College of Business and Public Policy
UNIVERSITY of ALASKA ANCHORAGE

Department of Economics Working Paper

WP 2019-03
August 2019

Geography, Geology, and Regional Economic Development

KEVIN BERRY
University of Alaska Anchorage

ALEXANDER JAMES
University of Alaska Anchorage

BROCK SMITH
Montana State University

BRETT WATSON
University of Alaska Anchorage

UAA DEPARTMENT OF ECONOMICS
3211 Providence Drive
Rasmuson Hall 302
Anchorage, AK 99508

http://econpapers.uaa.alaska.edu/

Geography, Geology, and Regional Economic Development*

Kevin Berry[†] Alexander James[‡] Brock Smith[§] Brett Watson[†]

August 5, 2019

Abstract

We examine long-run development effects of regional productivity shocks in the United States. We exploit the timing and location of large resource discoveries to measure exogenous variation in labor demand and consider heterogeneous effects based on environmental amenity and geographic isolation, developing novel measures of both. Using a dynamic event-study analysis we find that productivity shocks increase population both in the short and long-run, but this largely reflects the experience of low amenity, geographically isolated places that may otherwise struggle to develop. Moreover, this study offers several insights into the observed spatial pattern of development in the United States.

Keywords: Natural-Resource Discoveries; Regional Development; Long-Run Growth; Geography; Environmental Amenities

JEL Classification: Q32; Q33; R11

*A previous version of this paper was circulated with the title, "The long-run effects of resource discoveries and the role of amenity value". Thanks to Grant Jacobsen and Guy Michaels for helpful comments and reactions. Thanks to members of the Department of Economics and Public Policy at the University of Alaska Anchorage, the Department of Agricultural Economics and Economics at Montana State University, the Department of Agricultural Economics at Texas A&M University, the Department of Economics at the University of Alabama, participants of the 2017 CU Boulder Environmental Economics Workshop, the 2018 World Congress of Environmental and Resource Economists, and the 2019 UC Davis Alumni Conference for helpful comments and questions. All errors are our own.

[†]Department of Economics and the Institute of Social and Economic Research, University of Alaska Anchorage. Anchorage, AK. 99508.

[‡]Department of Economics, University of Alaska Anchorage. Anchorage, AK. 99508.

[§]Department of Agricultural Economics & Economics, Montana State University. Bozeman, MT., 59717. *Corresponding Author.* brock.smith1@montana.edu

"New roads, agriculture, employment, education, these are just a few of the things we can offer you, and I assure you ladies and gentlemen, that if we do find oil here, and I think there's a very good chance that we will, this community of yours will not only survive, it will flourish"—Daniel Plainview, There Will Be Blood.

1 Introduction

What determines where people choose to live? Within a country, where migration frictions are less pronounced, environmental amenities like warm temperature (Roback, 1982), mountain viewsheds (Groothuis, Groothuis, and Whitehead, 2007), clean air (Chay and Greenstone, 2005), and rainfall (Englin, 1996) clearly matter. But real wages are also important; positive labor demand shocks attract people from near and far away. This feature of human behavior is well documented in economics literature and is especially salient in resource-based development literature which affirms a robust relationship between natural-resource shocks, labor demand, and inward migration (Jacobsen and Parker, 2016; Alcott and Keniston, 2017; Feyrer, Mansur and Sacerdote, 2017; Richter, Salanguit and James, 2018; James and Smith, 2019). But this literature has overlooked the joint role of economic shocks and environmental amenity. Just as a high school degree is more important for people with low cognitive ability (Murnane, Willett, and Tyler, 2000), a boost to productivity might be especially important for otherwise less desirable locations.

Motivated by a simple analytical framework, we examine how large and localized productivity shocks—in the form of major oil and mineral discoveries—affect long-run population growth and regional economic development within the United States, and examine how these shocks interact with pre-existing geographic properties of the discovery site. Estimating long-run, truly posterior effects of economic shocks requires an analysis of events that happened long ago. However, doing so is often frustrated by the lack of available data; causal inference typically requires observing initial conditions. While census records provide historical subnational population estimates, county borders change over time and additional counties are added as new territories are acquired. To account for these issues, we make use of novel geospatial population count estimates measured at a 1km^2 resolution for decades between 1790 and 2010 constructed by Fang and Jawitz (2018). These data, along with our identification strategy, allow us to measure both short, medium and long-run effects (e.g., 100 years) of productivity shocks.[1]

We uniquely explore important heterogeneities in the effect of resource shocks along two dimensions: environmental amenity value and geographic isolation, constructing novel measures of both. Rather than relying on existing estimates of county-level environmental amenity which may reflect endogenous factors such as school quality (Bieri, Kuminoff and Pope, 2014), we use the idea that people "vote with their feet" (Tiebout, 1956) and construct a data-driven estimate of amenity value that is a function of purely exogenous, naturally-occurring characteristics like warm weather, mountain viewsheds, terrain ruggedness, and soil quality. We measure geographic isolation as the cost of traveling from a county to the nearest major market, defined as either an existing city, railroad, or marine highway. Rather than relying on a "as the crow flies" measure of distance, we develop a novel measure of geographic isolation that incorporates information about terrain ruggedness and water coverage between locations.

We separately consider discoveries of major oil fields and mineral deposits. The former effects tends to be more persistent (i.e., counties that discovered oil 100 years ago tend to be producing oil today, albeit, not necessarily from the same oil field) making it difficult to estimate truly posterior effects. However, mineral discoveries, while still creating significant labor demand, tend to be much smaller in value and are commonly exhausted within fifty years of discovery. This feature of the data allows us to test for path dependence, the idea that places develop because of historical influences that lead to agglomeration effects and persistent development.[2]

[1]It is important to note that our population analysis only applies to non-indigenous peoples. Our population data from Fang and Jawitz (2018) reconstructs spatial population distributions that do not consider Native Americans, because the US Census did not begin counting Native Americans until 1900. Therefore we stress that when we refer to uninhabited frontier locations, we mean uninhabited by non-indigenous peoples. Exploring how the population dynamics studied in this paper interacted with the pre-existing presence of Native Americans is beyond the scope of this paper but would be an interesting avenue for future study.

[2]Path dependence specifically refers to the idea that regional development reflects historical shocks that lead to agglomeration effects and persistent long run development. For example, Michaels and Rauch (2017) find that medieval towns were more likely to be found in the same location as Roman Empire-era towns in France than in Britain. They argue this is because the collapse of the empire temporarily ended all urbanization in

Our analysis is made up of two main parts: an event-study analysis of the effect of discovering both oil and minerals on population and a cross-sectional analysis of a broader set of modern-day economic outcomes including income, education attainment, poverty, unemployment, and the existence of transportation infrastructure. The event-study analysis indicates that, fifty years after discovering oil, population density in treatment counties is on average roughly 65% greater than it otherwise would have been. However, this average effect conceals important heterogeneities. When limiting our sample according to environmental and geographic favorability, we find that for relatively unfavorable counties (those with low amenity value or high transportation cost) the long-run treatment effect on population increases to roughly 170%. While the results for moderate favorability places are similar to the overall average treatment effect, effects for high favorability places are qualitatively small and statistically insignificant.

Examining the effect of mineral discoveries reveals similar heterogeneities, although the effects tend to be short lived with one exception: geographically isolated places with high transportation cost are more likely to have people living in them fifty years after discovering a mineral deposit nearby. One somewhat speculative interpretation of this finding is that resource extraction contributes to the fixed cost of infrastructure construction necessary for a place to become livable. We provide some supporting evidence for this by showing that otherwise geographically isolated places near mineral discoveries are, today, more likely to have access to transportation infrastructure like roads and railroads.

We carry out several robustness checks designed to test for endogeneity bias and reverse causality and explore the sensitivity of the results to various modeling assumptions. Our results prove to be quite robust and do not appear to be driven by endogenous natural-resource discoveries (whereby economic development leads to discovery). In particular, while discovery is potentially endogenous, the size of discovery is left to chance and we find that discoveries of large oil fields generate significantly larger effects than discoveries of small fields. We also reject the idea that our heterogeneous results are mechanical, whereby low-favorability places are less populated to begin with, and so experience larger percent increases in population

Britain, but not in France, suggesting a path-dependence effect.

following the discovery of a resource.

Taken together, we conclude that naturally-occurring amenity value, geographic features, and labor market opportunities jointly determine where people choose to live. We also find that the importance of one factor depends on the absence or presence of the other; a major resource shock in a place with high amenity value (e.g., southern California) may have minimal long-run effects. But a discovery in a less desirable, more geographically isolated place, such as North Dakota, may be critical for long-run economic development. In doing so, we make a number of important contributions to development, regional, and environmental economics literature.

Most directly, we contribute to a large literature that examines the short and long-run economic and political effects of resource extraction, booms, and discoveries (Michaels, 2011; Weber, 2012, 2014; Jacobsen and Parker, 2016; Muehlenbachs *et* al. 2015; Berman, *et.* al. 2017; Alcott and Keniston, 2017; Feyrer, Mansur and Sacerdote, 2017; Matheis, 2016). To our knowledge, we offer the first event-study analysis of historical oil and mineral discoveries in the United States, allowing for more robust causal inference. Additionally, the existing literature almost exclusively estimates average treatment effects. We demonstrate that important heterogeneities exist, raising questions about the external validity of any regional analysis. This may help to explain the wide variety of the estimated effects of resource booms, including the recent shale-energy boom literature that is synthesized by Marchand and Weber (2018). We also add to the literature on economic geography by providing novel estimates of local, naturally-occurring amenity value and geographic isolation that complements the work of Blomquist, Berger, and Hoehn (1988), Allen and Arkolakis (2014), Henderson, *et.* al. (2017), among others.

We also contribute to the extant literature on the relative importance of geographic properties vs. path dependence. As discussed by Bleakley and Lin (2012), "...it is difficult to disentangle the effects of state dependence (the presence of factors of production) versus serial correlation (the advantages that first attracted other factors)". Such considerations have motivated long-run analyses of short-term shocks including portaging (Bleakley and Lin, 2012), bombing campaigns (Davis and Weinstein, 2002), and the temporary location of railroads (Jedwab, Kerby, and Moradi, 2017). Historical mineral discoveries provide another opportunity to test for path dependence because mines are relatively short lived. Our study design allows us to jointly evaluate the role of both geography and path dependence, and we find some evidence for both. The importance of geographic fundamentals is demonstrated by the absence of resource shock effects in high-amenity or low-transportation cost locations (for both oil and mineral booms), along with the fact that these locations nonetheless have large average populations, implying that the development of these locations was inevitable to some extent. This finding complements much of the extant literature linking natural amenities and geographic features to economic development (Sachs, 2001; Miguel and Roland, 2011; Felkner and Townsend, 2011; McGregor and Wills, 2017). But our mineral results also suggest, albeit with somewhat mixed evidence, that some degree of path dependence exists for lower-favorability locations.

