by subtracting the mean and dividing by the standard deviation of the outcome variable among below-median ELF areas (a low diversity control group of sorts). The standardized outcome variables are denoted Y_k^*. With these, we form $Y^* = K^{-1}\sum_k Y_k^*$, a single index of outcomes, and we regress this index on ELF as in equation 3. The coefficient on ELF in this regression is the mean effect size. Note that we defined the outcome variables so that "better" is always positive; for instance, finding that ELF and disputes were positively correlated means disputes are lower in more diverse areas.

In terms of the sample, we drop all observations from Sierra Leone's six largest urban areas – Freetown, Bo Town, Kenema Town, Makeni, Bonthe Town, and Koidu – which together make up the vast majority of the country's urban population.[13] The nature of local collection action and public goods provision is qualitatively different in urban and rural areas – for instance, as a legacy of its settlement history, there are no chiefs in Freetown – and for reasons of comparability we thus focus on rural areas, where most population lives. As a robustness check, we also exclude chiefdoms in Kono district, the country's diamond mining center.

3.D.2 Local Measures of Public Goods, Collective Action, Social Capital and School Quality

The 2005 and 2007 National Public Services (NPS) Surveys are nationally representative surveys that asked over 6,000 respondents questions about their access to and satisfaction with public services.[14] The survey also contains questions designed to measure social capital, broadly defined. We create four broad categories of outcome variables; descriptive statistics see Appendix Table 3.A.4. The first grouping for the mean effects analysis is what we call *local collective action.* These outcomes include: road maintenance, known in Sierra Leone as road brushing, a locally organized activity to keep bush paths between villages passable, which is a critical public good especially in remote villages; participation in communal labor or other community projects (such as school construction); and attendance at community meetings, events where people voice concerns and make decisions about other local activities. These variables all capture some aspect of the effectiveness of local efforts to provide public goods. The local representative of the chiefdom authority often monitors these activities and has the power to fine non-participants (in road brushing, for instance), so we first look for diversity effects across chiefdoms (Table 3.3). Average participation in road brushing (by men) and in community meetings over the last month was quite high at around 40%, though there is wide variation across chiefdoms.

The second category of outcomes is *group membership*, which measures participation in community self-help groups, such as women's associations, youth groups, and religious groups. It also includes questions on groups with more economic significance, such as trade unions, school management groups, and credit groups. The latter may facilitate agricultural investment and boost farm productivity. Decisions to join these groups are made by individuals, and their choices plausibly reflect the degree of cooperation within a community. Most chiefdoms show high rates of community group participation at over 80% membership in at least one group, though average participation in credit groups and school groups was lower

[13] We omit one chiefdom (Kakua in Bo District), much of which is a neighborhood of Bo Town, one of Sierra Leone's largest cities. This leaves a main analysis sample of 146 chiefdoms.

[14] NPS data collection was designed so that half are administered to female respondents and half to male respondents, usually the head of household or her/his spouse. The surveys were originally intended to form a panel, but because of insufficient funding for respondent tracking, the matching rate is relatively low, and thus the data are treated as a repeated cross-section.

and more variable.

The third category is the *control of community disputes*. Respondents were asked questions about whether they were the victim of theft, physical attack, or were involved in land disputes. Obviously, in this case, in contrast to the previous two categories, higher values reflect worse local outcomes. In 2005, the average incidence of theft was quite high (27%), but by 2007 it had fallen substantially (though this may be due in part to a change in question wording across the two survey rounds). Physical attacks and land disputes were relatively infrequent. Traditional chiefs and their local representatives (e.g., village headmen) have explicit authority over public safety, and they also oversee the local courts which punish these offenses. The capability and performance of chiefly authorities may thus directly affect the control of community disputes. Chiefdom level diversity measures are also relevant as some disputes occur between neighboring EAs which may be dominated by different ethnic groups (e.g. disputes over cattle).

The fourth and final category is *trust*. Respondents were asked about the extent to which they trusted people in their community, as well as outsiders, local officials (chiefs and local councilors), and Members of Parliament in Freetown. Perhaps unsurprisingly, self-reported trust is much higher for members of respondents' own communities than for outsiders (at 91% versus 48%, respectively, in 2005). Trust for government officials is lower on average and falls noticeably between 2005 and 2007. Some of this decline may be explained by the end of the honeymoon period enjoyed by leaders in the immediate aftermath of the war but some is also the result of a change in question wording between survey rounds.[15] While the public goods measures we just described – road maintenance, communal labor, village meeting attendance, crime control and trust – are plausibly thought of as truly local, school quality is the result of a combination of village, chiefdom, local council, and central government decisions, as well as non-governmental organization (NGO) investments. For instance, the building of formal schools and hiring of government teaching staff are typically the responsibility of the Ministry of Education in Freetown, national reconstruction agencies and large Christian organizations, and thus are mainly determined by national policy or political concerns rather than by local collective action alone. Yet many communities supplement government provision by locally funding community teachers, paying for repairs and supplies and even building some community schools. Successful community organization can also impact the quality of public education through more indirect routes like lobbying the central government or attracting NGO support. Ethnic cooperation may also work through the provider sidei.e., if teachers show up to work more frequently when working in areas dominated by their own group.

School quality data was collected in the 2005 School Monitoring Survey. Enumerators made unannounced visits to a nationally representative sample of 338 schools and collected

[15] Wording changed for several questions between the 2005 to 2007 rounds, including the time period for the community meeting participation questions (i.e., annual versus monthly), trust questions, and control of community dispute questions. While the means of these variables change across rounds, it is still appropriate to group them together in the mean effect analysis since all variables are first normalized.

information on the quality of school buildings, the number of classes taught, whether teachers were present, and the availability of supplies for instruction. We employ data from the 281 schools not in Freetown or other large towns; descriptive statistics are in Appendix Table 3.A.5.

School outcomes were organized into three broad categories. The first set of school quality outcomes is *instructional supplies*. Enumerators recorded the number of desks, chairs, blackboards, and textbooks in use at the time of their visit. Together with school enrollment data, these allow us to construct a variety of per student input measures. Most supplies are either provided directly by central government or paid for though a small non-salary grant the central government sends to local schools (the so-called school fee subsidy). Communities can affect school supplies by effectively overseeing the school fee subsidy and ensuring it is spent properly on education (rather than being diverted or stolen), and by raising additional local funds, although this additional fundraising is limited in most communities.

The second category is *teaching quality* measures. Enumerators arrived unannounced at the primary schools and noted teacher absence; almost 40% of teachers were not present during these surprise visits, a remarkably high rate. If teachers were present, they also observed teacher classroom behavior upon arrival at the school (i.e., were they teaching, grading, sitting idly, chatting with other teachers, or talking on the phone, etc.), which allows us to compute the proportion of teachers who were actually working when the unannounced visit was made. On average, conditional on being present 80% of teachers were actually working when the enumerators arrived at a school.

The third category is *facilities quality*. Enumerators collected information on whether the school had a functioning toilet, electricity, and water supply, and whether the roof, floor, and walls of the school were made with sturdy building materials (e.g., concrete) rather than mud or thatch. Once again communities can raise additional funds locally to build or repair a school. Usually, however, communities only raise money to build temporary classroom structures when the central government has not yet built a permanent structure. The vast majority of schools in our sample are government built structures, so this category is plausibly one where local collective action is less important in practice.

3.E Impacts of Ethnic Diversity on Local Public Goods, Social Capital, Disputes and Schools

We first present estimates of the relationship between ethnic diversity and participation in road maintenance (brushing) across chiefdoms (Table 3.3). The first three columns contain OLS estimates, while the second three use the historical instrumental variables based on 1963 population census data. In column 1, we regress road brushing on ELF_j (and ethnicity share controls). The coefficient estimate on ELF_j is small and positive but not statistically significant. In column 2, we add controls for civil war conflict experiences and other socioeconomic and demographic controls. Most controls have little impact on estimated diversity effects, with the exception of the proportion of residents with some education, which is strongly positively correlated with road brushing, and the extent of civil war violence exposure, which is also positively related to road brushing in the chiefdom level analysis, echoing the perhaps surprising positive war impact findings in Bellows and Miguel (2009). Column 3 estimates interactions between ethnic diversity and war exposure, and finds that diversity effects are no different in areas that experienced worse war-related violence. The coefficients on ELF_j do not change substantially in the IV specifications (Table 3.3, columns 3-6), although some point estimates become slightly negative. Overall, ethnic diversity does not have a statistically significant impact on participation in road maintenance, one of the most important, time consuming and truly local and non-excludable public goods in rural Sierra Leone. Figure 3.4 presents these findings graphically, and Appendix Tables 3.A.6 and 3.A.7 show that the main findings are robust to different units of analysis (enumeration area and individuals, respectively).

We next assess whether the failure to find significant diversity effects is due to a lack of statistical power. One way to explore this question is to determine the magnitude that any diversity impact would need to have for us to detect it as statistically distinguishable from zero. Again consider road maintenance. From the IV specification with full controls in column 5 of Table 3.3, the estimated ethnic diversity effect on road maintenance participation is -0.083 with a standard error of 0.192. With 95% confidence, then, the true effect of diversity lies in the interval $[0.459, 0.293]$. If we perform the thought experiment of increasing ELF by one standard deviation (or roughly 0.2), the confidence interval implies that a change in road maintenance would lie inside $[0.09, 0.06]$ with 95% probability. Road maintenance participation has a standard deviation of 0.21, so we can reject the null hypothesis that a one standard deviation increase in diversity affects road maintenance by more than $\pm 0.3 - 0.5$ s.d., a moderate effect magnitude.

Table 3.4 reports mean effect estimates for the four groups of local outcomes – collective action, group membership, control of disputes, and trust – using both OLS and IV specifications, across different levels of aggregation (chiefdom-level in panel A and enumeration-area in panel B). It also reports the mean effect estimates for the three groups of school outcomes in panel A at the chiefdom-level (the sample does not allow estimation at the EA level). As with road brushing, the estimates remain close to zero for all four categories and almost none are significant at traditional confidence levels. Statistical precision falls in the IV specifications at the EA level, as expected given the weaker first stage (Table 3.2, Panel B).[16] The mean effects analysis for school supplies, the quality of teaching, and the quality of school buildings all tell a similar story: there are no significant effects of ethnic diversity in OLS or IV specifications, with or without controls (Panel A), and the "zero" estimates are precisely estimated.

[16] In unreported results, we did not find evidence of different effects in multiple-locality EAs (not shown).

Figure 3.5 reports 95% confidence intervals on the ethnic diversity effect estimates across all the variables that go into the mean effects indexes, with all variables standardized (to be mean zero and standard deviation one) to facilitate comparison. In all cases, we report confidence intervals based on IV specifications with the full set of controls (comparable to column 5 in Table 3.3). The confidence intervals for all outcomes intersect the vertical zero line, indicating that estimated diversity effects are not statistically significant. Moreover, the estimated zeros are again reasonably precise. Following the same exercise as above, the 95% confidence on the standardized effect size of a one standard deviation increase in ELF are: $[-0.4\sigma, 0.3\sigma]$ for the collective action mean effect, $[-0.2\sigma, 0.3\sigma]$ for the group membership mean effect, $[-0.1\sigma, 0.4\sigma]$ for the disputes mean effect, and $[-0.0\sigma, 0.2\sigma]$ for the trust measures. We view these as quite tightly estimated zero effects, such that even the moderate impacts falling outside these intervals can be ruled out with 95% confidence.

As a robustness check, we exclude the main diamond mining areas in the country's east (Kono district), and once again find no statistically significant ethnic diversity impacts on any of the four main mean effects categories (not shown). In a further robustness check, we created another diversity measure capturing the extent to which ethnic groups differ by language family rather than ethnic group. Recall from section 2.2 that the most salient distinction is between groups speaking Mande languages (e.g., Mende and others) versus Atlantic-Congo languages (Temne, Limba and others). We thus create a fractionalization index that captures the probability that two randomly sampled individuals speak languages from different families, and regressed our local public goods measures on this index. In a mean effects analysis (Appendix Table 3.A.8), some estimates are statistically significant in the enumeration area OLS results, but they are neither robust to including demographic controls, nor to the preferable IV approach, or analysis at the chiefdom-level.[17]

One concern with ELF is that it treats all ethnic differences identically, regardless of the history of group relations. This potentially introduces noise into the diversity measure, and might thus bias diversity estimates towards zero. We created an alternative diversity measure that only considers ethnic differences if particular pairs of groups had a history of armed conflict. Kup (1961) provides a detailed overview of how different ethnic groups came to settle the area that is today's Sierra Leone, and their conflicts during the 1400-1787 period. Let $\mathbf{s}_c$ denote a $(J \times 1)$ vector of ethnicity shares for chiefdom (EA) c. We can define the historical ethnic conflict index for chiefdom (EA) c as follows:

$$HCON_c = \mathbf{s}_c' \Gamma \mathbf{s}_c \tag{3.4}$$

where $\Gamma = [\Gamma_{jk}]$ is a $(J \times J)$ matrix with a typical element equal to 1 if groups j and k had historical conflicts with one another and zero otherwise. This matrix is depicted in Appendix

[17] Using the language-family diversity measure in the school quality regressions did not lead to significantly different results from those that use our initial ELF measure (not shown).

Table 3.A.9, with entries drawn from Kup (1961).[18] The correlation between HCON and ELF is high at 0.783 across chiefdoms. Again using a mean effects analysis (in Appendix Table 3.A.10), there are no robust impacts of this measure on local public goods outcomes.