This paper is organized as follows: Section 2 lays out a theory of how natural geographic properties interact with resource shocks to bring about long-run population growth. Section 3 discusses the two main identification strategies and Section 4 describes the estimation of both amenity value and transportation cost. The various data sources are discussed in Section 5 and the results are given in Section 6. Section 7 presents a series of robustness checks and Section 8 concludes.

2 Theoretical Motivation

To motivate the empirical analysis and develop hypotheses, we construct a model of resource extraction in a frontier space. We assume the resource is open access in that property rights are either weakly enforced or non-existent, potentially leading to a race to extract a natural resource in the frontier. There are two geographic regions consisting of the home and frontier regions. The home region provides potential laborers to the frontier who are price takers in the labor market. Let utility in the home region be fixed at $\bar{u}$. Utility in the frontier is a function of consumption, population, and natural-amenity value. Specifically, let utility of a representative person living on the frontier be given by

$$u_f = u(c, L, a), \tag{1}$$

where c is per capita consumption, L is population of the frontier, and a is amenity value in the frontier. We assume that utility is concave and strictly increasing in both consumption and amenity value. As in Blomquist, Berger and Hoehn (2017), when the population of the frontier is small, we assume that utility is increasing in population. Practically, this may reflect enhanced entertainment and medical services, or opportunities for courtship. Following Arnott and Stiglitz (1979), we assume the existence of a utility-maximizing population, L^*, and that population growth beyond L^* reduces utility on the frontier due to congestion, crime, or pollution, for example.

Each person living on the frontier works in a manufacturing sector. A natural resource is costlessly produced and both goods are exported in exchange for a composite consumption good given by

$$c = (1 - \tau(d))(m + \frac{R}{L}), \tag{2}$$

where c is the per capita consumption of the composite good, τ is an iceberg transport cost (itself an increasing function of distance, d, to a major market), m is per capita production of a manufactured good (taken as given), and R is the total value of a natural resource. Due to the assumed lack of property rights, the value of the resource is evenly distributed among all of the people living in the frontier. Substituting (2) into (1) and differentiating with respect to L implies the relationship between utility and population is ambiguous:

$$\frac{\partial u_f}{\partial L} = \frac{\partial u}{\partial L} - \frac{\partial u}{\partial c}(1 - \tau(d))\frac{R}{L^2} \lesseqgtr 0. \tag{3}$$

For a sparsely populated frontier, an increase in population offers a variety of benefits, but comes at the cost of sharing the resource among a larger number of people. Only for a sufficiently small value of resource production does the former effect outweigh the latter. In this case, utility in a sparsely populated frontier is increasing in population up to the utility maximizing population, L^*.

The equilibrium frontier population is determined by setting utility in the frontier equal to that in the home region which is assumed to be fixed at $\bar{u}$:

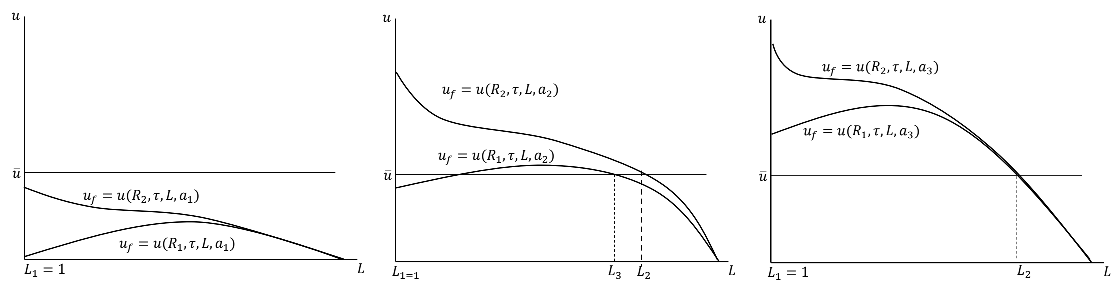
$$u((1 - \tau(d))(m + \frac{R}{L}), L, a) - \bar{u} = 0. \tag{4}$$

For the frontier to become populated, the first person to move to the frontier must find it advantageous to do so. Referring to panel (a) of Figure 1, if amenity value is low enough (or transport cost is high enough) no frontiersman will ever find it advantages to move. A resource discovery that increases the value of resource production from R_1 to R_2 increases the utility earned from living on the frontier, especially when the population is low (because the resource would be divided by a smaller number of people). However, in this case, the natural favorability of the place is so low that even a large resource discovery may not attract labor from the home region.

Next, consider panel (b) which describes a frontier space that is livable. Absence a resource discovery, amenity value is too low (transport cost too high) to attract labor from the home region. However, in this case, a resource discovery sufficiently increases the utility earned from living on the frontier and the first person moves. As more people follow, the resource is divided among a larger and larger group, causing utility earned by each person to fall. As congestion and pollution effects take hold, utility on the frontier is eventually driven down to the level earned in the home region and population settles at L_2, though this transition may take many years. Interestingly, a resource discovery and resulting shock need not be permanent in order to permanently populate a frontier. In the event that R_2 reverts back to R_1, the population of the frontier moves from L_2 to L_3, which is a stable equilibrium.

Now consider panel (c) which depicts utility in a hospitable frontier location with high amenity, or low transport cost, or both. The population of this frontier eventually swells to a population of L_2, even in the absence of a resource discovery. While a resource discovery makes this type of frontier even more appealing, the long-run trajectory of this type of place is largely unaffected, regardless of whether the resource shock is permanent or temporary. Potential examples of these types of frontiers are abundant in California. It is hard to imagine that the counties of Orange, San Luis Obispo, Kern, Los Angeles, Monterey, Santa Barbara, and Ventura would not be so heavily populated in the absence of major oil discoveries occurring in the late 1800s and early 1900s. And of course neighboring counties that did not experience large oil discoveries—such as San Diego, Santa Clara, and San Mateo are similarly populated.

Figure 1: A Resource Discovery in a Frontier, by Location Favorability



(a) Inhospitable (b) Livable (c) Hospitable

Utility on the frontier u_f and the reference home location $\bar{u}$, in inhospitable, livable, and, hospitable frontier locations. A resource discovery shock on the frontier increases R_1 to R_2, shifting u_f upwards. The equilibrium population L is L_1 prior to the resource shock, L_2 during the resource boom period, and L_3 after the resource is depleted.

In the following section, we empirically explore the predictions of the model that are outlined here. Namely, we evaluate how population evolves before, during, and after major resource discoveries in places with varying levels of environmental and geographic amenities. It is important to note, however, that ex-ante we have no way of knowing what constitutes an inhospitable, livable, or hospitable location as discussed in this section. It is possible that sufficiently low levels of amenity value do not exist in the continental United States for the effect of resource discoveries to have negligible effects on population. We nonetheless examine the effects of resource discoveries in counties with high, medium, and low levels of environmental and geographic amenity values, but again, these categories are not necessarily analogous to the three discussed here.

3 Identification Strategy

Our main analysis is an event-study framework that estimates the dynamics of population around the time that a resource discovery is made. To measure long-run development across a larger set of indicators, we also estimate a cross-sectional model that compares outcomes for resource-rich and poor areas in the 2000s, well after resource discoveries have occurred.

Our analysis separately considers two varieties of high-value non-renewable resources: oil fields and metallic mineral deposits.[3] Oil fields and mineral resources provide complementary evidence of the effect of non-renewable resource discovery on long-run economic development. Generally speaking, oil fields in the United States are geographically larger, economically more valuable, and have longer periods of utilization relative to mineral deposits. We therefore anticipate finding more significant and persistent treatment effects for oil. The unique geographic and temporal distribution of oil and mineral discoveries also lends to somewhat distinct empirical strategies. The complementary nature of these strategies show the robustness of our findings while also providing insight into the underlying mechanisms.

Our analysis considers heterogeneity across two dimensions: local amenity value and transportation cost to national and international markets. Local amenities are utility-improving exogenous characteristics relating to weather, proximity to mountains and coasts, and soil quality. Transportation cost is based on the distance from—and topographic features between—a particular location and the nearest market. We create a discrete three-point scale of locations based on their favorability in these two dimensions. The most favorable locations have high amenity values and/or low transportation cost. Unfavorable locations have low amenity values and/or high transportation costs. Moderate locations fall in-between. We assign locations into low, medium or high bins according to 20th/80th percentile cutoffs of our numerical amenity and transportation cost values (the construction of these measures is described in Section 4).

[3]Metallic minerals include the precious metals (e.g. gold, silver, platinum group metals), major metals, (e.g. iron, copper, lead, zinc, tin, aluminum), alloying metals (e.g. nickel, molybdenum, cobalt), rare and specialized metals (e.g. the lanthanides, gallium, lithium, uranium, beryllium). However, we exclude nearly all non-metallic mineral resources (e.g. sand/gravel, clay, stone, phosphate, pot ash). We also exclude coal resources (despite their regional significance) because the discovery of major bituminous coal basins in the United States pre-dates the country's independence.

To evaluate heterogeneous effects of resource discovery, we separately estimate both our event study and cross-sectional models within each amenity and transportation cost bin.

Our measure of local amenity is determined only by a county's climate and geography and is therefore exogenous with respect to the economic outcomes of interest. Our measure of transportation cost is based on both geographic factors and proximity to cities existing prior to the American oil and mineral booms. While the timing of discovery could be endogenous to economic conditions at the time of discovery, this is less of a concern in our long-run cross-sectional analysis than in past studies which have focused on short-run outcomes. Since America has been prospected for several centuries and only a few discoveries have been made since the 1950s, our treatment sample should closely correspond to the true distribution of resources within the United States. Further, our event study framework explores the population dynamics around the time of discovery, and we do not observe strong pre-existing trends in population effects leading up to discovery.[4]

3.1 Discovery Event Study Analysis

The event study examining effects of resource discoveries on local population density is specified below:

$$\ln(\text{Pop})_{i,s,t} = \alpha + \sum_{u \neq -10} \beta_u E_{i,t+u} + \sum_{y=1870}^{2000} \gamma_y X_i + \phi_i + \delta_{s,t} + \epsilon_{i,t} \tag{5}$$

where $\ln(\text{Pop})_{i,s,t}$ is the natural log of population density for county i in state s in census year t,[5] $E_{i,t+u}$ is a set of indicators equal to one if county i experienced a major discovery[6] u years ago, with u = -10 as the omitted category. γ_y is a set of year indicators and X_i is a set of fixed properties of county i. These properties include dummies for whether the

[4] We also address the concern of endogenous timing of discovery in Section 7 by documenting larger population effects for larger oil field discoveries.

[5] To account for shifting county borders over time, for each census year we find the population within present-day county borders using our spatial population data set. See Section 5.3 for further detail.

[6] As detailed in Section 5, a major oil discovery is of a field known to contain at least 100 million barrels. A major mineral deposit is one described as "world class" in the USGS Mineral Resources Data System.

county is in the low, medium, or high amenity bin, whether the county is in the low, medium, or high transport cost bin, and the population density as of 1870 to control for convergence effects. ϕ_i is county fixed effects and $\delta_{s,t}$ is state-year fixed effects. The coefficient of interest, β_u, represents the effect of discovering oil u years ago. We observe population every 10 years (corresponding to national censuses) from 1870 to 2000 (excluding 1960, which is missing from our population data. See Section 5.3), using 1870 as a cutoff because that is before the first major oil discovery and also when the vast majority of counties have non-zero population and thus are not dropped from our natural log specification. We assign the event year (u=0) as the first census year after a county makes its first discovery. Standard errors are clustered at the county level.

We first estimate Equation (5) for the full nationwide sample of counties to find an overall average effect of oil discoveries. We then estimate it separately for each transportation cost and amenity bin to study heterogeneous effects by location favorability. When analyzing a certain amenity bin, we retain our transportation cost-by-year interactions as controls (and vice-versa), which is important since amenity value and transportation cost are correlated. For the full sample, we expect β_u to be positive in the periods after discovery. We also expect the population effects in favorable locations to be smaller than in unfavorable locations.

We conduct an analogous exercise for mineral discoveries with several key differences. Whereas our data on oil fields is given at the county level[7], we know the precise spatial coordinates of mineral deposits (e.g. the mine shaft entrance or center of the ore-body). We leverage this property along with our geospatial data providing decennial population estimates at a 1km^2 cell resolution to analyze cell-level population effects in close proximity to mines. A cell-level analysis offers major advantages over a county-level one. First, mineral deposits are more geographically confined than oil fields, and cell-level analysis allows us to more accurately define treatments based on mine proximity. Second, since mineral deposits are more common in mountainous areas, they tend to appear in less densely populated and geographically larger counties with a mix of rugged and flat terrain (even within the western US, counties with a

[7]As discussed in Section 5, we know the counties in which major oil fields were discovered, but we do not know the precise locations of fields.

mine discovery are roughly twice as large on average as those without). Cell-level analysis allows us to more precisely control for local terrain ruggedness.

Another key difference for our mineral analysis is that the vast majority of major mineral discoveries occurred in the western U.S. at a time when it was extremely sparsely populated by non-indigenous peoples (the first gold rush in northern California occurred in 1848, the same year most of the western U.S. was annexed from Mexico). We therefore limit our mining sample to Western states.[8] We also use the sample period of 1850-1950, which covers the period from when most of the West became U.S. territory to at least a few decades after the vast majority of mineral discoveries.

For our baseline mineral specification, a cell is considered treated if it is within 30 miles of a major deposit and has also been exploited.[9] We choose 30 miles since deposits tend to be in especially rugged mountain areas and the associated mining towns sometimes locate fairly far away in the nearest flat land, but we also analyze effects for different radii in Appendix B.

We define the event year as the first census year after the mine is discovered, though for several mines this date is not available. For these cases, we use the first production date instead. If both the discovery year and first production year are unavailable, we do not include the mine in our analysis. This eliminates an additional 39 world-class deposits in the Western U.S. (after dropping non-producing deposits). Cells near multiple mines are assigned the earliest event year. We do not include cells with an event census year before 1860 or after 1950, so that we have at least one observation before and after treatment for each treated cell, and those are compared only to cells that are untreated. This leaves us with 129 world-class deposits with which to identify effects of mineral discovery.

Because the vast majority of cells are uninhabited by non-indigenous peoples at the beginning of our sample period, we also modify our outcome variable in Equation 5 to be a binary indicator variable that equals 1 if the cell contains a population of at least one and zero otherwise. We therefore consider this an analysis of extensive margin population effects

[8]New Mexico, Colorado, Wyoming, Montana, Arizona, Utah, Nevada, Idaho, California, Oregon, Washington.

[9]The USGS mineral deposit data indicates present-day development status. We only use mines with status "Producer" or "Past Producer" and do not use those with status "Occurrence" or "Prospect". This eliminates 15 of the 194 world-class deposits in the Western US.

in the American West. Figure A.7 shows how this cell-level population dummy evolves over space and time within our sample. The fact that the Western US is so sparsely populated during the mineral discovery period makes this setting particularly relevant for testing the predictions of the frontier model outlined in Section 2.

Finally, we also modify our control variables for the mineral analysis. First, we re-bin the amenity index and transportation cost measures for the Western U.S. cell-level sample, again using 20th/80th percentile cutoffs.[10] Second, we do not control for convergence effects since the outcome is an indicator dummy and the initial population is zero for the vast majority of the sample. Third, we create cell-level ruggedness bins (again using 20th and 80th percentile cutoffs) and additionally control for ruggedness[11] interacted by year; although average cell ruggedness is a component of our county-level amenity value that is applied to cells (see Section 4.1), we can measure local ruggedness much more precisely at the cell level, and this is plausibly a major factor in determining cell-level population growth. Finally, for the mineral analysis standard errors are clustered at the zip code level.[12]

3.2 Long-Run Cross Sectional Analysis

We supplement our event-study analysis with a cross-sectional analysis of outcomes in the 2000s. This approach offers two of advantages. First, it allows us to examine the effect of historical resource discoveries on a wider variety of outcomes including income, poverty, education attainment, and transportation infrastructure. Additionally, this specification further addresses the concern that the timing of discoveries is endogenous to overall regional economic development. For this cross-sectional analysis, treatment locations are those that ever discovered a resource, regardless of when the discovery was made. In this way treatments are arguably defined based off of exogenously determined geological factors only. As in our event study of population dynamics, we exploit the more spatially explicit locations of mineral

[10] Our amenity index is constructed at the county level, so we assign each cell within a county the same index score. See Section 4.

[11] See Section 4.2 for information about our ruggedness measure.

[12] An alternative would be county-level clusters, but for the regressions that restrict the sample to certain amenity or transport cost bins this results in more parameters than clusters.

resources by using high-resolution spatial data to analyze a complementary set of outcomes. For oil-rich counties, we analyze income, education, and manufacturing employment in the year 2000 at the county level. For mineral discoveries, we analyze the presence of population and transportation infrastructure located in 1km^2 cells within 5, 10 and 30 miles of a mineral discovery. For the county-level analysis of oil discoveries, we estimate the following equation for the full sample of counties and separately for each amenity and transport cost bin:

$$Y_{i,s} = \alpha + \beta * E_i + X_i + \delta_s + \epsilon_i, \tag{6}$$

where E_i is equal to one if the county ever made a major oil discovery, δ_s is state fixed effects, and X_i is a set of controls. Controls include dummies for each amenity and transport cost bin (where applicable, bin dummies are dropped when restricting the sample according to those same bins), the natural log of population density in 1890 and the share of manufacturing employment in 1890.[13] These latter two controls are intended to provide a pre-oil discovery baseline in population and manufacturing employment, the latter of which serves as a rough proxy of income or development at the time. We therefore roughly interpret these results as a two-period difference-in-difference over the long term. This analysis is similar to that of Michaels (2011) but extends it by including all lower 48 states and exploring heterogeneous effects by location favorability.

We examine several development outcomes as of 2000: population density, share of manufacturing employment, income, poverty rates, and education levels. Again, the theoretical model predicts smaller effects of resources on population density for favorable locations. To the extent that resource discoveries created agglomeration effects (Alcott and Keniston, 2017), we expect to find a positive effect on the share of labor in manufacturing, perhaps especially for discoveries made in counties with low transport cost. However, if sustained resource extractions works to pull labor out of manufacturing due to Dutch Disease, we might expect to find the opposite effect (Corden and Neary, 1982). To the extent that resource extraction persisted

[13] 1890 manufacturing employment data comes from the data set used in Michaels (2011), which the author generously provided us.

to the year 2000, or that population density yields persistent agglomeration effects, we may expect to find positive treatment effects for income and negative effects for poverty. Educational quality may improve as a results of bolstered local tax revenue and spending (Maurer, 2018), while at the same time enhanced returns to low skilled labor may increase dropout rates (Cascio and Narayan, 2017). There may also be important heterogeneous migration effects that influence measured educational attainment and income.

For the cross-section analysis of mineral discoveries, we estimate a variation of Equation 6 for the full sample and for each of the amenity and transportation cost bins. The treatment variable, E_i, is equal to 1 for cell i if that cell is within 30 miles of a world class mineral deposit discovered before 2000, and zero otherwise. We use this specification to estimate the extensive-margin effect of mineral discovery on population in the year 2000, defined as whether cell i has population ≥ 1. We also estimate the effect of mineral discovery on the presence of transportation infrastructure in cell i in the form of roadways and rail lines.

4 Amenity Value & Transport Cost

Our analysis considers heterogeneity across two dimensions of location favorability: local natural amenity values and transport cost to the nearest market. In this section we describe the construction of novel measures of both. Local amenities are utility improving characteristics such as temperature, humidity, soil productivity, and sunlight hours. Transport cost is based on the distance from—and topographic features in between—a particular location and the nearest market.

4.1 Estimation of Amenity Value

We construct an index of amenity value that reflects a large variety of strictly exogenous, naturally occurring environmental characteristics. These characteristics are theoretically utility improving for residents, and as we describe in our theoretical model, higher levels of amenity should, all else equal, result in larger population levels as people "vote with their feet". We leverage this intuition in constructing our index. Namely, we regress the total population living in a county on a vector of county-level amenity characteristics and state fixed effects. The resulting regression coefficients (excluding state fixed effects) are then used to construct a county-level amenity index, as described below.[14]

For a given census year, we flexibly model amenity-driven population using the following specification:

$$ln(Pop)_{i,t} = \sum_k f(a_{k,i}) + s_i + \epsilon_i, \quad \text{for} \quad t \in [1870, 2000], \tag{7}$$

where Pop_{it} is the natural log of population density in county i in decade t, s_i are state fixed effects, and a_k refers to a particular amenity k for county i. Recognizing that the effect of continuously measured amenities (such as temperature) may be highly non-linear, we flexibly model the effects of such amenities with a set of decile bins for each amenity. These amenities are average January temperature, average July temperature, average January sunlight, average annual humidity, average annual rainfall, soil quality, and terrain ruggedness. We also include dummy variables that identify counties on the coast, near (within 50-miles of) a mountain range, and within a mountain range. Finally, we control for a location's transportation cost using the estimates described in Section 4.2.