Another important dimension of social identity in Sierra Leone is religion. Unfortunately, the 1963 Census does not allow us to construct measures of historical religious diversity, so we rely on OLS estimates. There is no evidence of adverse effects of religious diversity on local collective action: for collective action, group membership and control of disputes, the point estimates on the religious diversity measure mean effects are negative but not statistically significant (Appendix Table 3.A.11).

3.F Explaining the weak relationship between diversity and local outcomes in Sierra Leone

In this section, we first discuss historical factors that affected the ethnic and economic cleavages in Sierra Leone, before turning to other factors, including the role of Krio as a lingua franca, that might serve to promote cooperation between groups (sections 6.1 and 6.2). Finally, we discuss the legacy of Britain's support for chiefs in section 6.3, which simultaneously preserve the salience of ethnicity while also promoting local collective action, although we do not find evidence that strong chiefs help promote local collective action in section 6.4.

3.F.1 Overview of Colonial History

One key difference between Sierra Leone and many other African countries is that the favored ethnic group during early colonialism, the Krio, were not truly indigenous. The Krio ethnic group are descendents of freed slaves who settled Freetown starting in the late 18th century. They were a powerful ethnic group during the 19th and first half of the 20th century but have since shrunk to demographic (and political) insignificance. Thus as Sierra Leone made its transition to independence in 1961, the primary source of political conflict shifted. As stated by Kandeh (1992), the salience of the Creole [Krio]-protectorate cleavage was eclipsed after independence by the rivalry between the Mendes of the south and Temnes of the north. The implications of this on Sierra Leone's political culture are many, and we argue it has plausibly helped shape inter-ethnic relations to the present day.

[18] In constructing this matrix, Kup (1961) provides data for all conflicts between 1460 (when the first European explorers came to Sierra Leone) and 1787 (when Sierra Leone was first colonized). From our reading of history, we assumed that the Krio were in conflict with all other groups. We also assumed that the Mandingo were in conflict with all other groups because they were notorious slave raiders, despite the fact that they are not mentioned frequently in the Kup (1961) text.

The Krio in the Colonial and Protectorate Period, 1787-1961

In 1787, with funding from English philanthropists including Granville Sharp, former slaves arrived at the peninsula of Freetown, now known as Sierra Leone's Western Area, negotiating purchases of land from local chiefs.[19] For a brief period, the Creoles, or Krio as they became known, governed themselves, but after attacks on the initial settlement by Temne warriors, Sharp needed to solicit additional funds to defend and repopulate the settlement. To do so, he aligned himself with commercial interests and in 1791 his investors formed the Sierra Leone Company, whose mission was to substitute legitimate commerce between Africa and Great Britain for the slave trade (Spitzer, 1974, p. 10). Under the company's 1800 Charter, directors could appoint government officials in Freetown. When the company went bankrupt in 1808, its lands were taken over by the British government and Sierra Leone became a British Colony.[20]

At that time, the colony of Sierra Leone referred only to the country's western peninsula. The rest of what is now Sierra Leone was never formally colonized but was instead annexed as a Protectorate in 1896 (see Figure 2). Residents of Freetown experienced direct British rule, which allowed many Krios to rise to positions of considerable authority in the colonial government, most notably on Sierra Leone's Legislative Council.[21] In contrast, in the Protectorate native Sierra Leoneans experienced indirect rule, a system that promoted chiefs loyal to the British, and institutionalized – and in many cases augmented – their autocratic power over their subjects, exacerbating inequality and reinforcing social divisions. The divergence in the governmental structures of the Colony and the Protectorate was reflected in vast social differences between Freetown residents – who were Christians, often literate in English, and saw themselves as defenders of Western civilization – and those who lived "up-country". According to Kandeh (1992, p. 83), "protectorate Africans were commonly referred to by Creoles and colonial authorities as aborigines, natives, savages, naked barbarians, and many other kindred epithets".[22] Thus it may not be surprising that when both Mende and Temne chiefs revolted in 1898 during the so- called "Hut Tax Wars", they targeted Krio traders and settlers as well as British officials.[23]

One consequence of the violence experienced during the "Hut Tax Wars" was a growing British realization of the widespread animosity between the Krio and the numerically much larger ethnic groups in the interior, and as a result the British began to limit Krio political influence. Before the 1898 uprising, Krios had been appointed to positions of power throughout

[19] For a narrative account of the settling of the colony by freed slaves, including many who gained their freedom by fighting with the British during the American Revolution, see Schama (1995).

[20] This historical account closely follows Collier (1970) and Spitzer (1974).

[21] See Kandeh (1992) and Wyse (1989) for a discussion of this point.

[22] Spitzer (1974) presents draws on newspaper articles, speeches, and books from Krio scholars of the day to document the pervasive racism exhibited by Freetown Krios towards their "up-country" brethren.

[23] There were multiple origins of the "Hut Tax Wars" including both the imposition of an unpopular new tax in the Protectorate, as well as sharper limits on the internal slave trade; see Grace (1975).

the Protectorate, serving as "African Assistant District Commissioners" in many districts. However, because of growing ethnic tensions, they were not well received up-country; one colonial official at the time noted that "Freetown Creoles were worse than useless as Administrative Officers in the Sierra Leone Protectorate where they were both hated and despised" (Wyse, 1989, p. 27). Relations between the Krio and the British, too, began to deteriorate. In 1917, it became official policy to remove Krios from their limited positions of authority in the Protectorate, and during this period Wyse (1989) finds instances in which talented Krio were overlooked for local professional positions in the clergy and medicine in favor of less qualified British whites.

By 1924, the British allowed representatives from the Protectorate to have seats on the Sierra Leone Legislative Council. Three Paramount chiefs (two Mende and one Temne) were initially appointed to the Council, an event that provoked Krio outrage. Moreover, after the large railway strike of 1926, which was driven by Krio labor organizing, the Colonial Governor dissolved the Freetown City Council, the most important vehicle for Krio political interests (Wyse, 1989). The Krio objected to the growing strength of other ethnic groups in the colonial government well into the 1950s, but by then their influence had waned.

While ethnic divisions in sub-Saharan Africa have often been exacerbated by colonialism – the political rise of the favored minority Tutsi in Rwanda being perhaps the most notorious example – in Sierra Leone, the British took steps to curb Krio political power, at least temporarily preventing the dominance of one ethnic group over others. The country's two largest ethnic groups, the Mende and Temne that today dominate Sierra Leone numerically and politically, spent the colonial period united in their opposition to Krio dominance rather than battling each other for supremacy.[24]

One of the principal legacies of Sierra Leone's settlement by former slaves, and its long history as a slave trading outpost, is the language now called Krio, which is now believed to be spoken (mainly as a second language) by 95% of the population (Oyètádé and Luke, 2008). While its exact origins are uncertain[25], the popularity of the Krio language throughout Sierra Leone is clear. Speakers of the leading indigenous ethnic languages have adopted Krio,

[24] The political marginalization of the Krio is a striking contrast to the supremacy of their analogs in Liberia, the Americo-Liberians. Liberia was never colonized, but in 1822, the capital Monrovia was settled by former U.S. slaves. These individuals and their descendants dominated Liberian politics until they were overthrown in 1980. Recent political violence in Liberia is the result, at least in part, of resentments between Americo-Liberian elites and up-country tribes, divisions that were dampened in Sierra Leone by British policies marginalizing the Krio.

[25] Schama (1995) claims that Krio evolved from the language used by native (non-Krio) Sierra Leoneans to communicate with slave traders in the 16th and 17th centuries: "A pidgin English, much coloured with pidgin Portuguese, had been a lingua franca on the coast for at least a century since the slavers had first leased Bance Island" (Schama, 1995, p. 202). (Oyètádé and Luke, 2008) argue instead that it is closely related to the language spoken by Jamaican Maroons (descendents of escaped slaves), and was transplanted to Freetown when they resettled there. A related view is that Krio evolved as a language through which Freetown's disparate groups could communicate.

and Krio has had a major impact on spoken Mende and Temne as well as other languages. The widespread knowledge of Krio in Sierra Leone – despite the fact that the vast majority of adults in the country have no formal schooling – facilitates trade, communication and potentially cooperation across ethnic lines. That Krio is an indigenous language may help provide a common feeling of national identity.

The high degree of interethnic marriage in Sierra Leone, especially in urban areas (Davies, 2002), may also be an indication of favorable ethnic relations and historical interaction, while also potentially promoting inter-ethnic cooperation in the next generation. While large-scale statistical evidence on inter-marriage is limited, it is reinforced by suggestive genetic evidence. Jackson et al. (2005) study the nucleotide sequences of mitochondrial DNA in different ethnic groups and find no statistically significant differences between the sequences found in the Mende, Temne, and Loko groups (although there were some significant differences between these groups and the Limba). The lack of a detectable genetic difference between the country's two largest groups, the Mende and Temne, is especially noteworthy.[26]

3.F.2 Politics and Civil War in Post-Independence Sierra Leone, 1961-present

The major political parties in post-independence Sierra Leone have always had clear ethnic ties (Casey, 2009). The first two prime ministers, brothers Milton Margai (prime minister 1961-64) and Albert Margai (1964-67), were leaders of the Sierra Leone People's Party (SLPP) and members of the Mende ethnic group that dominates southern Sierra Leone. Albert Margai was a notoriously corrupt leader who, in attempting to intimidate opposition candidates from the largely northern African People's Congress (APC) in 1967 parliamentary elections, began to weaken the country's nascent democratic institutions.

The election winner, Siaka Stevens, an ethnic Limba (a northern group), survived a subsequent coup attempt organized by pro-Margai officers, and went on to dismantle all remaining democratic checks and balances. Sierra Leone became a one-party state in 1978, and Stevens is widely accused of plundering the country's resources for his own personal gain, while providing few public services (Reno, 1995). Stevens handed over power to his weak successor Joseph Momoh (another Limba) in 1985.

Sierra Leone's civil war started in 1991 and lasted until 2002. An estimated 50,000 people were killed, over half of the population was displaced from their homes, and thousands were victims of assaults, rapes and amputations (Human Rights Watch, 1999). Partially as a result of widespread discontentment with government corruption and ineffectiveness, a small group of rebels entering the country from Liberia in 1991 were successful in gaining recruits. As their numbers swelled in 1992, these rebels, known as the Revolutionary United Front

[26] Tishkoff et al. (2009) contains a detailed discussion of genetic diversity both within and across African populations, and documents the genetic signatures that characterize many African groups.

(RUF), spread the armed conflict throughout the country. Some scholars claim that the RUF's initial motivations were partly idealistic, and that they promoted an egalitarian non-ethnic national identity within the group (Richards, 1996). Another important factor in the RUF's rise was access to diamond wealth. Mining diamonds in Sierra Leone requires no machinery or technology since these alluvial stones sit close to the surface in dried riverbeds, and thus any group that controlled a diamond-rich area could extract and sell diamonds for considerable profits.

One feature of the war that has drawn attention was the frequent cooperation between the rebels and the Sierra Leone Army (SLA). These two groups often coordinated their movements to avoid direct battles, and at times worked out mutually beneficial profit-sharing arrangements in diamond areas. As a result, civilians were the main victims of the violence. For protection against RUF and SLA terror, many communities eventually organized local fighting groups that became known collectively as the Civil Defense Forces (CDF). CDF fighters were overwhelmingly civilians and relied primarily on local fundraising for supplies. While there were numerous manifestations throughout the country, the CDF's command and organization was often linked to traditional chiefs and secret religious societies. Neither the nationally organized RUF and SLA nor the locally organized CDF used ethnicity as a rallying cry.

Following the brutal 1999 rebel attack on Freetown, a deployment of United Kingdom and United Nations troops finally brought the war to an end. These foreign troops conducted a disarmament campaign and secured a peace treaty in early 2002. Donor and NGO assistance has since played a major role in reconstructing physical infrastructure, and resettling internally displaced people (almost all of whom had returned home by 2003). While an SLPP president ruled from 1996 through 2007, the APC candidate won the 2007 presidential election. While it is still too soon to know if stability has returned for good, the peaceful alternation of power suggests that democratic consolidation is occurring.

3.F.3 The Legacies of Colonial "Decentralized Despotism" and Slavery

British rule led to the strengthening of traditional chiefly authorities. These rulers had the explicit backing of British military might against any local challengers, dramatically bolstering their political standing relative to the pre-colonial period, provided they remained loyal to their British overlords. This authority translated into unchecked power and growing wealth for chiefs around Africa, and Sierra Leonean chiefs are perhaps the epitome of this tendency (Mamdani, 1996). Except in rare cases where villages are roughly equally split between two ethnicities and there are two village chiefs each representing their own ethnicity, village chiefs and Paramount chiefs have authority over both their own tribe and over other ethnic groups.

Paramount Chiefs in colonial Sierra Leone were the local executive, legislative and judicial authority. They had the power to fine, imprison, banish, and even kill, and their network of section chiefs and (male) elders stretched into every village in the country. Chiefs were also prominent in the domestic slave trade, which flourished in Sierra Leone legally until the late 1920s, and informally for decades afterwards. Powerful chiefs owned dozens of slaves, allowing them to plant vast tracts of farmland. Even after the formal end of slavery, Chiefs were able to press local youth to donate labor to their large farms. Chiefs also laid early claim to much of Sierra Leone's diamond wealth, which was being discovered mid-century, and to this day claim royalties on local diamond finds.

One of the more intriguing hypotheses about the origins of civil war in Sierra Leone is that the conflict had its roots in the legacy of the internal slave trade led by chiefs. Being one of the best natural harbors on Africa's western coastline, Freetown was for centuries a major Atlantic slave trade outpost, and Sierra Leone was long affected by slave raids tied to it. Although the Atlantic trade largely ended by the mid-19th century, local warlords continued carrying out slave raids in the region until the turn of the 20th century, especially in the Mano River area marking today's Sierra Leone-Liberia border.