To calculate the index value for a given county, we multiply the coefficient values from Equation (7) by the county's amenity variables, but exclude transportation cost (so the index is conditional on transport costs) and the state fixed effects. Therefore we are identifying amenity values using within-state variation, but then applying these values across the country without including state averages, which are impacted by many other unobservable factors.[15]

[14] Our estimation of amenity value fits into a large literature that estimates the value of non-marketed amenities. Following Rosen (1974, 1979), hedonic price models assume that environmental amenities are capitalized into local prices. Examining wages and housing prices in the United States, Albouy et al. (2016) estimate willingness to pay (WTP) for a variety of exogenous environmental amenities including sunshine and temperature. Population density is also strongly correlated with exogenous, non-marketed environmental amenities such as temperature, sunshine, and precipitation (Albouy and Stuart, 2014). Because we are interested in estimating historical non-marketed amenity value, and lacking detailed historical data on wages and land prices, we use predicted population density as a proxy for amenity value.

[15] Suppose we did not include state fixed effects in Equation (7). Then this analysis would be vulnerable to over-fitting. For example, the Eastern U.S. is much more densely populated than the west, which may be for many reasons unrelated to amenities. Without state fixed effects the model would simply weight amenity traits

Finally, we allow for the effect of these amenities to change in value over time by estimating Equation (7) separately for each decade, and likewise construct the index separately by decade. The index is then averaged over time to provide a single value of amenity that reflects both contemporary and historical preferences for various amenities over our sample period. For our cell-level mineral analysis, cells are assigned the amenity index value of the county they lie within. We use this information in our analysis of resource discoveries by assigning counties (or cells) to either high, medium or low amenity groups based on 20th and 80th percentile cutoffs.

Panel (a) of Figure A.4 describes environmental amenity value by county. For comparison, panel (b) maps population density in 2000. Several notable features emerge. First, the correlation between our amenity index and observed population density in 2000 is quite strong. Our model predicts that coastal counties should be more heavily populated, as well as mid-western counties east of the Rocky mountain range. The model also predicts that the Salt Lake Valley and Denver county (which are relatively flat places, but near mountains) should be more heavily populated relative to nearby counties, which we also observe. We also predict that the northern part of the U.S. should be less populated due largely to below average temperatures. There is some error in the model. For example, whereas we predict southeastern Arizona should be sparsely populated due to its very arid and hot climate, it is actually densely populated. We similarly predict that parts of northern California and southern Oregon should be more heavily populated than they are. For the year 2000, the correlation between our predicted log of population density and actual log density is 0.57.

Several sources of data are utilized to construct the amenity index. Data on temperature were collected from the USDA Natural Amenities Scale. Total January sunlight hours is measured as an average from 1941-1970.[16] Data on annual rainfall (averaged over 1961-1990) by county comes from the USGS.[17] County-level soil quality data come from Schaetzl, Krist,

that are common in the east. Because we are using many amenity traits, the within-sample predictions would likely be very good even if all predictors were in fact irrelevant. But by conditioning on state fixed effects, we eliminate correlations based on broad regional patterns of population density, and identify amenity value based only on within-state variation. Then, excluding state fixed effects from the amenity index calculation ensures that we only base the index on observable amenity traits that apply equally to all counties.

[16]See Figure A.3 for a spatial description of each of these USDA inputs to the amenity value index.

[17]The data is available at: https://catalog.data.gov/dataset/united-states-average-annual-precipitation-

and Miller (2012) and measures exogenous "natural native soil productivity" which reflects soil characteristics such as soil tilth, clay mineralogy, organic carbon content, and the presence of root-impeding layers. Our spatial mountain range shapefile comes from "Landforms of the World", from ArcGIS online. The measure of terrain ruggedness for 1km^2 cells is described in the following subsection. For the county-level amenity index, we use the average ruggedness measure for cells within a county.

4.2 Estimation of Transportation Cost

This subsection describes the construction of the transport cost variable used in our main analysis. First, we describe the nature of the transportation cost problem. We then discuss a discrete formulation of the problem and how it is parameterized.

Broadly, we are interested in the variable cost, $C_{f,m}$, of moving goods to and from a frontier location f, to a nearest market, m. One approach would be to proxy for $C_{f,m}$ using "as-the-crow-flies" distance between f and m as in Redding and Venables (2004), for example. However, this distance ignores potentially important topographic features (mountains, marshland, large lakes and rivers) which may exist between the frontier and markets. In a practical U.S. context, Denver, Colorado is approximately as close to the port of San Francisco as it is to the port of Houston (approximately 1,500 kilometers). However, between Colorado and San Francisco are the Rocky Mountains, Great Basin Ranges, and Sierra Nevada Mountains which pose significantly higher transportation costs than the relatively flat Denver-Houston route.

A second option might be to refine the distance proxy by calculating the cost along realized modern transportation networks (as in Donaldson and Hornbeck, 2016). This method acknowledges topography between f and m via the realization of the network; engineers incorporate the cost of building transportation networks in rugged terrain. A problem with this method in our setting is that the spatial pattern of the contemporary transportation network is potentially endogenous to historical resource discoveries as intermittent cities develop nearby.

1961-1990-direct-download.

Given the shortcomings of these alternatives, we construct an exogenous measure of transportation cost that exploits important topographic features between places. Our estimate is based on a transportation cost framework common in linear and network programming. The transportation cost problem conceptualizes the United States as a grid of 1km x 1km cells, with each cell i having associated cost, c_i, which is incurred to pass through it. Transportation cost is a function of the distance across the cell (which in our case is fixed at 1km), and the topography of the cell. Each frontier location f seeks the least cost path to any market location m across this grid. Because the actual cost of transportation within any given cell is unobserved, we develop an approximation of the transportation model's cost parameters inferred by the structure of the modern rail network. From the solution to the transportation cost problem, the likelihood of the presence of modern rail in a given cell is a decreasing function of the cost associated with traveling through that cell. We estimate this relationship using a logistic regression model of the form:

$$\text{rail}_{i,2017} = f(\text{terrain}_i) + \epsilon_i, \tag{8}$$

where the binary variable $\text{rail}_{i,2017} = 1$ if cell i contains rail as of 2017 and 0 otherwise. $f(\text{terrain}_i)$ is a function of the ruggedness of the terrain within the cell and the percentage of the cell containing water. Intuitively, both uneven terrain and bodies of water such as lakes and rivers pose significant and expensive challenges for the construction of a railroad. To measure ruggedness, we use the National Elevation Dataset (NED), which provides elevation measures at a 30x30 meter resolution. We define ruggedness as the standard deviation of elevation within each 1km^2 cell in the grid. To avoid making assumptions about linearity, we include indicator variables for every percentile of ruggedness within the sample. The percentage of water within a cell is measured by decile-bin indicators for having no water, between 0-10% water, 10-20% water, and so on.

We use the estimated model to predict fitted probability values of observing rail in a given cell, i.e. $\text{P}(rail|terrain) = \hat{rail}$ for the entire grid. Next, we transform the model's fitted values into an approximation of cost, using the inverse of rail probability: $\hat{c}_i = 1/\hat{rail}$. This transformation is somewhat arbitrary, but makes the reasonable assumption that a cell with half the probability of containing a rail line is twice as costly to traverse (and for our purposes only relative cell costs are important since we ultimately evaluate heterogeneous treatment effects by relative differences in total transport costs). Intuitively it makes sense that the relationship between the probability of rail and rail construction cost is non-linear. When construction cost is very high, increasing construction cost arguably has little effect on the probability of rail. But when cost is low, increasing cost is more likely to make alternative routes more appealing from a cost-minimization standpoint. Additionally, simulation analysis in Appendix C shows that this transformation produces results that align closely to the rail structure we are trying to predict.

Figure A.8 presents the estimates from Equation 8 of the change in rail probability associated with values of terrain ruggedness and water. We generally find a consistent negative relationship between rail probability and ruggedness, with an especially steep drop in probability at more extreme values. However, we do oddly find a steeply increasing rail probability for the first few percentiles (i.e., the most flat).[18] The percentage of water coverage is also generally negatively associated with rail probability.

With the transportation cost problem parameterized, we populate our U.S. grid with our estimated cost values and solve for the lowest-cost path between every cell and the lowest-cost market destination.[19] We define market destinations as rail lines existing as of 1870, marine highways, and pre-existing cities of at least 5000 people.[20] In this way, our transportation cost measure captures the difficulty of constructing a new rail line to access the existing rail or waterway network (which then allow goods to be transported to any market destination at relatively low cost) or to a city plausibly large enough in size to serve as a market in its own

[18]This fact is at least partially explained by the fact that many relatively flat cells lie very near the coast, and rail lines are rarely so close to the ocean. Many other flat cells are observed in the remote salt flat region of Utah and Nevada, and also the upper reaches of the Northern Midwest where glaciers carved the landscape. Because few people live in these largely remote, and yet relatively flat terrains, there are few rail lines present. However, the overall relationships shown in Figure A.8 are robust to state fixed effects, further suggesting that near perfectly flat terrain tends to be indicative of somewhat inhospitable land (aside from coastal cells)

[19]This is done with the "Cost Distance" tool in ArcGIS.

[20]For the analysis of oil discoveries we use cities of at least 5000 as of 1900, since this is shortly before the vast majority of oil discoveries. For the analysis of mineral discoveries we use cities as of 1850, reflecting the timing of most mineral discoveries.

right. The cost of traversing this lowest-cost path is then our cell-level transportation cost measure.

In addition to the simulation exercise described in the Appendix, we further evaluate the accuracy and soundness of our transportation cost measure by comparing the predicted lowest-cost paths between cities to the actual rail network in four separate case studies. We do not expect to find that predicted paths perfectly match actual ones; by design our prediction is based purely off of exogenously determined geographic features of the landscape and in reality, transportation infrastructure reflects a number of endogenous factors including the location of pre-existing markets. We nonetheless use these case studies to shed some light on the predictive power of our estimates, and to test whether our model simply predicts linear least cost paths. Figure A.5 shows these four instructive examples in the western US, where the prevalence of mountains will often make the lowest-cost path indirect. The first example, Santa Fe to Salt Lake City, shows that the modern rail network is endogenous to regional economic development. We estimate the lowest cost path as a relatively straight line through mostly flat terrain. In reality, the rail network heads north out of Santa Fe to Denver Colorado, or west to Los Angeles, before heading to Salt Lake City, Utah. The Santa Fe, New Mexico to Salt Lake City, Utah connection is likely not important enough to build an additional more direct route. A similar argument likely holds for the Boise-Sacramento connection. However, our predicted path aligns very well with the actual rail lines between Denver, Colorado and Las Vegas, Nevada, which avoids a route that is much more direct but covers more rugged terrain. Finally, the Boise-Helena route offers perhaps the most straightforward test of our cost parameterization: both cities are important enough to be rail hubs, there are few or no other major cities in the area that might divert rail lines and there are mountains blocking the most direct path. In this case both our predicted path and the actual rail network go around the mountains and through the valley to the south in a very similar way, demonstrating that our procedure is at least in some cases applying an appropriately high penalty of mountain terrain in finding the lowest-cost path.

Figure A.6 shows cell-level transportation costs, along with 1870 rail lines, marine highways, and cities of at least 5000 (green dots). The results in panel (a) limit the cities that serve as markets to those that had populations greater than 5,000 by 1850 (used for the mineral deposit results), while for panel (b) cities with populations greater than 5,000 by 1900 also serve as markets (used for the oil results). Note how costs increase rapidly as one moves east from the headwaters of the Columbia River in Washington State into Idaho and Montana, while costs rise more gradually as one moves west from Minnesota into the Dakotas. These figures show the locations with the highest transportation costs are Montana, Wyoming, North Dakota, the Four-corners area (New Mexico, Utah, Colorado and Arizona) and Western Texas. Population expansion between 1850 and 1900 reduced the isolation of these locations to some degree, but many areas in the Western United States still remained highly isolated.

5 Data

In this section we describe the data sources used in the estimation of equations (5) and (6). We collected additional data from a variety of sources for the estimation of county-level amenity value and transport cost. That data is discussed in detail in the corresponding amenity and transport cost estimation subsections in Section 4.

5.1 Oil Discovery Data

Data on large, historical oil field discoveries were collected from the Oil and Gas Journal Data Book (2000). The book identifies, by year and field name, all U.S. discoveries of oil fields containing at least 100 million barrels of oil up to the year 1999, of which there were 263. Oil fields were matched by name to U.S. counties using the Energy Information Administration (EIA) Oil and Gas Field Code Master List, 2015. Matching oil field discoveries to U.S. counties decreases the number of observed discoveries to 231.

Panel (a) of Figure A.1 shows the spatial distribution of the major oil field discoveries. The majority of discoveries took place in a handful of states: Texas, California, Wyoming, Montana, Utah, Oklahoma, Louisiana, and New Mexico. Though a number of discoveries were also made in North Dakota, Mississippi, Kansas, Colorado, Arkansas, Alabama, and Illinois. Figure A.2 describes the temporal distribution of the discoveries. The earliest discovery was made in 1880, and the latest in 1988. The large majority of the discoveries were made early in the 20th century. For example, roughly 85% occurred prior to 1950 and 45% occurred prior to 1930. Only 9 discoveries (3.4%) occurred post 1970.

For context, we estimate the gross in-situ value of oil fields at the time of their discovery. Using data from BP's Statistical Review of World Energy, we estimate that the median discovery has a value of $2.6B in 1998 USD. The 5th percentile field is $1.2B and the 95th percentile is $18B.

5.2 Mineral Discovery Data

Data on the location (geographic coordinates), date, and significance of metallic mineral discoveries comes from the USGS's Mineral Resources Data System (MRDS). MRDS characterizes the size of mineral discoveries as being of "world class significance" or not. This is a qualitative designation by the USGS, but is based on the total endowment of contained commodity, which includes all past production and reserves. Each commodity is considered separately for this designation. In other words, the tonnage required for world class significance is different for gold than for iron. As described in Section 3, a number of deposits in MRDS do not have a known discovery date. In this case, we assign the date the deposit first came into commercial production as the treatment year. If this date is also missing, we drop the deposit from the sample.

As shown in Figure A.1, mineral resources exist primarily in the western United States. While the eastern United States contains a significant number of individual coal mines and non-metallic mineral quarries, we do not consider these resources in our analysis. Figure A.2 describes the temporal distribution of the world-class discoveries within the western US. The earliest discovery was made in 1800 (though the next one wasn't until 1848), and the latest in 1981. The majority of the discoveries were made late in the 19th century. Roughly 74% occurred prior to 1900, and 43% occurred prior to 1880. A little over one third of the deposits produce gold as the primary commodity, but the sample contains a number of copper, zinc, iron and molybdenum deposits as well. As with oil, we estimate the gross in-situ value of world class mineral deposits at the time of their discovery, with mineral price data assembled from USGS Series 140. For the 65 deposits with sufficient data on grade and tonnage, we estimate that the median discovery has a value of \$302m in 1998 USD. Relative to oil fields, mineral deposits have a much larger range of values. The 5th percentile deposit is \$0.6m and the 95th percentile is \$31.5B.

5.3 Economic & Demographic Data

We estimate the long-run effect of resource discovery on nearby population and other socioeconomic outcomes, including per capita income, population density, labor shares, transportation infrastructure, and poverty rates. Our population data comes from Fang and Jawitz (2018), which provides geospatial population count estimates at a 1km^2 resolution for every decade from 1790-2010 except for 1960, for which digital urban population data are missing. Estimates are based on county-level census counts, which are then down-scaled to the 1km^2 level using five models of increasing complexity. We use the most complex model "M5", which is demonstrated to be the most accurate. This specification utilizes the spatial extent of urbanization and non-inhabitable areas (e.g. water bodies and mountains), topographical features and measures of "socioeconomic desirability" to disaggregate county population counts from the decennial Census. For our county-level analysis of oil resource discoveries, we aggregate the 1km^2 resolution population maps up to modern county boundaries. In this way, we have a consistent spatial unit for analysis over time. For our analysis of mineral discoveries, we use the 1km^2 data directly.

Data for our cross-sectional analysis come from several sources. All socioeconomic outcome variables are measured in the year 2000, which is three years prior to the start of the oil price boom of the 2000s, and roughly 20 years after the previous oil price boom of the late 1970s. Labor shares, per capita income, and population data were collected from the Bureau of Economic Analysis (BEA), Regional Economic Accounts. Population density was computed by dividing population by county land area measured in square miles (land area data were collected from the U.S. Census Bureau, USA Counties Database). Using per capita income as a measure of prosperity may be misleading for a couple reasons. First, regional inflation can lead one to overstate regional real income levels. Second, average income levels say little about how income is distributed. We therefore use the poverty rate as an additional outcome variable, with data collected from the Census Bureau's USA Counties Database. County-level data for the percentage of people with only a high school degree, or with at least a college degree, were collected from the Census Bureau and American Community Survey. Unemployment data across counties were collected from the Bureau of Labor Statistics, Local Area Unemployment Statistics.

The unconditional means of these various economic outcomes, as well as the amenity and transport cost estimates, are presented for oil rich and oil poor U.S. counties in Table A.1. Cell-level population statistics for mineral-rich and mineral-poor cells are given in Table A.2.

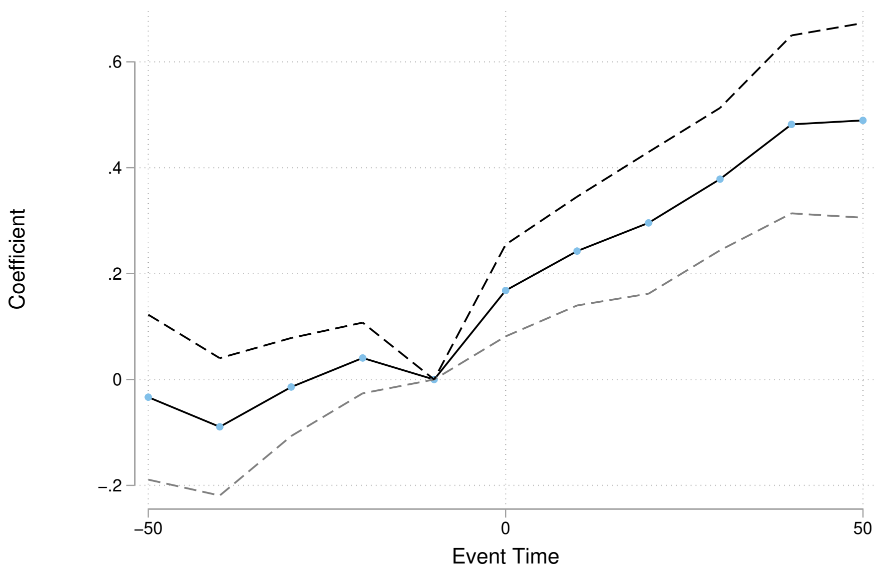
6 Results

6.1 Dynamic Analysis of Population

On average, large-scale resource discoveries cause local population growth that can last well past (>50years) the time of discovery. Decomposing this average effect for both oil and mineral discoveries, we show oil discoveries generate persistently higher local populations relative to their non-discovery counterparts. The population-level differential around mineral discoveries is more transient, peaking around 20 years after discovery. Further dis-aggregating the results reveals geographic heterogeneity in the population dynamics. Resource locations with favorable amenities and market access develop similarly to favorable non-resource locations. The average population effect of resource discovery are instead driven by the growth experienced in moderate and unfavorable locations. Finally, we show the size and longevity of resource discoveries also shape the population growth levels and dynamics.

Estimated from Equation (5), Figure 2 plots the estimated effects of oil discovery on logged population density by event-time for the 50 years before and after a discovery (with the first census decade prior to discovery as the reference period), along with 95% confidence intervals. Prior to discovering the resource, both oil-rich and oil poor counties conditionally grow at similar rates. However, population density in oil-rich counties is 20% larger than the counterfactual in the decade a discovery is made, and 65% larger after five decades.[21] Not all types of oil-rich counties experience this effect, however. Figure 3 plots the population dynamics associated with oil discovery when estimating Equation (5) with the sample limited to each of the six amenity and transportation cost groups. Unfavorable locations are those with low amenity values or high transportation costs, favorable locations have the reverse characteristics, and moderate locations fall in between. Unfavorable locations experience by far the largest population effects from oil discoveries, while moderate locations also see significant, albeit relatively smaller, effects. Favorable locations experience no significant effects. These findings persist through the end of the sampled post-discovery period with no indication of attenuating effects 50 years after discovery.

Figure 2: Log Population Density effect of Oil Discovery.