The British colonial government's Protectorate Ordinance of 1896 attempted (at least nominally) to limit the internal slave trade, leading to outrage and formal protests among chiefs. This dissent soon gave way to violence in the Hut Tax War, and as a result, the colonial authorities in practice backed off their attempts to contain domestic slavery, hoping that the institution would eventually fade away. The institution lasted several more decades before it was finally outlawed in 1927. Yet Grace (1975) argues that, in practice, the formal legal ban on domestic slavery did little to change the social hierarchy and economic inequalities that existed between former masters and subjects. Richards (2005) similarly argues that much of the inequality in rural Sierra Leone today is a persistent effect of domestic slavery in the early 20th century, and views the Sierra Leone civil war as a sort of belated slave revolt, in which the descendants of slaves took up arms against the descendants of their masters. Perhaps as a result, the RUF explicitly targeted Chiefs for assassination during the war (Bellows and Miguel, 2009).

While the role of domestic slavery in the origins of the war is somewhat controversial, the arbitrary and undemocratic nature of the Chiefdom system, and the lack of voice for young men in particular, are widely held to have played a role in fueling the social discontent that contributed to the RUF uprising, There was, as a result, some public discussion after the civil war about major reforms to chieftancy institutions, but there have not been any meaningful changes since 2002. As discussed above, our survey data indicate that chiefs remain by far the most influential local authorities in rural Sierra Leone today.

3.F.4 Empirical Evidence on the Role of "Strong" Chiefs

A leading explanation for why ethnic diversity may not undermine public goods provision in rural Sierra Leone is the presence of a strong third-party enforcer, the traditional chiefly authorities. Habyarimana et al. (2007) find evidence in the lab for the importance of third-party enforcement in sustaining public goods provision in a Ugandan sample, echoing Fehr and Gächter (2000). In Sierra Leone, Chiefs have explicit responsibility for enforcing participation in public goods provision and can levy fines on free-riders. They also have responsibility for dealing with theft and disputes, which in turn can influence levels of trust.

In 2008, we surveyed every Paramount Chief in Sierra Leone, and collected information on age, tenure in office, and education. This allows us to use several different proxies for the political strength of chiefs in our analysis, both as stand-alone regressors and in interaction with ethnic diversity. Table 3.5 reports the mean effects results (for the same four NPS categories as above, in panels A-D) for chiefdom ELF, Paramount Chief tenure (years since the last election), whether or not the chief was an interim ruler in 2008 (ruling only until the position could be filled on a permanent bases through the traditional selection process)[27], and the interactions between ELF and Chief tenure, and ELF and interim status, on local public goods provision. Overall, we find no significant relationship with either Paramount Chief tenure or interim status and local collective action, or their interactions with ELF, for any of the four sets of local outcomes. We similarly investigated the relationship between village chief characteristics and outcomes at the EA level, and also fail to find any significant relationships (not shown).

These findings undercut the third-party enforcement theories advanced by Habyarimana et al. (2007), but two major caveats are worth keeping in mind. First, the proxies for Chief strength (tenure in office, interim status, and education) may be missing important dimensions of political influence, and this mismeasurement of actual influence could lead to attenuation bias towards zero. Second, to the extent that nearly all Sierra Leone chiefs – even the weakest ones – have sufficient authority to punish free-riders in ethnically diverse areas, chief strength impacts would not be apparent in the cross-section. That said, the findings in Table 3.5 suggest that the other factors discussed above, including historical inter-ethnic ties and a ubiquitous common language (Krio), are likely to be more important than strong chiefs in limiting the negative impacts of ethnic diversity on local outcomes in Sierra Leone.

3.G Conclusion

Sierra Leone is one of Africa's poorest countries and was devastated by over a decade of civil war. It does not, however, fit the stereotype of a country torn apart by tribal hatred, where different ethnic groups are unable to cooperate to provide public goods. When war came, it did not divide the country along ethnic (or religious) lines, and we show in this paper that ethnically diverse communities have levels of collective action and trust that are statistically indistinguishable from homogeneous communities. Many basic public goods are

[27] This information was collected from the local government ministry's official database of ruling chiefs. Summary statistics for these variables can be found in Appendix Table 3.A.12.

provided through local collective action and the outcomes that we study – road maintenance, communal labor, self-help groups, control of crime and school infrastructure – are important determinants of rural Sierra Leoneans' households' well-being and thus worthy objects of study.

The results hold when we address endogenous residential sorting by instrumenting for current ethnic fractionalization levels with historical levels, and restricting the sample to rural areas with stable ethnic composition since the colonial period. The civil war generated considerable migration and enables us to carefully examine the process of residential sorting. Our analysis of migration decisions demonstrates that many Sierra Leoneans have a strong preference to relocate to areas where co-ethnics also live. Importantly, the strength of co-ethnic residential preferences varies across individuals: educated people show more residential sorting towards ethnically diverse areas. To the extent that people with more education also have greater engagement in collective action (as our data suggest), this endogenous sorting could bias OLS estimates of the relationship between ethnic diversity and local collective action, confirming the usefulness of our novel IV approach.

The evidence that ethnicity plays a central role in migration decisions is an example of how Sierra Leone is far from being a post-ethnic society. The puzzle, therefore, is how ethnic identity can play such an important factor in decisions such as where to live and how to vote, but was not a leading factor in the conduct of the civil war nor the provision of local public goods. A positive interpretation is that it is possible to preserve strong ethnic identities and still achieve inter-ethnic cooperation, perhaps because the common bonds of language and national identity are stronger than the centripetal pull of tribe. We discuss how historical factors may have contributed to this result, for example through the spread of a lingua franca (Krio) that is unique to Sierra Leone yet not the first language of either of the country's two largest and most politically powerful ethnic groups (the Mende and Temne). Another potentially important factor is the colonial legacy of cooperation between these two groups against a common foe, the once- dominant Krio settler community who are now numerically and politically inconsequential. Genetic similarity and extensive intermarriage may also help solidify cross-group affinity.

The alternative, and less sanguine, explanation is that rural Sierra Leone communities overcome ethnic fractionalization because of the vice-like grip of traditional chiefs, who regulate collective action in rural areas and levy heavily fines those who do not take part. Yet we do not find that proxies for Chief strength affect local public goods, casting doubt on the importance of this potential mechanism.

Scholars have now identified several African cases where high levels of ethnic diversity do not impede successful local collective action. By learning from such cases, we hope to generate insight into how to address ethnic divisions in other societies where they remain a concern. In this regard, the story that emerges from Sierra Leone is different in important respects from others described in the literature. Like Tanzanians, Sierra Leoneans are bound together by a common national language that they strongly feel is theirs, yet the two countries differ fundamentally in their local and national institutions and how these interact. In contrast to Tanzania, the high level of interethnic cooperation in Sierra Leone is not the result of a modernizing approach that dismantled chiefdom authorities and replaced them with elected local institutions. Unlike in Zambia, successful local collective action across diverse ethnic groups is maintained in Sierra Leone even when the groups are national political rivals.

While it is difficult – and potentially unwise – to draw general conclusions about how to achieve inter-ethnic cooperation in a continent as diverse as Africa, Sierra Leone provides evidence that ethnic differences can be highly salient in some aspects of life and yet not undermine local public goods provision, an encouraging message for other diverse societies. It also provides a stark counterexample to the view that underdevelopment in Africa is inextricably connected to tribal conflict. Looking forward, it is still possible that the post-war transition to democracy, with tightly contested recent national elections fought largely along ethnic lines, will increasingly exacerbate ethnic tensions in Sierra Leone (consistent with the findings in Eifert et al., 2010), perhaps gradually undermining the cooperation documented in this paper. More optimistically, the strong local inter-ethnic cooperation that we document may continue to provide a robust bulwark against the exploitation of ethnic divisions by national politicians.

Table 3.1: Migration across chiefdoms (1990 to 2007) and ethnic composition (conditional logit)

	(1)	(2)	(3)	(4)
Distance between chiefdoms	-0.021 (0.001)***	-0.013 (0.003)***	-0.024 (0.001)***	-0.014 (0.003)***
Co-ethnic population share	2.184 (0.107)***	2.225 (0.260)***		
Ethnolinguistic fractionalization (ELF)			1.504 (0.092)***	2.277 (0.274)***
Any education × Distance		0.009 (0.003)***		0.009 (0.003)***
Any education × Co-ethnic		-1.597 (0.220)***		
Any education × ELF				2.498 (0.240)***
Experienced war violence × Distance		-0.056 (0.008)***		-0.063 (0.009)***
Experienced war violence × Co-ethnic		1.959 (0.585)***		
Experienced war violence × ELF				-4.995 (0.731)***
Ruling family member × Distance		0.003 (0.002)***		0.003 (0.003)
Ruling family member × Co-ethnic		0.392 (0.246)**		
Ruling family member × ELF				-0.519 (0.272)*
Chiefdom population (1985)	11.615 (0.258)***	11.886 (0.258)***	9.675 (0.268)***	9.016 (0.272)***
Chiefdom population Density (1985)	-0.009 (0.001)***	-0.009 (0.001)***	-0.008 (0.001)***	-0.007 (0.001)***
Distance to a road	-0.053 (0.007)***	-0.049 (0.007)***	-0.04 (0.006)***	0.224 (0.006)***
Distance to a city	-0.813 (0.020)***	-0.833 (0.021)***	-0.652 (0.018)***	-0.654 (0.006)***
Attacks and battles in the civil war	0.015 (0.002)***	0.014 (0.002)***	0.016 (0.002)***	0.014 (0.003)***
Any Mining in chiefdom	-0.012 (0.004)***	-0.014 (0.004)***	-0.025 (0.003)***	-0.025 (0.003)***
Log Pseudolikelihood	-1.314	-1.213	-1.366	-1.238
Pseudo R^2	0.772	0.789	0.763	0.785
Number of Individuals	5488	5488	5488	5488
Number of chiefdoms/locations	154	154	154	154

Notes: Estimation computed on a conditional logit model using weighted maximum likelihood, which addresses the endogenous stratification problem (see Appendix A). */**/*** denotes significantly different from zero at 90/95/99% confidence. Distances are measured in km between centroids. Chiefdom population is measured in thousands. Any education is an indicator variable for any schooling.

Table 3.2: First Stage Regressions

Panel A: Chiefdom-level analysis

	Dependent Variable				
	ELF	**% Mende**	**% Temne**	**(% Mende)^2**	**(% Temne)^2**
ELF (1963)	0.797 (0.089)***	-0.064 (0.077)	0.091 (0.050)*	-0.011 (0.085)	0.05 (0.028)*
% Mende (1963)	-0.817 (0.199)***	2.081 (0.191)***	-0.297 (0.088)***	1.436 (0.254)***	-0.151 (0.053)***
% Temne (1963)	0.161 (0.192)	-0.169 (0.148)	1.055 (0.104)***	-0.302 (0.203)	0.223 (0.088)**
(% Mende)^2 (1963)	0.807 (0.232)***	-1.204 (0.231)***	0.315 (0.099)***	-0.566 (0.300)*	0.165 (0.059)***
(% Temne)^2 (1963)	-0.261 (0.214)	0.152 (0.167)	-0.052 (0.112)	0.313 (0.226)*	0.788 (0.091)***
N (chiefdoms)	146	146	146	146	146
R^2	0.668	0.940	0.985	0.886	0.987

Panel B: Enumeration area-level analysis (NPS EAs only)

	Dependent Variable				
	ELF	**% Mende**	**% Temne**	**(% Mende)^2**	**(% Temne)^2**
ELF (1963)	0.429 (0.110)***	-0.028 (0.060)	0.117 (0.064)*	0.006 (0.064)	0.076 (0.051)
% Mende (1963)	-0.139 (0.184)	2.033 (0.161)***	-0.286 (0.129)**	1.607 (0.187)***	-0.269 (0.106)**
% Temne (1963)	0.094 (0.217)	-0.193 (0.134)	0.941 (0.211)***	-0.311 (0.156)**	0.544 (0.198)***
(% Mende)^2 (1963)	0.157 (0.208)	-1.127 (0.185)***	0.324 (0.141)**	-0.723 (0.212)***	0.298 (0.117)**
(% Temne)^2 (1963)	-0.086 (0.241)	0.187 (0.150)	0.09 (0.229)	0.318 (0.173)*	0.454 (0.214)**
N (EAs)	444	444	444	444	444
R^2	0.179	0.917	0.894	0.866	0.879

Notes: OLS regressions, robust standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence.

Table 3.3: Ethnic diversity and road maintenance (brushing) across chiefdoms

	OLS regressions			IV regressions		
	(1)	(2)	(3)	(4)	(5)	(6)
Ethnolinguistic fractionalization (*ELF*)	0.05 (0.159)	0.052 (0.146)	0.041 (0.179)	0.032 (0.218)	-0.083 (0.192)	-0.224 (0.330)
Civil war victimization index		0.289 (0.098)***	0.283 (0.112)**		0.338 (0.096)***	0.291 (0.123)***
Female respondent share		-0.509 (0.313)	-0.509 (0.314)		-0.464 (0.309)	-0.457 (0.305)
Youth (age 16-35) respondent share		-0.154 (0.216)	-0.153 (0.218)		-0.081 (0.205)	-0.070 (0.210)
Middle aged (age 36-50) respondent share		-0.31 (0.193)	-0.308 (0.199)		-0.251 (0.188)	-0.226 (0.203)
Muslim share		0.169 (0.074)**	0.169 (0.074)**		0.159 (0.074)**	0.158 (0.075)**
Any education share		0.481 (0.165)***	0.474 (0.186)**		0.473 (0.173)***	0.402 (0.187)**
Average socioeconomic status index		-0.263 (0.202)	-0.264 (0.202)		-0.301 (0.211)	-0.298 (0.212)
Community leader respondent share		0.172 (0.120)	0.172 (0.121)		0.171 (0.119)	0.166 (0.123)
Civil war victimization index × *ELF*			0.04 (0.384)			0.378 (0.586)
N (chiefdoms)	146	146	146	146	146	146
R^2	0.016	0.208	0.208	0.007	0.181	0.175

Notes: Robust standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. Shares for Mende, Temne, and their squares are included in the specification but coefficient estimates are not shown. The instrumental variables are listed in Table 3.2. All regressions are estimated with survey weights, where each chiefdom observation is weighted by the inverse of its sampling probability.