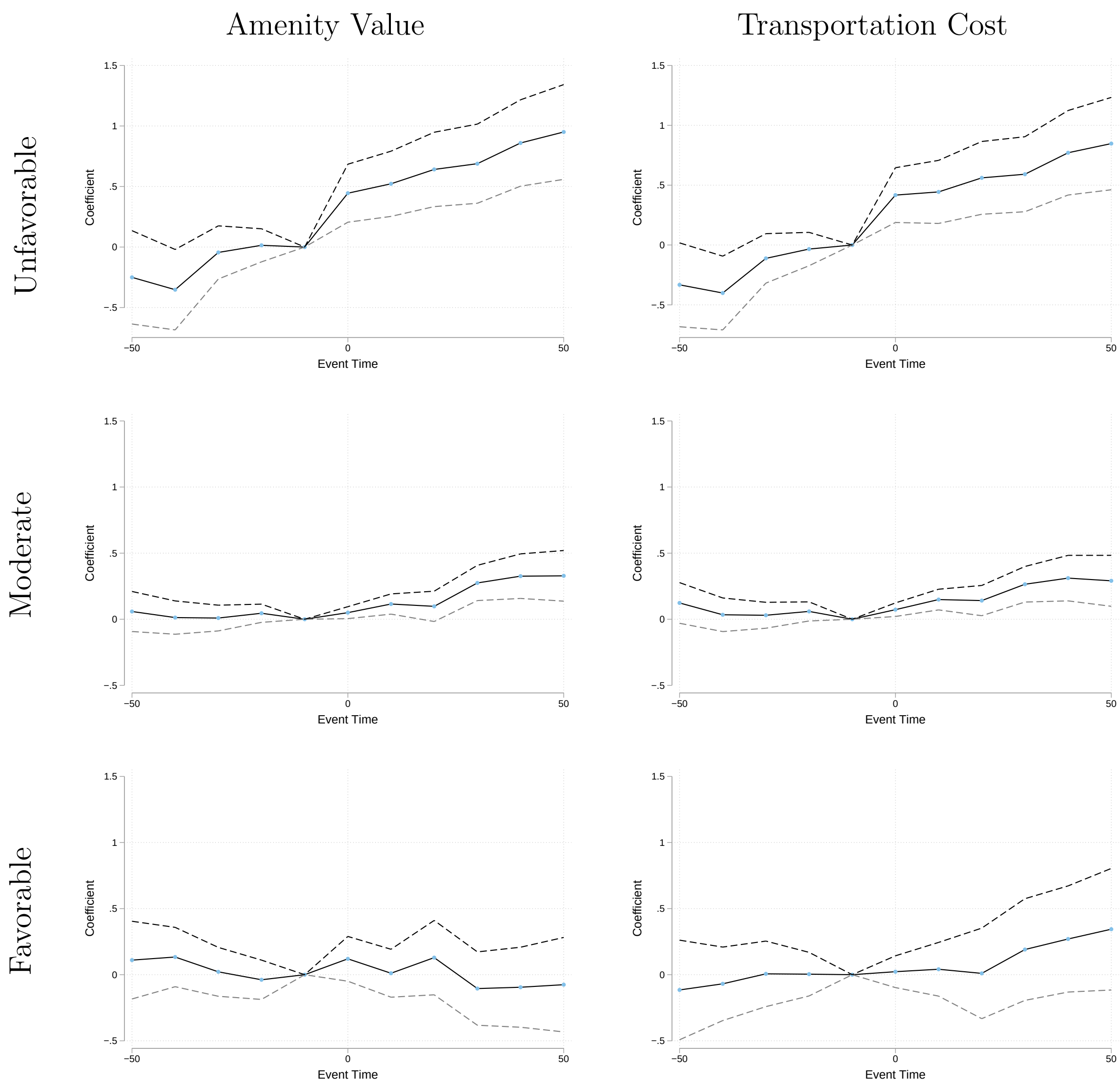


The graph plots the average effects of oil discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. In the pooled dataset population density is 20% (65%) larger in oil rich counties one (five) decades after discovery relative to counties without a discovery.

Figure 3 shows very similar effects for the low amenity and high transport cost bins. This is due to similar treatment groups for both bins. Although they are constructed independently,

[21]Note that the dependent variable is the natural log of population density. As such, β implies a $e^{\beta} - 1$ percent change in population density.

Figure 3: Heterogeneity in Population Dynamics of Oil-Discovery Counties.

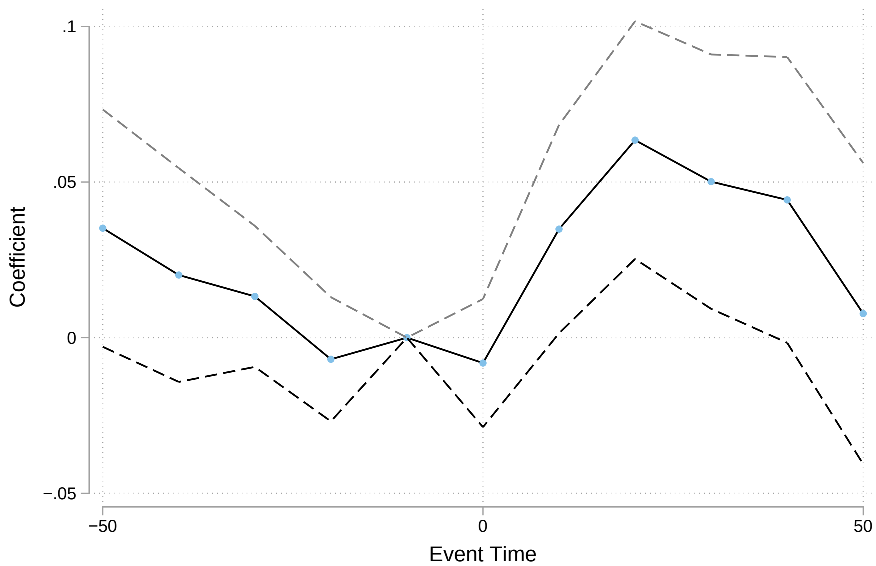


The graphs plot the average effects of oil discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. The data is partitioned based on the 20th and 80th percentile cutoffs into high, medium, or low amenity and transportation groups, and separate regressions are run for each of the six groups.

there is correlation in favorability between amenity and transportation cost measures. For the full sample, there are still many counties that are counted as unfavorable by one measure but not by the other. But within the treatment group, 90% of low-amenity counties are also high transport cost, largely due to a disproportionate number of counties in western Texas. Therefore, for the oil discovery analysis we cannot strongly separately identify effects for low-amenity vs. high transportation cost bins. This is less of an issue for the mineral discovery analysis that follows, as less than half of the low-amenity treatment cells are also high transportation cost. In any case, here the conclusion remains that effects are largest for low-favorability counties.

We next turn to population effects of mineral discoveries estimated from Equation (5) at the cell level and with the modifications discussed in Section 3. Figure 4 plots the estimated effects of mineral discoveries for the five decades before and after discovery. Specifically, it plots the effect on the probability that a given cell will be populated by at least one person if a mineral resource is discovered within 30 miles. A potential concern with identification of the effect of discovery on population is that population growth in particular places may lead to resource discovery (rather than the reverse). Figure 4 alleviates this concern by showing that prior to discovery, conditional population for mineral-discovery areas is actually trending down relative to non-discovery areas (though it is roughly flat in the two decades prior). Two decades after discovery, treated cells are over 5% more likely to be populated. However, this effect declines and is statistically insignificant after five decades.

Figure 4: Population effect of Mineral Discovery, 5 decades before and after

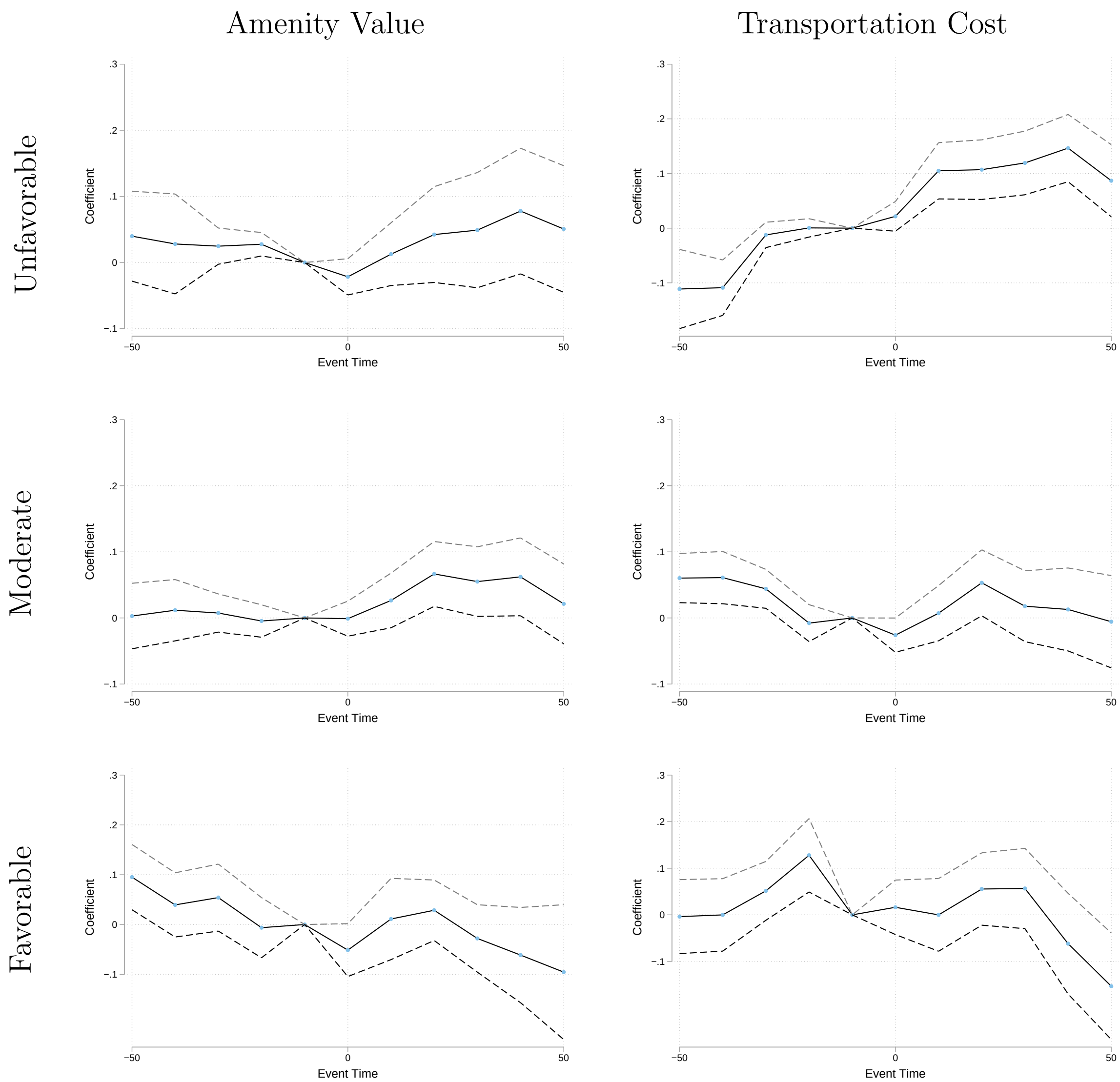


The graph plots the average effects of a "world class" mineral discovery on the probability that a 1km^2 cell within 30 miles of the discovery is populated. Effects are estimated using Equation 5. 95% confidence intervals are included.

Figure 5 plots the estimated effects of mineral discovery when limiting the sample to each of the six amenity and transportation cost groups. Unfavorable locations experience positive population effects after mineral discovery, particularly for those places with high transportation costs. After five decades, low amenity cells within 30 miles of a mineral discovery are approximately 7% more likely to be populated while high transportation cost cells within 30 miles are 15% more likely to be populated. For moderately favorable locations, treated cells are approximately 5% more likely to be populated two to four decades after mineral discovery, with the effect disappearing after five decades. We observe similarly transient effects for moderate transportation cost cells. Finally, in favorable locations, there is no evident population boom after discovery but there is some evidence of a negative effect in four to five decades after discovery. The negative results for these more favorable locations is likely a driver of the boom-and-bust pattern we observe in the full-sample effects in Figure 4.