Table 3.4: Ethnic diversity and local outcomes: mean effects analysis

Panel A: Chiefdom-level analysis

	OLS regressions		IV regressions	
	(1)	**(2)**	**(3)**	**(4)**
Collective Action mean effect	0.221 (0.682)	0.165 (0.630)	-0.175 (1.025)	-0.525 (0.850)
Group Membership mean effect	0.458 (0.515)	0.256 (0.334)	0.124 (0.780)	-0.395 (0.429)
Disputes mean effect	0.494 (0.575)	0.461 (0.558)	0.646 (0.708)	0.647 (0.691)
Trust mean effect	0.474 (0.356)	0.228 (0.341)	0.274 (0.361)	-0.146 (0.348)
School Supplies mean effect	-0.441 (0.452)	-0.078 (0.483)	-0.638 (0.603)	-0.219 -0.642
Teaching Quality mean effect	0.499 (0.302)	0.423 (0.321)	0.192 (0.372)	0.084 (0.401)
School Building Quality mean effect	0.135 (0.440)	-0.250 (0.410)	0.109 (0.527)	-0.326 (0.505)
Regression controls	No	Yes	No	Yes
Number of Chiefdoms	146	146	146	146

Panel B: Enumeration area-level analysis (NPS EAs only)

	OLS regressions		IV regressions	
	(1)	**(2)**	**(3)**	**(4)**
Collective Action mean effect	0.346 (0.373)	0.050 (0.397)	0.784 (1.193)	0.312 (1.226)
Group Membership mean effect	0.469 (0.320)	-0.032 (0.228)	0.718 (0.921)	-0.570 (0.698)
Disputes mean effect	0.790 (0.387)**	0.471 (0.345)	0.931 (1.043)	0.566 (1.070)
Trust mean effect	0.043 (0.187)	-0.139 (0.196)	0.075 (0.616)	-0.415 (0.682)
Regression controls	No	Yes	No	Yes
Number of Chiefdoms	444	444	444	444

Notes: Each entry is the coefficient estimate on ethnolinguistic fractionalization (ELF) from a separate regression. Standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. See Appendix 3.D for details on the mean effects analysis. The instrumental variables are listed in Table 3.2. The regression controls are like those in Table 3.3, columns 2 (OLS) and 5 (IV). All regressions are estimated with survey weights, where each observation is weighted by the inverse of its sampling probability.

The components of the "Collective Action" category are participation in road brushing, community labor, and community meetings. The components of the "Group Membership" category are member of any community group, a credit group, and a school group. The components of the "Disputes" category are the incidence of any local assault dispute, land dispute, or dispute involving theft. The components of the "Trust" category include trust of people in own community, people outside community, local councilors, and the central government. Descriptive statistics for these outcomes are in Appendix Table 3.A.2. The components of the School supplies category are the average number of desks per student, chairs per student, benches per student, blackboards per student, and textbooks per student. The components of the "Teaching Quality" category are the teacher / student ratio, the percentage of teachers present during surprise visit, and the percentage of teachers actually working during surprise visit. The components of the "School Building Quality" category are the percentage of schools with toilets, with electricity, with piped water, and with sturdy buildings. Descriptive statistics are presented in Appendix Table A3.

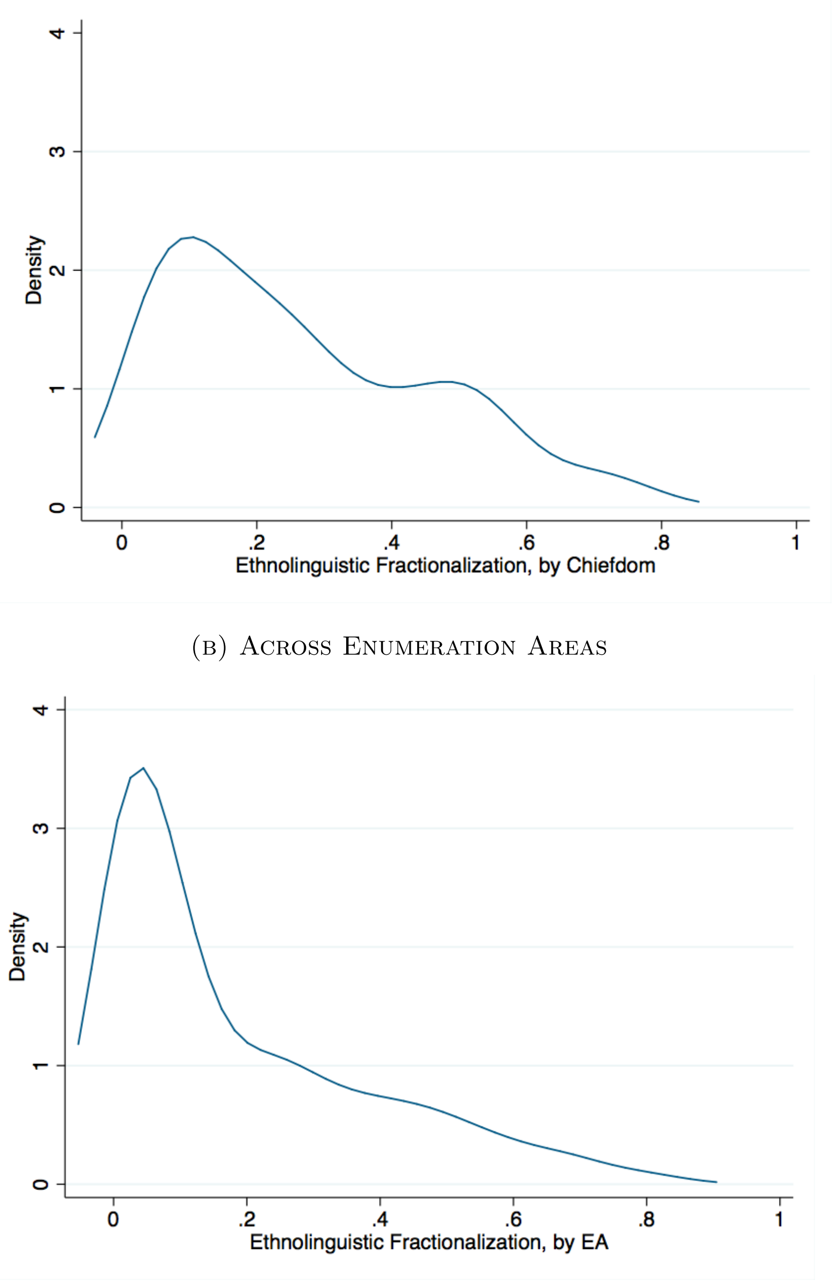
Table 3.5: Paramount chief characteristics and ethnic diversity

	(1)	**(2)**	**(3)**
Panel A: Dep. var: Collective Action mean effect			
ELF	-0.525	-0.792	-0.175
	(0.850)	(0.942)	(0.899)
Paramount Chief Tenure (in years)		-0.011	
		(0.015)	
ELF × Paramount Chief Tenure		0.055	
		(0.045)	
Interim Paramount Chief			0.208
			(0.266)
ELF × Interim Paramount Chief			-0.967
			(0.878)
Panel B: Dep. Var.: Group Membership mean effect			
ELF	-0.395	-0.671	-0.054
	(0.429)	(0.492)	(0.418)
Paramount Chief Tenure (in years)		-0.02	
		(0.009)*	
ELF × Paramount Chief Tenure		0.041	
		(0.026)	
Interim Paramount Chief			0.282
			(0.162)
ELF × Interim Paramount Chief			-0.798
			(0.511)
Dep. Var.: Disputes mean effect			
ELF	0.647	1.089	0.865
	(0.691)	(0.715)	(0.681)
Paramount Chief Tenure (in years)		0.012	
		(0.013)	
ELF × Paramount Chief Tenure		-0.067	
		(0.039)	
Interim Paramount Chief			0.292
			(0.223)
ELF × Interim Paramount Chief			-0.304
			(0.793)
Panel D: Dep. Var.: Trust mean effect			
ELF	-0.146	-0.348	0.122
	(0.348)	(0.407)	(0.360)
Paramount Chief Tenure (in years)		-0.017	
		(0.011)	
ELF × Paramount Chief Tenure		0.029	
		(0.033)	
Interim Paramount Chief			0.263
			(0.156)
ELF × Interim Paramount Chief			-0.514
			(0.457)

Notes: Standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. See Appendix 3.D for details on the mean effects analysis. All columns are IV specifications, and the instrumental variables are listed in Table 3.2. Regression controls like those in Table 3.3, column 5 (IV) are included in all specifications. N = 146 chiefdoms for all specifications. All regressions are estimated with survey weights, where each chiefdom observation is weighted by the inverse of its sampling probability.

The components of the "Collective Action", "Group Membership", "Disputes", and "Trust" categories and the descriptive statistics are in Appendix Table 3.A.4. The summary statistics for the Paramount Chief characteristics are presented in Appendix Table 3.A.12.

Figure 3.1: Ethno-linguistic fractionalization in Sierra Leone (non-parametric densities)

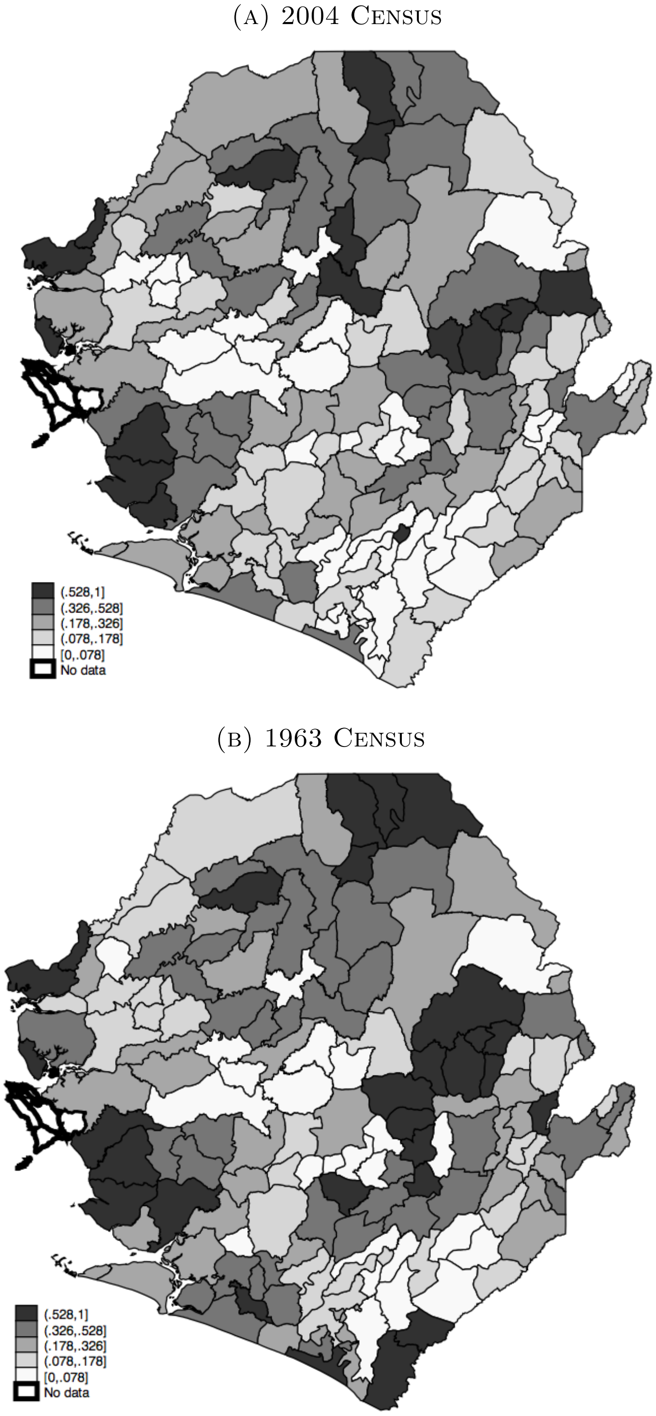


(a) Across Chiefdoms

(b) Across Enumeration Areas

Notes: The data source for both panels is the 2004 Population Census. Both use a Gaussian kernel with bandwidth set to minimize integrated mean squared error. The mean of ELF across chiefdoms (panel A) is 0.264, with a standard deviation of 0.196. The mean of ELF across EAs (panel B) is 0.185, with a standard deviation of 0.199.