Overall, the pattern of geographic heterogeneity for minerals is broadly consistent with the the evidence provided from the oil discoveries; locations with less favorable geography tend to

Figure 5: Population Dynamics of 30 mile Mineral-Discovery Buffer, 5 decades before and after



The graph plots the average effects of a "world class" mineral discovery on the probability that a 1km^2 cell within 30 miles is populated over time and is estimated by Equation 5. 95% confidence intervals are included. The data is partitioned based on the 20th and 80th percentile cutoffs into high, medium or low amenity and transportation groups, and separate regressions are run for each of the six groups.

experience larger gains from economic booms than locations with more favorable geography. However, the primary difference is that the positive population effects are not persistent out to 50 years for mineral discoveries. There are several possible reasons for this. First, our definitions of "significant" oil and mineral discoveries are not evaluated based on the same measures of their importance. We classify significant oil discoveries as those with more than 100 million barrels of reserves, while we adopt the USGS's qualitative definition of whether a particular discovery is "world class". The difference in average value of these resources at the time of their discovery is an order of magnitude. As discussed in Section 5, the median oil discovery has an in-situ value of \$2.6B compared to \$300m for minerals. Consistent with the idea that more valuable resources generate larger population shocks, we illustrate in Figure 6 that larger oil discoveries (those above the median number of in-situ barrels discovered) generate larger population effects.

There are also differences in the longevity of resource extraction. All oil discoveries in our sample were still producing at some level in the year 2000 (70 years after initial average discovery), while the median world-class mineral deposit in our sample had a 48 year mine life. It could be the case that the larger and more persistent oil discovery effect is being driven by the fact that these locations are still actively producing after 50 years while the mineral resources are, on average, depleted around this time. To test this hypothesis, we find a subset of oil-rich counties that did experience some depletion effects. While we do not observe annual production levels, we can observe extractive industry employment from 1969-2000 using Bureau of Economic Analysis data on employment by sector. In this period of available data, there was a large, positive employment shock from 1975 to 1990 in response to the oil embargo and other global geopolitical forces. All but 8 oil-rich counties in our data saw their employment in the extractive sector, both in levels and per capita terms, increase during this period. This boom was shared similarly in both older and more recently discovered oil fields, and in larger and smaller fields. For the remaining 8 counties, we cannot say their fields had been exhausted by this 1975-1990 period or even exhausted 50 years after discovery, but we can say they experienced this price run-up much differently than similarly sized fields discovered around the same time. If the persistence in the oil discovery effect was driven purely by robust and continuing extraction or subsequent commodity booms, we would be less likely to find such a persistent effect in these 8 counties. Figure B.9 shows the results for our main event study analysis when restricting the treatment group to these 8 counties. While the results are underpowered due to the small treatment group, they are suggestive of persistent and positive effects. Fifty years after discovery in these 8 counties, the magnitude of the treatment effect is similar to what we observe in the overall sample of counties.

These results are consistent with the idea that there may be more intrinsic differences between oil and mineral resources in terms of their effects on development. As mentioned above, the economic value of a typical oil field is much larger than that of a typical mineral deposit. Further, there are differences in the labor requirements of the start-up and operating periods of resource extraction, creating differences in the timing and intensity of labor demand. There are also differences in the transportation infrastructure needed to move the commodities to export markets.

6.2 Cross-Sectional Analysis of Development Outcomes

6.2.1 Long-Run Effects of Oil Discoveries

For oil, we observe a number of economic outcomes at the county level in the year 2000. We report the average effect of oil discovery for the full sample and separately for different favorability bins on these various outcomes in Table 1. The treatment effect (the effect of discovering a major oil field at any point in the past) is conditioned on the low and medium amenity and transportation cost bins. As such, the reference categories are the high amenity and high transportation cost bins. We also control for the share of manufacturing employment and population density in 1890.

The estimated average effect of oil discovery on population density is .443 log points, or roughly 55%, a result that aligns well with the event study analysis. Unsurprisingly, population density is significantly higher in favorable locations with low transportation cost or high amenity value counties. For our other outcomes, we find that as of the year 2000 historical oil discoveries are associated with more poverty and a less educated population. The latter result is consistent with the finding of Cascio and Narayan (2017) that local fracking booms led to increased high school dropout rates as the returns to low-skilled labor increased. We also find that low amenity value and high transport cost is associated with less income, less manufacturing employment (as a share of total employment), more poverty and more unemployment. This finding complements the existing literature that documents a positive relationship between the location of income and amenity value (Brueckner *et* al., 1999). Taken together, we find that both environmental amenity and geographic isolation are significant determinants of regional development and explain a significant proportion of the spatial pattern of regional development in the United States. The long-run effect of oil discoveries appears to be most significant for population density, and tends to have a negligible effect on the broader set of outcome variables.

Table 1: Average Effects of Oil Discoveries, Amenity Values, and Transportation Costs

	ln(Pop Dens.)	ln(Med. Inc.)	Share Man. Emp	Poverty Rate	% W/out HS Deg.	% W Coll. Deg.	Unemp. Rate
Variable:	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)
Had Oil	.443***	-.010	-.010*	.066**	2.540***	-.293	.107
	(.117)	(.018)	(.006)	(.033)	(.737)	(.596)	(.162)
Man Emp 1890	6.926***	.042	-.038	1.030***	-7.542*	19.62***	1.169*
	(.617)	(.111)	(.043)	(.214)	(3.945)	(4.861)	(.679)
ln(Pop. Dens.) 1890	.695***	.032***	.004***	-.030***	-.801***	1.776***	-.183***
	(.034)	(.004)	(.001)	(.009)	(.174)	(.200)	(.033)
Med. Trans.	.158**	.120***	.020***	-.189***	-2.459***	.349	-.405***
	(.076)	(.014)	(.006)	(.025)	(.514)	(.508)	(.105)
Low Trans.	.474***	.170***	.013*	-.221***	-3.727***	1.938***	-.439***
	(.089)	(.017)	(.007)	(.032)	(.622)	(.632)	(.120)
Med. Amen.	-.343***	-.074***	.005	.140***	.889**	-1.411***	.316***
	(.051)	(.012)	(.005)	(.023)	(.398)	(.455)	(.072)
Low Amen.	-.518***	-.081***	-.021***	.199***	-.623	.165	.403***
	(.094)	(.018)	(.007)	(.034)	(.597)	(.678)	(.139)
N	2,532	2,508	2,356	2,507	2,532	2,532	2,532

Note: Had Oil is equal to unity for counties that at any point discovered a major oil field. Robust standard errors are provided in parentheses below each estimate. Omitted favorability categories are 1) high amenity value bin and 2) high transportation cost bin. "Man Emp 1890" is the share of the labor force in 1890 in manufacturing. All outcome variables are measured in the year 2000. State fixed effects are included in all regressions. We include all counties with a non-zero population in 1870 to be consistent with the event study analysis.

The heterogeneous cross-sectional results are given in Table 2. Note that the "Total" column heading reports the average oil treatment effect from Table 1 for reference. The following six columns give the estimated treatment effect after limiting the sample of counties accordingly. For example, we estimate the effect of historical oil discoveries in low transportation cost counties by limiting the sample to only those counties in the low transportation cost bin, and these results are given under the column heading, "Low Trans.".

For population density, we find that the effect of oil discoveries on population density is especially large in counties with high transportation cost and low and medium amenity value, which is consistent with the dynamic result. For example, by the year 2000, historical oil discoveries led to an almost 200% increase in population density in low amenity counties, but had no effect in high amenity ones. A discovery similarly led to a roughly 115% increase in population density in high transportation cost counties, but again, had no effect in low transportation cost counties.

We generally do not find significantly heterogeneous effects for other outcomes with the exception of percentage without a high school diploma, where the effects are concentrated in low-favorability locations, perhaps reflecting that the reduced wage gap between low and high-skilled workers is even smaller in less favorable places.

6.2.2 Long-Run Effects of Mineral Discoveries

For mineral discoveries we again evaluate treatment effects at the 1km^2 cell level, and limit the sample to cells in western states. As with the cross-sectional results for oil, we first report the average effect of mineral discoveries, amenity value and transportation costs. We then present the heterogeneous treatment effects separately. We consider four different outcome variables: 1) an indicator for positive population in the year 2000, 2) an indicator for the presence of a road measured in the year 2017,[22] 3) road density (meters of road within a cell) in 2017, and 4) an indicator for the presence of railroad in 2017. As with our baseline event study analysis, treatment cells are defined as those within thirty kilometers of mineral discoveries. We condition on indicators for medium and high transportation cost and amenity bins, and for

[22]Geospatial road density data provided at the 1km^2 cell level comes from the NOAA National Geophysical Data Center, and is found on ArcGIS Online.

Table 2: Heterogeneous Effect of Oil Discovery by Amenity Value and Transportation Cost

ln(Population Density)							
	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Coef.	.443***	.172	.244*	.767***	1.05***	.256**	-.202
(Std. Err.)	(.117)	(.290)	(.133)	(.267)	(.315)	(.128)	(.206)
N	2,532	565	1,635	327	377	1,608	542
ln(Median Income)							
	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Coef.	-.010	-.005	-.019	-.026	-.011	-.022	.039
(Std. Err.)	(.018)	(.050)	(.024)	(.030)	(.032)	(.023)	(.052)
N	2,508	560	1,618	325	377	1,594	532
Manufacturing Employment Share							
	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Coef.	-.010*	-.013	-.015*	.003	-.001	-.014*	-.003
(Std. Err.)	(.006)	(.022)	(.008)	(.009)	(.008)	(.008)	(.019)
N	2,356	542	1,533	275	326	1,509	516
Poverty Rate							
	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Coef.	.066**	-.032	.080**	.091	.099	.058	.000
(Std. Err.)	(.033)	(.094)	(.038)	(.073)	(.086)	(.037)	(.103)
N	2,507	560	1,618	324	376	1,594	532
% Without High School Diploma							
	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Coef.	2.540***	-1.246	2.418***	5.147***	5.095***	2.071**	-.602
(Std. Err.)	(.737)	(2.12)	(.873)	(1.821)	(1.892)	(.897)	(.996)
N	2,532	565	1,635	327	377	1,608	542
% With College Degree							
	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Coef.	-.293	1.661	-.715	-2.04**	-1.974	-.271	.002
(Std. Err.)	(.596)	(2.280)	(.714)	(1.02)	(1.27)	(.718)	(1.21)
N	2,532	565	1,635	327	377	1,608	542
% Unemployment Rate							
	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Coef.	.107	-.206	.091	.469	.221	.053	-.213
(Std. Err.)	(.162)	(.360)	(.186)	(.419)	(.393)	(.189)	(.435)
N	2,532	565	1,635	327	377	1,608	542

Note: Each entry in the table is the estimated effect of being an oil discovery county. All outcomes are measured in the year 2000. Robust standard errors are provided in parenthesis below each estimate. Column headings (Total, Low Trans., Med Trans., etc.) describe which counties are being used in the estimation. For example, "Low Trans." gives the effect of discovery after limiting the control and treatment group of counties to those found in the low transportation cost bin. The "Total" column gives the results using all counties in the sample. Results are conditioned on manufacturing employment in 1890, the natural log of population density in 1890, the medium and low transport cost and amenity value dummies, and state fixed effects. We include all counties with a non-zero population in 1870 to be consistent with the event study analysis.

medium and high "inverse rail probability" bins which parsimoniously and flexibly account for cell-level ruggedness and water coverage (i.e. the fitted values from Equation 8 are discretized by 20th and 80th quantiles). We again cluster standard errors at the zip code level.