Figure 3.2: Ethnic Diversity by Chiefdom

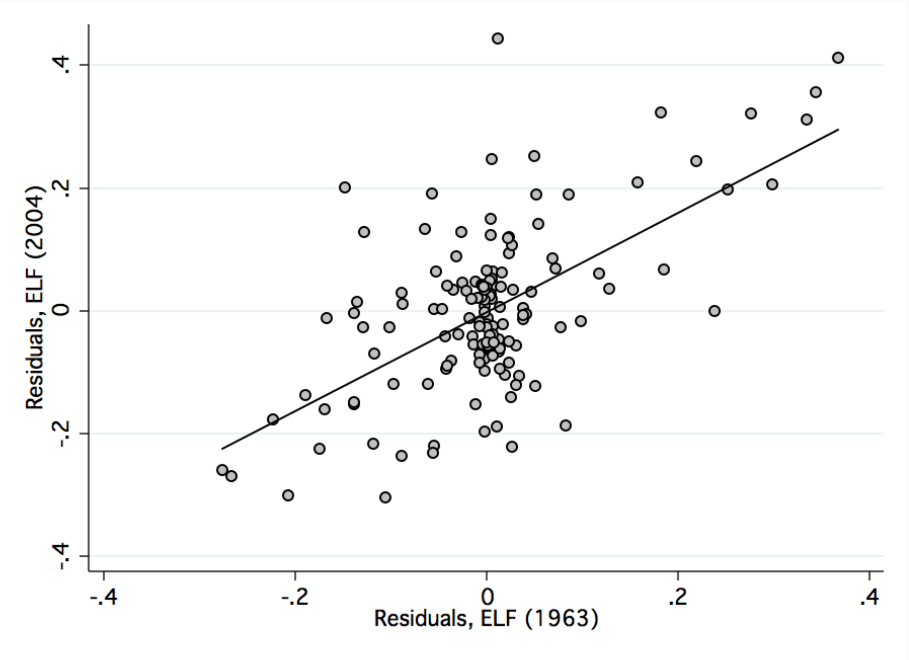


(a) 2004 Census

(b) 1963 Census

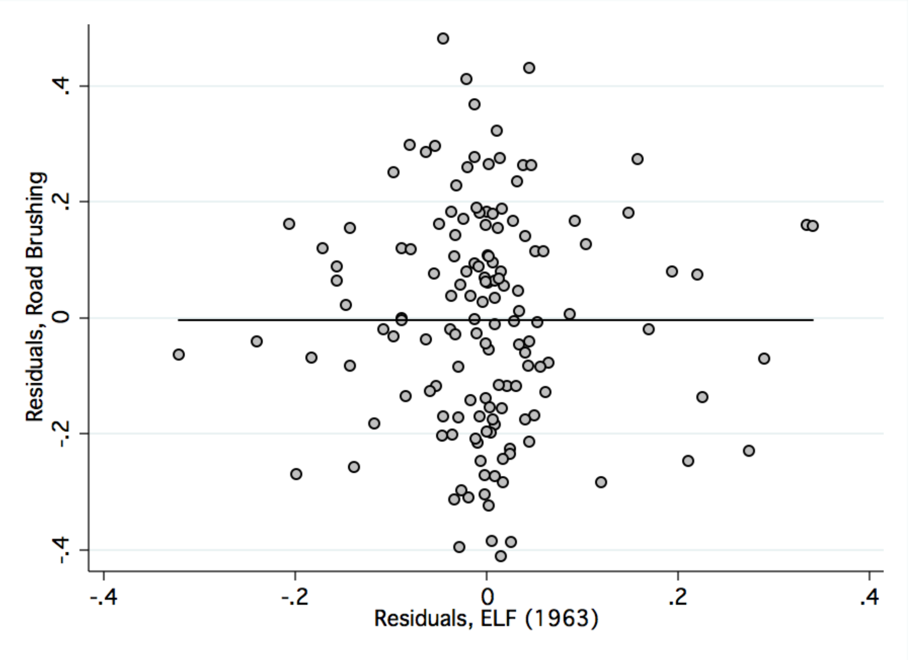
Notes: The mean of ELF across chiefdoms in 2004 (Panel A) is 0.264, with a standard deviation of 0.195. and in 1963 (Panel B) is 0.304, with a standard deviation of 0.205.

Figure 3.3: Chiefdom ethno-linguistic fractionalization in 2004 versus ethno-linguistic fractionalization 1963 (residual plot)



Notes: This figure is a residual-on-residual plot, a graphical representation of our first stage. The y-axis displays residuals from a regression of 2004 ELF on 1963 ethnic share controls. The x-axis plots residuals from a regression of 1963 ELF on 1963 ethnic share controls. The regression fit corresponds to Table 2, column 1 (panel A).

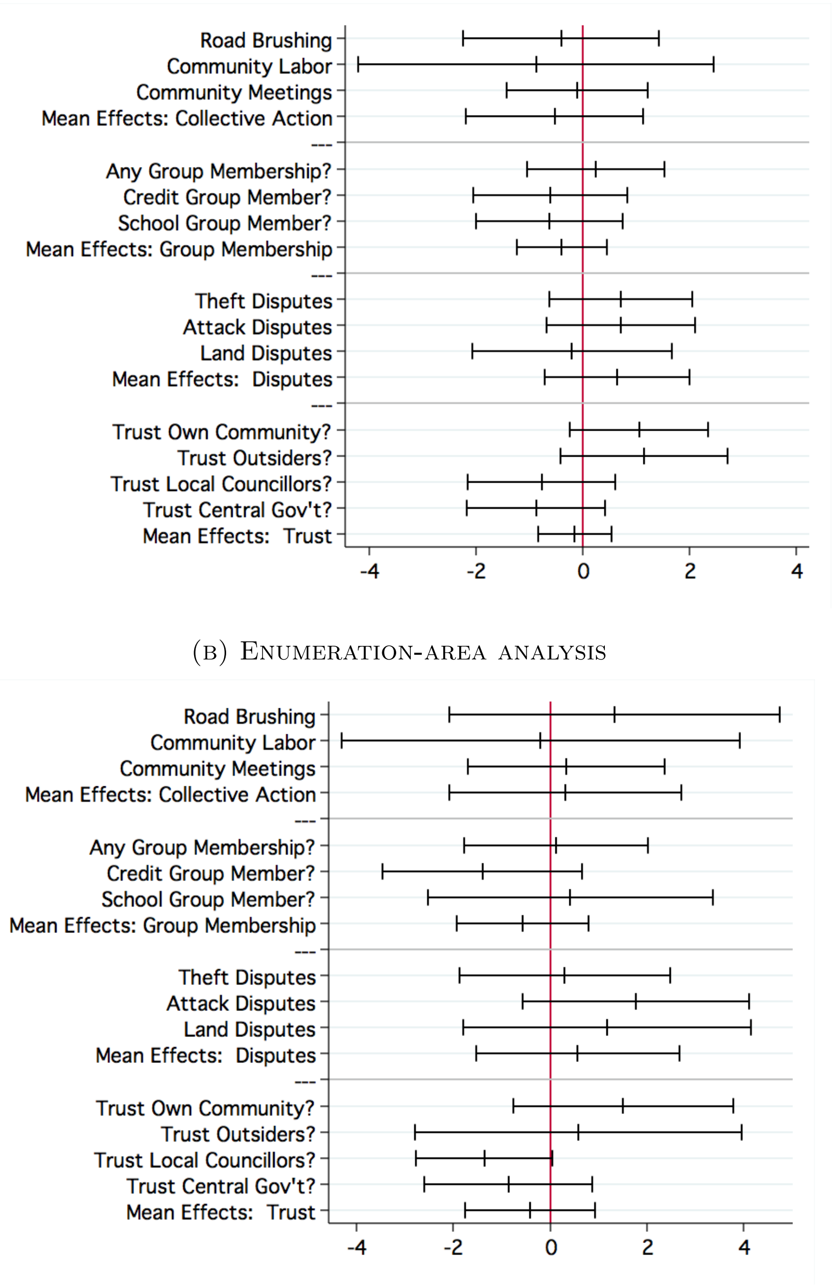
Figure 3.4: Chiefdom road maintenance participation in 2007 versus ethnolinguistic fractionalization in 1963 (residual plot)



Notes: This figure is a residual-on-residual plot, a graphical representation of the reduced form. The y-axis displays residuals from a regression of road maintenance on 1963 ethnic share controls and other controls from Table 3, column 5. The x-axis plots residuals from a regression of 1963 ELF on 1963 ethnic share controls and other controls from Table 3, column 5.

Figure 3.5: Point Estimates and 95% Confidence Intervals for the Effects of Ethnic Diversity on Local Outcomes, Pooled 2005 and 2007

(a) Chiefdom-level analysis



(b) Enumeration-area analysis

Notes: Dependent variables were standardized before regressions to make confidence intervals more comparable. Individual estimates and confidence intervals taken from IV specifications with full controls, analogous to Table 3, column 5. Mean effects are produced in Table 4, Column 4. All regressions are estimated with survey weights, where each chiefdom (or enumeration area) observation is weighted by the inverse of its sampling probability.

3.A Appendix Tables and Figures

Table 3.A.1: Ethnic Population Shares in Sierra Leone

Ethnic Group (tribe)	**1963 census**	**2004 census**
Mende	0.309	0.322
Temne	0.298	0.318
Limba	0.084	0.083
Kono	0.048	0.044
Kuranko	0.037	0.041
Sherbro	0.034	0.023
Fullah	0.031	0.037
Susu	0.031	0.029
Lokko	0.030	0.026
Kissi	0.022	0.025
Madingo	0.023	0.024
Krio	0.019	0.014
Yalunka	0.007	0.007
Krim	0.004	0.002
Vai	0.003	0.001
Other	0.021	0.006

Notes: Estimation computed on a conditional logit model using weighted maximum likelihood, which addresses the endogenous stratification problem (see Appendix A). */**/*** denotes significantly different from zero at 90/95/99% confidence. Distances are measured in km between centroids. Chiefdom population is measured in thousands. Any education is an indicator variable for any schooling.

Table 3.A.2: Additional descriptive statistics for migration analysis (individual sample)

	N	**Mean (SD)**
Moved between chiefdoms/locations (1990 to 2007)	5488	0.265 (0.442)
Moved between districts (1990 to 2007)	5488	0.165 (0.371)
Distance moved (in kilometers)	5488	19.717 (48.365)
Distance moved (in kilometers), if moved	1457	74.267 (69.002)
Any education	5488	0.372 (0.483)
Ruling family member	5488	0.262 (0.440)
Any member of 1990 HH made a refugee (left country)?	5488	0.232 (0.422)

Notes: Source NPS 2007 Survey.

Table 3.A.3: Ethnic Diversity Regressions, 1963

	Dependent variable: ***ELF* (from 1963)**
% of population literate (1963)	1.523 (0.936)
% of population formally employed (1963)	-0.207 (0.144)
N (chiefdoms)	142
R^2	0.028

Note: OLS regressions, robust standard errors in parentheses. Dependent variables in column header. */**/*** denotes significantly different from zero at 90/95/99% confidence. The estimated constant term is not shown.

Table 3.A.4: Descriptive statistics for local outcomes, 2005 and 2007 NPS surveys

	2005 NPS	2007 NPS
Collective Action	Mean (SD)	Mean (SD)
Participation in road brushing	.	0.396 (0.209)
Participation in community labor (farm, school)	.	0.185 (0.145)
Participation in community meetings	0.766 (0.143)	0.422 (0.214)
Group Membership		
Member of any community group	0.873 (0.151)	0.811 (0.125)
Member of a credit group	0.159 (0.131)	0.159 (0.107)
Member of a school group	0.209 (0.160)	0.220 (0.154)
Control of Disputes		
Any local assault disputes	0.021 (0.037)	0.044 (0.060)
Any local land disputes	.	0.038 (0.049)
Any local dispute involving theft	0.271 (0.169)	0.053 (0.060)
Trust		
Trust of people in own community (index)	0.906 (0.098)	0.780 (0.178)
Trust of people outside community (index)	0.479 (0.187)	0.386 (0.175)
Trust of local councilors (index)	0.641 (0.165)	0.285 (0.163)
Trust of the central government (index)	0.630 (0.169)	0.346 (0.189)
Regression Controls		
Youth (ages 16-35) respondent share	0.415 (0.118)	0.353 (0.114)
Middle aged (ages 36-50) respondent share	0.340 (0.121)	0.375 (0.122)
Female respondent share	0.491 (0.025)	0.494 (0.065)
Muslim	0.798 (0.232)	0.785 (0.236)
Any education share	0.264 (0.142)	0.230 (0.132)
Community leader respondent share	0.509 (0.183)	0.493 (0.152)
Average socioeconomic status index	0.205 (0.074)	0.223 (0.092)
Civil war victimization index	0.406 (0.182)	0.429 (0.161)

Notes: Source 2005 and 2007 NPS Surveys. $N = 146$ chiefdoms. Standard deviations in parentheses. The Civil war victimization index is the average across three indicators: "Were any members of your HH killed?", "Were any members of your HH injured/maimed?", and "Were any members of your HH made refugees?". The Ethnic minority share refers to being a minority in that chiefdom. The Average socioeconomic status is an index composed of having a wage paying job, durables ownership, and the household water source.

There were some important changes in question wording across the 2005 and 2007 survey rounds that can explain changes in survey response patterns over time. The "Participation in community meetings" question was asked for the past one year in 2005, and for the past month in 2007, and this likely explains the higher mean attendance rate reported in 2005. The 2005 control of disputes questions were significantly more detailed (in separately prompting for specific types of theft, i.e., of livestock, household items, etc.) than the 2007 questions, and this likely explains the higher survey means in 2005.The wording of the trust questions also changed between 2005 and 2007, with the 2005 questions having a more set of possible responses (ranging from 1 to 5) and the 2007 questions restricted to an indicator variable. While the means of these variables change considerably across rounds, it is still appropriate to group these variables together in the mean effect analysis, since all variables are demeaned and normalized.

Table 3.A.5: Descriptive statistics for school quality outcomes

	2005 School Survey
School Supplies	**Mean (SD)**
Average number of desks per student	0.306 (0.286)
Average number of chairs per student	0.184 (0.412)
Average number of benches per student	0.297 (0.248)
Average number of blackboards per student	0.044 (0.036)
Average number of textbooks per student	0.899 (1.028)
Teaching Quality	
Teacher / student ratio	0.035 (0.033)
Percentage of teachers present during surprise visit	0.608 (0.251)
Percentage of teachers actually working during surprise visit (conditional on being present at school)	0.799 (0.252)
School Building Quality	
Percentage of schools with toilets	0.632 (0.484)
Percentage of schools with electricity	0.006 (0.076)
Percentage of schools with piped water	0.092 (0.290)
Percentage of schools with sturdy buildings	0.410 (0.455)

Notes: Source 2005 Primary School Surveys. Standard deviations in parentheses. $N = 146$ chiefdoms.

Table 3.A.6: Ethnic diversity and road maintenance (brushing) across enumeration areas

	OLS regressions			**IV regressions**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Ethnolinguistic fractionalization (ELF)	0.075 (0.122)	-0.033 (0.132)	-0.030 (0.247)	0.457 (0.435)	0.317 (0.414)	1.972 (4.857)
Civil war victimization index		0.169 (0.082)**	0.170 (0.096)*		0.140 (0.084)*	0.539 (1.093)
Female respondent share		-0.247 (0.110)**	-0.247 (0.110)**		-0.253 (0.105)**	-0.232 (0.115)*
Youth (ages 16-35) respondent share		0.184 (0.090)**	0.184 (0.089)**		0.174 (0.089)*	0.159 (0.109)
Middle aged (ages 36-50) respondent share		0.014 (0.090)	0.014 (0.090)		0.004 (0.092)	-0.018 (0.108)
Muslim share		0.040 (0.055)	0.040 (0.055)		0.004 (0.063)	-0.034 (0.147)
Any education share		0.100 (0.089	0.100 (0.089)		0.056 (0.126)	0.048 (0.155)
Average socioeconomic status index		0.012 (0.102)	0.012 (0.103)		-0.019 (0.108)	-0.022 (0.127)
Community leader respondent share		0.211 (0.058)***	0.211 (0.058)***		0.21 (0.058)***	0.23 (0.092)***
Civil war victimization index × *ELF*			-0.006 (0.397)			-2.55 (6.948)
N (EAs)	444	444	444	444	444	444
R^2	0.02	0.10	0.10	0.00	0.08	0.00

Notes: Robust standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. Shares for Mende, Temne, and their squares are included in the specification but coefficient estimates are not shown. The instrumental variables are listed in Table 3.2. All regressions are estimated with survey weights, where each chiefdom observation is weighted by the inverse of its sampling probability.

Table 3.A.7: Ethnic diversity and road brushing (maintenance), individual-level analysis

	OLS regressions			**IV regressions**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Ethnolinguistic fractionalization (ELF)	0.077 (0.122)	-0.016 (0.131)	-0.018 (0.244)	0.461 (0.434)	0.221 (0.398)	0.923 (4.290)
Civil war victimization index, household		0.176 (0.080)**	0.176 (0.097)*		0.150 (0.084)	0.324 (0.992)
Female respondent		-0.200 (0.017)***	-0.200 (0.017)***		-0.201 (0.017)***	-0.201 (0.016)***
Youth (ages 16-35) respondent		0.153 (0.020)***	0.153 (0.020)***		0.152 (0.020)***	0.152 (0.021)***
Middle aged (ages 36-50) respondent		0.093 (0.018)***	0.093 (0.018)***		0.092 (0.018)***	0.091 (0.019)***
Muslim respondent		0.019 (0.027)	0.019 (0.027)		0.010 (0.030)	0.004 (0.047)
Any education indicator		0.030 (0.023)	0.030 (0.023)		0.026 (0.026)	0.026 (0.027)
Socioeconomic status index, household		0.049 (0.048)	0.049 (0.048)		0.038 (0.061)	0.035 (0.074)
Community leader indicator (respondent)		0.088 (0.019)***	0.088 (0.019)***		0.088 (0.019)***	0.090 (0.023)***
Civil war victimization index × ELF			0.003 (0.407)			-1.100 (6.257)
N (individuals)	4414	4318	4318	4386	4318	4318
R^2	0.00	0.08	0.08	0.00	0.07	0.07

Notes: Robust standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. Shares for Mende, Temne, and their squares are included in the specification but coefficient estimates are not shown. The instrumental variables are listed in Table 3.2. All regressions are estimated with survey weights, where each chiefdom observation is weighted by the inverse of its sampling probability.

Table 3.A.8: Language family diversity and local outcomes, mean effects analysis

Panel A: Chiefdom-level analysis

	OLS regressions		IV regressions	
	(1)	**(2)**	**(3)**	**(4)**
Collective Action mean effect	-0.300	-0.172	-0.432	-0.252
	(0.432)	(0.368)	(0.535)	(0.530)
Group Membership mean effect	0.589	0.847	0.018	0.415
	(0.305)*	(0.208)**	(0.401)	(0.322)
Disputes mean effect	0.128	0.209	0.452	0.579
	(0.470)	(0.466)	(0.510)	(0.464)
Trust mean effect	0.203	-0.185	-0.886	-0.944
	(0.366)	(0.373)	(0.374)*	(0.357)**
Regression controls	No	Yes	No	Yes
Number of Chiefdoms	146	146	146	146

Panel B: Enumeration-area analysis

	OLS regressions		IV regressions	
	(1)	**(2)**	**(3)**	**(4)**
Collective Action mean effect	-0.235	-0.274	0.970	0.538
	(0.306)	(0.306)	(1.878)	(1.640)
Group Membership mean effect	0.706	0.534	0.905	1.631
	(0.239)**	(0.187)**	(1.014)	(1.033)
Disputes mean effect	0.932	0.574	-0.232	0.470
	(0.276)**	(0.278)**	(1.511)	(0.784)
Trust mean effect	0.559	-0.494	-2.277	-1.930
	(0.156)**	(0.170)**	(1.124)	(0.846)**
Regression controls	No	Yes	No	Yes
Number of Chiefdoms	444	444	444	444

Notes: Each entry is the coefficient estimate on the language family diversity index from a separate regression. Standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. See Appendix 3.D for details on the mean effects analysis. The instrumental variables are analogously defined conflict indices, using the 1963 data. The regression controls are like those in Table 3.5, columns 2 (OLS) and 5 (IV). All regressions are estimated with survey weights, where each chiefdom (or enumeration area) observation is weighted by the inverse of its sampling probability.

The components of the "Collective Action" category are participation in road brushing, participation in community labor, and participation in community meetings. The components of the "Group Membership" category are member of any community group, member of a credit group, and member of a school group. The components of the "Disputes" category are the incidence of any local assault dispute, any local land disputes, or any local dispute involving theft. The components of the "Trust" category include trust of people in own community (index), trust of people outside community (index), trust of local councilors (index), and trust of the central government (index). Descriptive statistics for each of these outcomes are presented in Appendix Table 3.A.4.

Table 3.A.9: Historical Ethnic Conflict Matrix: 1500-1800

	Fullah	Kissi	Kono	Kuranko	Krim	Krio	Limba	Loko	Mandingo	Mende	Other	Sherbro	Susu	Temne	Vai	Yalunka
Fullah	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1
Kissi	0	0	1	1	0	1	1	0	1	0	0	0	0	0	1	0
Kono	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0
Kuranko	0	1	0	0	0	1	0	1	1	0	0	0	1	0	0	1
Krim	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0
Krio	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
Limba	0	1	0	0	0	1	0	1	1	1	0	0	0	1	0	0
Loko	0	0	0	1	0	1	1	0	1	0	0	0	1	1	0	0
Mandingo	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Mende	0	0	0	0	0	1	1	0	1	0	1	1	0	1	0	0
Other	0	0	1	0	1	1	0	0	1	1	1	0	0	1	1	0
Sherbro	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
Susu	1	0	0	1	0	1	0	1	1	0	0	0	0	1	0	0
Temne	0	0	0	0	0	1	1	1	1	1	1	0	1	0	0	0
Vai	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0
Yalunka	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0

Notes: Kup (1961) and authors calculations. Each entry in this table equal to one if the two ethnic groups have had historical conflict and is zero otherwise. In constructing this matrix, Kup (1961) provides data for all conflicts between 1460 (when the first European explorers came to Sierra Leone) and 1787 (when Sierra Leone was first colonized). From our reading of history, we assumed that the Krio were in conflict with all other groups. We also assumed that the Mandingo were in conflict with all other groups because they were notorious slave raiders, despite the fact that they are not mentioned frequently in the Kup (1961) text. Results are not sensitive to either of the latter two assumptions.

Table 3.A.10: Historical conflict index and local outcomes, mean effects analysis

Panel A: Chiefdom-level analysis

	OLS regressions		**IV regressions**	
	(1)	**(2)**	**(3)**	**(4)**
Collective Action mean effect	-1.052 (0.478)**	-0.578 (0.448)	-1.667 (0.764)*	-1.229 (0.825)
Group Membership mean effect	-0.552 (0.408)	0.292 (0.362)	-0.134 (0.777)	0.137 (0.561)
Disputes mean effect	-0.721 (0.554)	-0.689 (0.619)	1.058 (0.847)	1.069 (0.745)
Trust mean effect	0.995 (0.392)**	0.985 (0.432)**	1.095 (0.514)*	0.915 (0.456)*
Regression controls	No	Yes	No	Yes
Number of Chiefdoms	146	146	146	146

Panel B: Enumeration-area analysis

	OLS regressions		**IV regressions**	
	(1)	**(2)**	**(3)**	**(4)**
Collective Action mean effect	-0.573 (0.424)	-0.332 (0.399)	-3.358 (1.349)*	-2.592 (1.401)*
Group Membership mean effect	-0.584 (0.419)	-0.075 (0.325)	-1.855 (1.134)	-0.446 (1.037)
Disputes mean effect	0.583 (0.512)	0.536 (0.478)	0.767 (1.693)	1.234 (1.490)
Trust mean effect	0.553 (0.254)**	0.567 (0.267)**	3.108 (1.021)**	3.414 (1.076)**
Regression controls	No	Yes	No	Yes
Number of Chiefdoms	444	444	444	444

Notes: Each entry is the coefficient estimate on the language family diversity index from a separate regression. Standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. See Appendix 3.D for details on the mean effects analysis. The instrumental variables are analogously defined conflict indices, using the 1963 data. The regression controls are like those in Table 3.5, columns 2 (OLS) and 5 (IV). All regressions are estimated with survey weights, where each chiefdom (or enumeration area) observation is weighted by the inverse of its sampling probability.

The components of the "Collective Action" category are participation in road brushing, participation in community labor, and participation in community meetings. The components of the "Group Membership" category are member of any community group, member of a credit group, and member of a school group. The components of the "Disputes" category are the incidence of any local assault dispute, any local land disputes, or any local dispute involving theft. The components of the "Trust" category include trust of people in own community (index), trust of people outside community (index), trust of local councilors (index), and trust of the central government (index). Descriptive statistics for each of these outcomes are presented in Appendix Table 3.A.4.

Table 3.A.11: Religious diversity and local outcomes, mean effects analysis

Panel A: Chiefdom-level analysis

	OLS Regressions	
	(1)	**(2)**
Collective Action mean effect	0.046	0.091
	(0.768)	(0.661)
Group Membership mean effect	0.334	0.324
	(0.497)	(0.320)
Disputes mean effect	-0.312	-0.296
	(0.383)	(0.367)
Trust mean effect	-0.190	-0.152
	(0.276)	(0.270)
Regression controls	No	Yes
Number of Chiefdoms	146	146

Panel B: Enumeration-area analysis

	OLS Regressions	
	(1)	**(2)**
Collective Action mean effect	-0.004	-0.098
	(0.466)	(0.449)
Group Membership mean effect	-0.012	-0.158
	(0.335)	(0.276)
Disputes mean effect	-0.159	-0.345
	(0.345)	(0.326)
Trust mean effect	-0.436	-0.324
	(0.298)	(0.275)
Regression controls	No	Yes
Number of Chiefdoms	444	444

Notes: Each entry is the coefficient estimate on religious fractionalization from a separate regression. Standard errors in parentheses. */**/*** denotes significantly different from zero at 90/95/99% confidence. See Appendix 3.D for details on the mean effects analysis. The regression controls are like those in Table 3.5, column 2 (OLS). All regressions are estimated with survey weights, where each chiefdom (or enumeration area) observation is weighted by the inverse of its sampling probability.

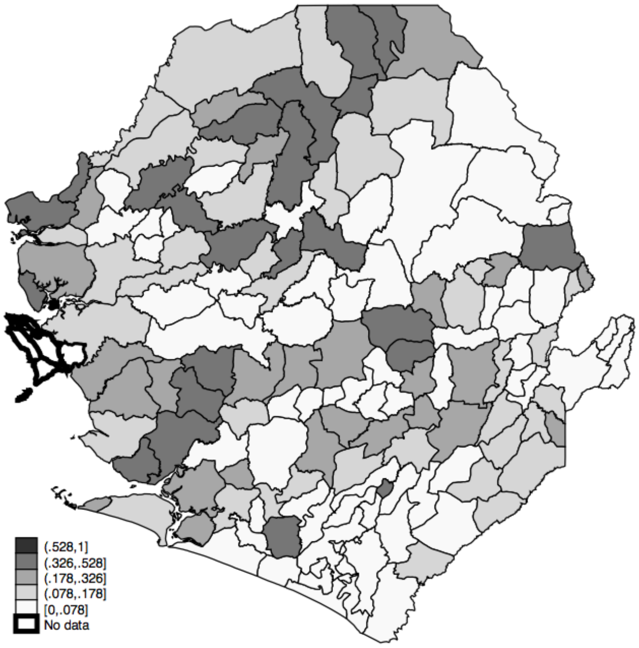
The components of the "Collective Action" category are participation in road brushing, participation in community labor, and participation in community meetings. The components of the "Group Membership" category are member of any community group, member of a credit group, and member of a school group. The components of the "Disputes" category are the incidence of any local assault dispute, any local land disputes, or any local dispute involving theft. The components of the "Trust" category include trust of people in own community (index), trust of people outside community (index), trust of local councilors (index), and trust of the central government (index). Descriptive statistics for each of these outcomes are presented in Appendix Table 3.A.4.

Table 3.A.12: Descriptive statistics for chief strength measures

	Mean (SD)
Age	60.60 (9.44)
Any education indicator	0.79 (0.41)
Number of years in office	10.59 (9.49)
Membership on the National Council of Paramount Chiefs	0.66 (0.48)
Number of years since last Paramount Chief election	18.91 (14.49)
Interim Paramount Chief	0.33 (0.47)

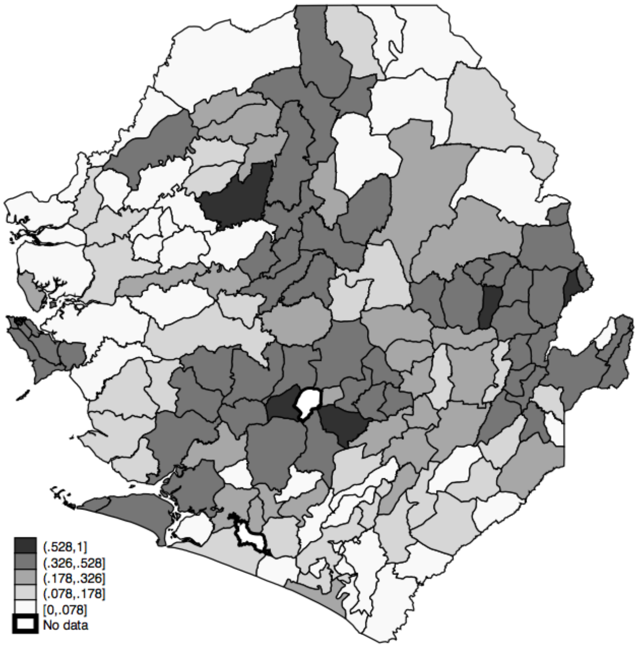
Source: 2008 Chief and Local Councilors Survey. Standard deviations in parentheses. $N = 146$ chiefdoms.

Figure 3.A.1: Historical Ethnic Conflict Index by Chiefdom, 2004



Notes: The mean of $HCON$ across chiefdoms is 0.123, with a standard deviation of 0.123.

Figure 3.A.2: Religious Diversity by Chiefdom, 2004



Notes: The mean of religious diversity across chiefdoms is 0.229, with a standard deviation of 0.179.

3.B Discrete Choice Models with Choice-Based Sampling and Survey Weights

Manski and Lerman (1977) discuss the estimation of discrete choice models with choice-based sampling and endogenous stratification. One approach is to use a weighted maximum likelihood estimator, with weights corresponding to the ratio of population strata probabilities to sample strata probabilities. Another approach is just to maximize the same likelihood function while including a full set of alternative specific constants, but because we included choice variables that are constant across choices, we could not implement this approach. Table 1 reports results using the weighted maximum likelihood procedure.

The NPS surveys, while nationally representative, were designed to oversample smaller chiefdoms, so all regressions reported do use weights to make the sample more reflective of the national population. The sampling probability of each enumeration area (EA) was taken directly from the survey design; EAs were randomly sampled within each district local council area. Hence, the probability that an EA was sampled is just the number of EAs selected per local council area divided by the total number of EAs in that local council area. To obtain the chiefdom-level sample weights, we can compute the probability that chiefdom j was not sampled as follows:

$$1 - H_j = \prod_{e=1}^{N_j} \left(1 - Pr\left\{EA_e \text{ sampled}\right\}\right)$$

where we use the fact that the probability that all EAs were not selected within a chiefdom can be written as the product of the probabilities that each EA was not selected, by independence. Solving the above equation for H_j yields the chiefdom sampling probability. The NPS 2007 has observations on 6,345 individuals, but 408 observations were dropped due to missing information on 1990 residence and 449 because of other missing covariates, leaving an estimation sample of 5,488 individuals. There are 149 chiefdoms and the dataset contains the full set of pairwise combinations of chiefdoms and individuals (817,712 observations). In equation 1, the error term ε_{ij} is distributed i.i.d. extreme value (type 1).

3.C Mapping 1963 Chiefdoms to 2004 Chiefdoms

The mapping between chiefdoms in 1963 and 2004 was generally quite straightforward, as almost all chiefdoms had the same geographic boundaries and did not change their names over the period. Therefore, the construction of 1963 ethnicity shares for chiefdoms as they were defined in 2004 was not problematic. However, there were a few instances in which this was not the case:

- Chiefdoms that unified between 1963 and 2004
  - The 1963 census documents separate what is now Jawie Chiefdom, Kailahun district (Chiefdom ID 1102) into Jawi Lower Chiefdom and Jawi Upper Chiefdom.
- Chiefdoms that split apart between 1963 and 2004
  - Panga Kabonde Chiefdom, Pujehun District,was split into Panga Kabonde Chiefdom (3404) and Sowa Chiefdom (3411).
  - Marampa Masimera Chiefdom, Port Loko District, was split into Marampa Chiefdom (2408) and Masimera Chiefdom (2409).
  - T.M.S. Dibia Chiefdom, Port Loko district, was split into T.M.S. Chiefdom (2411) and Dibia Chiefdom (2403).

For chiefdoms that unified between 1963 and 2004, ethnicity shares were calculated using totals from both areas. For example, we calculated the 1963 ethnicity shares of Jawie chiefdom as the ethnicity shares using totals from Jawi Lower and Jawi Upper Chiefdoms. For chiefdoms that split apart between 1963 and 2004, the 1963 ethnicity shares of the offspring chiefdoms were calculated to be equal to the shares of the "parent" chiefdom in 1963.

3.D Mean Effects Analysis

Katz et al. (2007) discuss two distinct approaches for testing hypotheses about the effect of one covariate on a group of outcomes. All results presented in this paper follow the approach outlined below in which mean effects are constructed from a single index regression. Another way to compute a mean effect size is to jointly estimate regressions of the form in equation 3 for all dependent variables in a grouping using a stacked OLS system (or a SUR system). This allows for separate covariate adjustment for each dependent variable, unlike the procedure outlined in the text. The results are unchanged with the alternative procedure (not shown).

The approach taken in the text is to first form groupings of related outcome variables, denoted by Y_k, $k = 1, ..., K$ (e.g. measures of local collective action). We then standardize each of the outcome variables by subtracting the mean and dividing the standard deviation of the outcome variable for below median ELF areas, our quasi control group. Call each of these standardized outcome variables Y_k^*. With these, we form a single index,

$$Y^* = \sum_k Y_k^*$$

and we regress this on ELF and controls, as in equation 3. The coefficient on ELF in this regression is the mean effect size. This regression can be computed using OLS, with robust standard errors which are clustered when appropriate, as well as using IV methods. This approach is intuitive and is easy to implement computationally.

3.E Data Appendix

3.E.1 National Public Services Surveys, 2005 and 2007

The 2005 and 2007 surveys conducted by the Institutional Reform and Capacity Building Project (IRCBP) provide individual level measures of conflict victimization and measures of local institutional outcomes. The surveys were designed to be nationally representative and representative at the district level, although not necessarily at the lower levels of disaggregation that we analyze. Data is missing for Gbonkolenken chiefdom, which leaves 151 chiefdoms and a total of 539 enumeration areas in all. The sample size is 5,278 households in 2005 and 5,193 households in 2007.

3.E.2 Sierra Leone School Monitoring Survey, 2005

This survey was conducted by IRCBP as part of their ongoing evaluation of local public service provision in Sierra Leone. The school monitoring survey featured two unannounced visits, in which the enumerators collected information on the number of teachers present, the number of children in school, whether the school was open, etc. In addition to this surprise component, enumerators also asked detailed questions regarding schools finances and operations. A total of 288 schools were surveyed, and we use chiefdom averages. There are 104 (out of a total 149) chiefdoms that have school data.

3.E.3 Sierra Leone Household Census, 1963 and 2004

Sierra Leone Household Censuses in 1963 and 2004 were designed to count all individuals in Sierra Leone. For the 1963 data, we digitized data on total population by ethnicity, literacy rates and proportion of population with formal sector employment. We are grateful to Statistics Sierra Leone for their cooperation in sharing the data.

3.E.4 Sierra Leone Chief and Local Councilors Survey, 2008

This survey was conducted by IRCBP, and surveyed all Paramount Chiefs and all elected Local Councilors in Sierra Leone, gathering information on their demographic and socioeconomic characteristics, as well as a range of their political and social views.

3.E.5 No Peace Without Justice (NPWJ) Report, 2004

A measure of conflict intensity that focuses on troops and soldiers is provided by the number of attacks and battles in each chiefdom. This measure was coded from the No Peace Without Justice (NPWJ) conflict mapping report. NPWJ is a non-profit organization that works to promote an effective international criminal justice system and to support accountability mechanisms for war crimes. The conflict mapping report seeks to record all violations of humanitarian law that occurred over the entire conflict period. The factual analysis section of the report is organized chronologically by district, and it reports the chiefdom where each incident occurred, allowing for the construction of chiefdom level war violence measures. The report is available online at: `http://www.npwj.org`.

The measure used in our analysis is the number of attacks and battles that occurred within each chiefdom. An attack is defined to be an incident in which an armed group came into a village briefly, burned houses, raped or killed residents. It is common for attacks to be part of a larger military campaign and thus for human rights violations to be committed on a large scale (e.g. during these attacks RUF forces burnt down fifty houses, killed nine people, abducted an unknown number of people and amputated a mans hand with an axe p. 189). A battle is defined to be a confrontation between two armed groups (e.g. On 25 February, the RUF made a successful counter-attack at the rutile mining site, dislodging the SLA forces based there. p. 430). Battles need not directly involve violence against civilians, although they sometimes do. There were 1,995 violent incidents recoded in the NPWJ report, and 1,363 of these incidents were classified as either an attack or a battle. To give the reader some sense of who the perpetrators of violence against civilians are, of the 968 recorded attacks over 95% were committed by RUF rebels and less than two percent by CDF. The majority of the battles took place between RUF and CDF troops.

3.E.6 Geographic Information Systems (GIS) Data

GIS data provides measures of resources and infrastructure in Sierra Leone. This data is managed and produced by Sierra Leone Information Systems and the Development Assistance Coordination Office (SLIS/DACO) in Freetown. GIS coordinates of all government registered industrial mining sites were combined with firm descriptions from site licenses to determine to location of all registered diamond mining sites. Non-diamond industrial mining plots, including rutile, bauxite, silver, gold, and assorted minerals, are also observed and included as controls in our regression analysis. Because of unregistered and illegal mining, these measures of mining activity may understate the true extent of diamond mining in Sierra Leone. However, since the civil war ended, the government of Sierra Leone has made a concerted effort to document and register all of the mining in the country, as these resources are a major source of government revenue. GIS data was also used to construct measures of road density, river density, distance of the chiefdom to Freetown, and the land area of each chiefdom.

Bibliography

Alesina, A., R. Baqir, and W. Easterly (1999): "Public Goods and Ethnic Divisions," *Quarterly Journal of Economics*, 114, 1243–1284.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic Growth*, 8, 155–194.

Alonso, W. (1964): *Location and Land Use: Toward a General Theory of Land Rent*, Cambridge: Harvard University Press.

Altonji, J., T. Elder, and C. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Anderson, J. E. and E. Van Wincoop (2004): "Trade Costs," *Journal of Economic Literature*, 42, 691–751.

Arellano, M. (1987): "Computing Robust Standard Errors for Within-Group Estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

Asia Foundation (2008): "The Cost of Moving Goods: Road Transportation, Regulations and Charges in Indonesia," Survey report.

Azis, I. J. (1990): "Analytic Hierarchy Process in the Benefit-Cost Framework: A Post-Evaluation of the Trans-Sumatra Highway Project," *European Journal of Operations Research*, 48, 38–48.

Baldwin, K. and J. D. Huber (2010): "Economic versus cultural differences: Forms of ethnic diversity and public goods provision," Working Paper.

Baum-Snow, N. (2007): "Did Highways Cause Suburbanization?" *Quarterly Journal of Economics*, 122, 775–805.

Bellows, J. and E. Miguel (2009): "War and Local Collective Action in Sierra Leone," *Journal of Public Economics*, 93, 1144–1157.

Bennett, C. R., A. Chamorro, C. Chen, H. de Solminihac, and G. W. Flintsch (2007): "Data Collection Technologies for Road Management," Technical report, East Asia Pacific Transport Unit, World Bank.

Berry, S. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242–262.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Blalock, G. and P. J. Gertler (2008): "Welfare Gains from Foreign Direct Investment Through Technology Transfer to Local Suppliers," *Journal of International Economics*, 74, 402–421.

Blomquist, G. C., M. C. Berger, and J. P. Hoehn (1988): "New Estimates of Quality of Life in Urban Areas," *American Economic Review*, 78.

Busso, M., J. Gregory, and P. M. Kline (2010): "Assessing the Incidence and Efficiency of a Prominent Place Based Policy," NBER Working Paper 16096.

Carlton, D. (1983): "The Location and Employment Choices of New Firms: An Econometric Model with Discrete and Continuous Endogenous Variables," *Review of Economics and Statistics*, 65, 440–449.

Casey, K. W. (2009): "Determinants of Voting Choice and Political Party Investment in Sierra Leone," Working Paper.

Chamberlain, G. (1984): "Chapter 22 Panel data," in *Handbook of Econometrics, Volume 2*, ed. by Z. Griliches and M. D. Intriligator, Elsevier, 1247 – 1318.

——— (1992): "Comment: Sequential Moment Restrictions in Panel Data," *Journal of Business & Economic Statistics*, 10, 20–26.

Chay, K. Y. and M. Greenstone (2005): "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 113, 376–424.

Chipman, J. S. (1970): "External Economies of Scale and Competitive Equilibrium," *Quarterly Journal of Economics*, 84, 347–385.

Collier, G. (1970): *Sierra Leone: Experiment in Democracy in an African Nation*, New York: New York University Press.

Coughlin, C. C., J. V. Terza, and V. Arromdee (1991): "State Characteristics and the Location of Foreign Direct Investment within the United States," *Review of Economics and Statistics*, 73, 675–683.

Davidson, J. S. (2010a): "Driving Growth: Regulatory Reform and Expressways in Indonesia," *Regulation and Governance*, 4, 465–484.

——— (2010b): "How to Harness the Positive Potential of KKN: Explaining the Variation in the Private Sector Provision of Public Goods in Indonesia," *Journal of Development Studies*, 46, 1729–1748.

Davies, V. A. B. (2002): "War, Poverty and Growth in Africa: Lessons from Sierra Leone," in *Centre for the Study of African Economies (CSAE) 5th Annual Conference Proceedings.*

Davis, D. and D. Weinstein (2003): "Market Access, Economic Geography and Comparative Advantage: an Empirical Test," *Journal of International Economics*, 59, 1–23.

Deichmann, U., K. Kaiser, S. V. Lall, and Z. Shalizi (2005): "Agglomeration, Transport, and Regional Development in Indonesia," World Bank Policy Research Working Paper 3477.

Dijkstra, E. W. (1959): "A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, 1, 269–271.

Donaldson, D. (2010): "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure," Working Paper.

Dunning, T. and L. Harrison (2010): "Cross-cutting Cleavages and Ethnic Voting: An Experimental Study of Cousinage in Mali," *American Journal of Political Science*, 104, 21–39.

Easterly, W. and R. Levine (1997): "Africas Growth Tragedy: Policies and Ethnic Divisions," *Quarterly Journal of Economics*, 112, 1203–1250.

Eifert, B., E. Miguel, and D. N. Posner (2010): "Political Competition and Ethnic Identification in Africa," *American Journal of Political Science*, 54, 494–510.

Ellison, G. and E. L. Glaeser (1997): "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy*, 105, 889–927.

——— (1999): "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?" *American Economic Review*, 89, 311–316.

Fehr, E. and S. Gächter (2000): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90, 980–994.

Grace, J. (1975): *Domestic Slavery in West Africa, with Particular Reference to the Sierra Leone Protectorate, 1896-1927*, London: Frederick Muller Limited.

Gramlich, E. M. (1994): "Infrastructure Investment: A Review Essay," *Journal of Economic Literature*, 32, 1176–1196.

Gyourko, J., M. Kahn, and J. Tracy (1999): "Chapter 37 Quality of Life and Environmental Comparisons," in *Handbook of Regional and Urban Economics, Volume III*, ed. by E. S. Mills and P. Cheshire, Elsevier, 1413–1454.

Habyarimana, J., M. Humphreys, D. Posner, and J. Weinstein (2007): "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review*, 101, 709–725.

——— (2009): *Coethnicity: Diversity and the Dilemmas of Collective Action*, New York: Russell Sage.

Harris, C. D. (1954): "The Market as a Factor in the Localization of Industry in the United States," *Annals of the Association of American Geographers*, 44, 315–348.

Head, K. and T. Mayer (2004): "Market Potential and the Location of Japanese Investment in the European Union," *The Review of Economics and Statistics*, 86, 959–972.

Head, K., J. Ries, and D. Swenson (1995): "Agglomeration Benefits and Location Choice: Evidence from Japanese Manufacturing Investments in the United States," *Journal of International Economics*, 38, 223–247.

Helpman, E. (1998): "The Size of Regions," in *Topics in Public Economics: Theoretical and Applied Analysis*, ed. by D. Pines, E. Sadka, and I. Zilcha, Cambridge University Press: Cambridge, 33–54.

Henderson, J. V. and A. Kuncoro (1996): "Industrial Centralization in Indonesia," *The World Bank Economic Review*, 10, 513–540.

Hill, H. (2000): *The Indonesian Economy*, Cambridge: Cambridge University Press.

Hord, H. H. V. (1966): "The Conversion of Standard Fruit Company Banana Plantations in Honduras from the Gros Michel to the Giant Cavendish Variety," *Tropical Agriculture*, 43, 269–275.

Human Rights Watch (1999): "Sierra Leone: Getting Away with Murder, Mutilation, and Rape," Survey report.

Jackson, B., J. Wilson, S. Kirbah, S. Sidney, J. Rosenberger, L. Bassie, J. Alie, D. McLean, W. Garvey, and B. Ely (2005): "Mitochondrial DNA genetic diversity among four ethnic groups in Sierra Leone," *American Journal of Physical Anthropology*, 128, 156–163.

Jackson, M. (1974): "The Structure and Significance of the Kuranko Clanship," *Africa: Journal of the International African Institute*, 44, 397–415.

Jacoby, H. G. (2000): "Access to Markets and the Benefits of Rural Roads," *The Economic Journal*, 110, 713–737.

Kandeh, J. D. (1992): "Politicization of Ethnic Identities in Sierra Leone," *African Studies Review*, 35, 81–99.

KATZ, L., J. KLING, AND J. LIEBMAN (2007): "Experimental analysis of neighborhood effects," *Econometrica*, 75, 83–119.

KEEN, D. (2005): *Conflict and Collusion in Sierra Leone*, New York: Palgrave.

KLINE, P. (2010): "Place Based Policies, Heterogeneity, and Agglomeration," *American Economic Review: Papers and Proceedings*, 100, 383–387.

KOEPPEL, D. (2008): *Banana: The Fate of the Fruit That Changed the World*, New York: Hudson Street Press.

KRUGMAN, P. (1991): "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99, 483–499.

KUP, A. P. (1961): *A History of Sierra Leone, 1400-1787*, Cambridge: Cambridge University Press.

LANGER, A. (2010): "Demographic Preferences and Price Discrimination in New Vehicle Sales," Job Market Paper, Unpublished.

LEINBACH, T. R. (1989): "Transport Policies in Conflict: Deregulation, Subsidies, and Regional Development in Indonesia," *Transportation Research Part A: General*, 23, 467–475.

LIU, X., M. E. LOVELY, AND J. ONDRICH (2010): "The Location Decisions of Foreign Investors in China: Untangling the Effect of Wages Using a Control Function Approach," *Review of Economics and Statistics*, 92, 160–166.

MAMDANI, M. (1996): *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*, Princeton: Princeton University Press.

MANSKI, C. AND S. LERMAN (1977): "The estimation of choice probabilities from choice based samples," *Econometrica*, 45, 1977–1988.

MARQUARDT, S. (2001): " 'Green Havoc': Panama Disease, Environmental Change, and Labor Process in the Central American Banana Industry," *American Historical Review*, 106, 49–80.

MICHAELS, G. (2008): "The Effect of Trade on the Demand for Skill: Evidence from the Interstate Highway System," *The Review of Economics and Statistics*, 90, 683–701.

MIGUEL, E. (2004): "Tribe or Nation?: Nation Building and Public Goods in Kenya versus Tanzania," *World Politics*, 56, 327–362.

MIGUEL, E. AND M. K. GUGERTY (2005): "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya," *Journal of Public Economics*, 89, 2325–2368.

MILLS, E. (1967): "An Aggregative Model of Resource Allocation in a Metropolitan Area," *American Economic Review*, 57, 197–210.

Montalvo, J. G. and M. Reynal-Querol (2005): "Ethnic Polarization, Potential Conflict and Civil War," *American Economic Review*, 95, 796–816.

Moretti, E. (2010): "Local Labor Markets," in *Forthcoming, Handbook of Labor Economics*, Elsevier.

Muth, R. (1969): *Cities and Housing: the Spatial Pattern of Urban Residential Land Use*, Chicago: University of Chicago Press.

Nevo, A. (2000): "A Practitioner's Guide to Estimation of Random-Coefficient Logit Models of Demand," *Journal of Economics and Management Strategy*, 9, 513–548.

Oliver, R. A. and A. Atmore (2001): *Medieval Africa, 1250-1800*, Cambridge: Cambridge University Press.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Ostrom, E. (1990): *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge: Cambridge University Press.

Oyètádé, B. and V. Luke (2008): "Sierra Leone: Krio and the Quest for National Integration," in *Language and National Identity in Africa*, ed. by A. Simpson, Oxford: Oxford University Presss, 122–140.

Posner, D. (2004): "The Political Salience of Cultural Difference: Why Chewas and Tumbukas are Allies in Zambia and Adversaries in Malawi," *American Political Science Review*, 98, 529–545.

——— (2005): *Institutions and Ethnic Politics in Africa*, Cambridge: Cambridge University Press.

Redding, S. J. and D. M. Sturm (2008): "The Costs of Remoteness: Evidence from German Division and Reunification," *American Economic Review*, 98, 1766–1797.

Reno, W. (1995): *Corruption and State Politics in Sierra Leone*, Cambridge: Cambridge University Press.

Richards, P. (1996): *Fighting for the Rainforest: War, Youth and Resources in Sierra Leone*, Portsmouth, NH: Heinemann for the International African Institute.

——— (2005): "To Fight or to Farm? Agrarian Dimensions of the Mano River Conflicts (Liberia and Sierra Leone)," *African Affairs*, 104, 571–590.

Roback, J. (1982): "Wages, Rents, and the Quality of Life," *Journal of Political Economy*, 90, 1257–1278.

Rosen, S. (1979): "Wage-Based Indexes of Urban Quality of Life," in *Current Issues in Urban Economics*, ed. by P. Mieszkowski and M. Straszheim, Johns Hopkins University Press.

Samuelson, P. A. (1954): "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments," *The Economic Journal*, 64, 264–289.

Sappington, J. M., K. Longshore, and D. Thompson (2007): "Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study using Bighorn Sheep in the Mojave Desert," *Journal of Wildlife Management*, 71, 1419–1426.

Sayers, M. W., T. D. Gillespie, and W. D. Paterson (1986): "Guidelines for Conducting and Calibrating Road Roughness Measurements," World Bank Technical Paper 46.

Schama, S. (1995): *Rough Crossings: Britain, the Slaves and the American Revolution*, New York: HarperCollins.

Sjöberg, Ö. and F. Sjöholm (2004): "Trade Liberalization and the Geography of Production: Agglomeration, Concentration, and Dispersal in Indonesia's Manufacturing Industry," *Economic Geography*, 80, 287–310.

Soluri, J. (2000): "People, Plants, and Pathogens: the Eco-Social Dynamics of Export Banana Production in Honduras," *Hispanic American Historical Review*, 80, 463–501.

——— (2005): *Banana Cultures: Agriculture, Consumption, and Environmental Change in Honduras and the United States*, Austin: University of Texas Press.

Spitzer, L. (1974): *The Creoles of Sierra Leone: Responses to Colonialism, 1870-1945*, Madison: University of Wisconsin Press.

Stover, R. H. and N. W. Simmonds (1987): *Bananas*, New York: John Wiley & Sons.

Tishkoff, S., F. Reed, F. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. Hirbo, A. Awomoyi, J. Bodo, O. Doumbo, et al. (2009): "The Genetic Structure and History of Africans and African Americans," *Science*, 324, 1035–1044.

Vigdor, J. (2002): "Interpreting Ethnic Fragmentation Effects," *Economics Letters*, 75, 271–276.

wa Thiongo, N. (2009): *Something Torn and New: An African Renaissance*, New York: Basic Civitas Books.

Wardlaw, C. (1929): "Virgin Soil Deterioration: the Deterioration of Virgin Soils in the Caribbean Banana Lands," *Tropical Agriculture*, 6, 243–249.

——— (1972): *Banana Diseases, including Plantains and Abaca*, New York: Humanities Press.

Wie, T. K. (2000): "The Impact of the Economic Crisis on Indonesia's Manufacturing Sector," *The Developing Economies*, 38, 420–453.

World Bank (2007): "A Decade of Action in Transport: An Evaluation of World Bank Assistance to the Transport Sector, 1995-2005," World bank independent evaluation group report.

WYSE, A. (1989): *The Krio of Sierra Leone: An Interpretive History*, London: C. Hurst & Co.

YATCHEW, A. (1997): "An Elementary Estimator of the Partial Linear Model," *Economics Letters*, 57, 135–143.

YU, J., E. CHOU, AND J. YAU (2006): "Development of Speed-Related Ride Quality Thresholds using International Roughness Index," *Transportation Research Record*, 1974, 47–53.