The full-sample average effect estimates are given in Table 3. We find that historical mineral discoveries have an insignificant effect on our measured outcomes with the exception of the presence of rail-lines, where the effect is qualitatively small. This is consistent with the event study analysis which found the average effect of a mineral discovery on population is insignificant after just 50 years. The average cell in Table 3 was treated over 100 years before. As in the county level analysis for oil, we also examine the main effects of location favorability. Similarly, we find cells with less favorable geography (more rugged cells that fall into the "High Inv. Rail Prob" bin, those with high transportation cost, and those with low amenity value) are less likely to have people living in them, and are also less likely to have significant transportation infrastructure.

Table 4 presents the heterogeneous treatment effects of mineral discovery based on transportation cost and amenity value. For each outcome we estimate the average treatment effect (the "'total" column) in addition to the six heterogeneous effects. As with the results for oil, each entry in Table 4 represents a separate regression with the sample limited according to the column heading. For robustness, we use three different treatment definitions: cells within 30, 10, and 5 miles of a mineral discovery (similar to the event study robustness check in Section 7).

The first three rows of Table 4 show that historical mineral discoveries generally have little effect on the likelihood that a cell is home to at least one person, with the exception of high transportation cost cells with treatment defined as within 10 miles of a deposit (the effect for high transportation cost cells within a 5-mile radius is qualitatively large but not statistically significant). Interestingly, the most consistent heterogeneous effect found for the three transportation infrastructure outcomes is also for high transportation cost cells in closer proximity to the deposit. One possible explanation for this is that resource revenues, along with the need to transport the resources to larger markets, contributes to building transportation infrastructure in places that are less likely to have infrastructure otherwise, and this may contribute to population growth in the longer run. However, caution is warranted with this interpretation; in the event study analysis using the 5/10-mile treatment radius

Table 3: Average Effect of Mineral Discoveries, Amenity Values, and Transportation Costs

	Pop> 0	Roads> 0	Road Density	Rail> 0
Variable:	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)
Treated	.027	.0002	-8.799	.003*
	(.023)	(.009)	(26.515)	(.001)
Med. Trans.	-.145***	-.080***	-385.60***	-.020***
	(.023)	(.011)	(38.604)	(.001)
High Trans.	-.092**	-.108***	-457.74***	-.020***
	(.040)	(.017)	(53.463)	(.002)
Med. Amen.	.074***	-.007	17.454	.005***
	(.026)	(.010)	(28.256)	(.001)
High Amen.	.244***	.047***	391.31***	.014***
	(.037)	(.014)	(48.393)	(.002)
Med. Inv. Rail Prob.	-.051***	-.061***	-367.42***	-.040***
	(.008)	(.004)	(16.716)	(.001)
High Inv. Rail Prob.	-.083***	-.318***	-934.62***	-.058***
	(.014)	(.008)	(23.669)	(.001)
N	2,752,848	2,752,848	2,752,848	2,752,848

Note: Treated is equal to unity for 1km^2 cells within thirty miles of a mineral discovery. Standard errors clustered by zip code are provided in parentheses below each estimate. Road Density is defined as meters of road per square kilometer. Pop> 0, Roads> 0, and Rail> 0, are equal to unity if a cell had at least one person living in it, had at least one mile of road, or at least one mile of railroad, respectively. "Inv. Rail Prob" represents the inverse of the predicted probability that a cell has a railroad (see the discussion on transportation cost estimation) and measures cell ruggedness. The omitted categories are low amenity value, low transportation cost, and low inverse rail probability.

shown in Figures B.4-B.5, we do not find population effects after 50 years, so the long-run population effects for these treatment definitions are somewhat ambiguous.

7 Robustness Checks

We carry out a battery of tests that highlight the overall robustness of the main findings. Starting with the results for oil discoveries, we first consider the possibility that oil discoveries generated significant spatial spillovers (as documented by Feyrer, Mansur and Sacerdote; 2017; James and Smith, 2019) which would cause treatment effects to be biased downward.

Table 4: Heterogeneous Effect of Mineral Discovery by Amenity Value and Transportation Cost

	Total	Low Trans.	Med. Trans.	High Trans.	Low Amen.	Med. Amen.	High Amen.
Population>0 (30 Mile Radii)							
Coef.	.027	.036	.044	.030	.0006	.043	.108
(Std. Err.)	(.023)	(.034)	(.031)	(.036)	(.060)	(.030)	(.304)
Population>0 (10 Mile Radii)							
Coef.	.038	-.024	.029	.082*	.053	.037	.035
(Std. Err.)	(.031)	(.045)	(.041)	(.045)	(.070)	(.038)	(.064)
Population>0 (5 Mile Radii)							
Coef.	.042	-.039	.047	.067	.061	.044	.028
(Std. Err.)	(.037)	(.061)	(.050)	(.061)	(.078)	(.047)	(.079)
Roads 2017>0 (30 Mile Radii)							
Coef.	.0002	.018	.018	-.016	-.016	.016	-.010
(Std. Err.)	(.009)	(.016)	(.012)	(.014)	(.021)	(.012)	(.017)
Roads 2017>0 (10 Mile Radii)							
Coef.	.042***	.034*	.056***	.038**	-.009	.078***	-.012
(Std. Err.)	(.011)	(.018)	(.016)	(.017)	(.019)	(.015)	(.024)
Roads 2017>0 (5 Mile Radii)							
Coef.	.093***	.031	.114***	.093***	.031	.128***	.045
(Std. Err.)	(.015)	(.024)	(.022)	(.019)	(.025)	(.019)	(.031)
Road Density 2017>0 (30 Mile Radii)							
Coef.	-8.79	-121.29	3.09	28.54	9.438	22.87	-129.70**
(Std. Err.)	(26.515)	(93.71)	(33.64)	(39.43)	(59.10)	(32.71)	(64.14)
Road Density 2017>0 (10 Mile Radii)							
Coef.	71.29*	-218.30	85.95*	103.97**	25.82	151.050***	-246.14**
(Std. Err.)	(38.66)	(155.00)	(51.04)	(50.89)	(61.85)	(48.57)	(101.07)
Road Density 2017>0 (5 Mile Radii)							
Coef.	214.36***	-263.91	286.50***	205.46***	81.50	321.80	-147.47
(Std. Err.)	(58.22)	(194.38)	(85.80)	(63.80)	(86.97)	(73.32)	(156.03)
Rail 2017>0 (30 Mile Radii)							
Coef.	.003*	-.007	.0004	.011***	.005*	.004**	-.003
(Std. Err.)	(.001)	(.005)	(.001)	(.002)	(.002)	(.002)	(.003)
Rail 2017>0 (10 Mile Radii)							
Coef.	.008***	.003	.0004	.012***	.0004	.014***	-.009
(Std. Err.)	(.003)	(.015)	(.003)	(.004)	(.004)	(.004)	(.006)
Rail 2017>0 (5 Mile Radii)							
Coef.	214.36***	-263.91	286.50***	205.46***	81.50	321.80***	-147.47
(Std. Err.)	(58.22)	(194.38)	(85.80)	(63.80)	(86.97)	(73.32)	(156.03)
N	2,752,848	550,508	1,651,756	550,584	538,688	1,663,536	550,624

Note: Each entry in the table is the estimated effect of being a treated 1km^2 cell. Standard errors clustered by zip code are provided in parentheses below each estimate. Column headings (Total, Low Trans., etc.) describe which cells are being used in the estimation. For example, "Low Trans." gives the treatment effect when limiting the sample to cells in the low transportation cost bin. The "Total" column gives the results using all cells in the sample. Results are conditioned on amenity bin dummies, transportation cost dummies, inverse rail probability dummies (see the discussion on transportation cost estimation), and state fixed effects. S

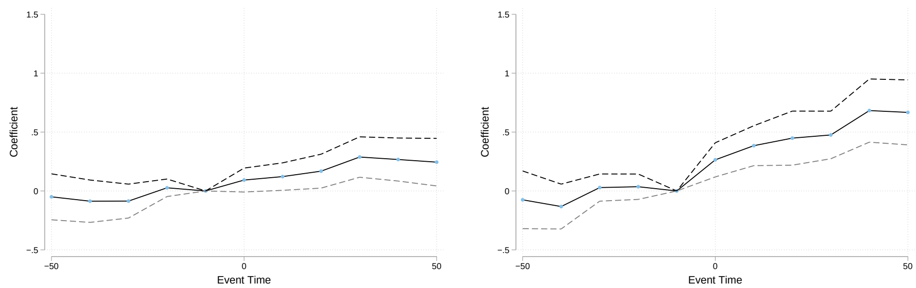
We address this concern by dropping all counties that are contiguous to treatments and re-estimate our baseline set of equations. These results are given in Figure B.1. For the full sample, fifty years after discovery, the average treatment effect is estimated to be 0.5, which is nearly identical to that found using the full set of counties. The effects by favorability are also very similar to before, suggesting the baseline results are not biased due to spatial spillovers.

Of clear concern is the idea that the timing of a discovery is endogenous to economic development within the region. If oil fields are discovered as a result of population growth, our results are potentially explained by reverse causality. We argue that this is unlikely affecting our results as first we note that pre-existing trends in the baseline event-study analyses are largely insignificant, and second, to the extent that most major oil fields had been discovered by the year 2000 (a fairly reasonable assumption), the treatment definition used in the cross-sectional analysis is exogenous and based largely off of geological factors. In addition to these arguments, we also examine whether estimated treatment effects are sensitive to the size of the fields being discovered. More specifically, we re-estimate baseline estimation equation 5 for both "high" and "low" oil endowments, defined as being above or below the treatment group median, respectively.[23] While exploration and search effort are potentially endogenous variables, conditional on discovering oil, whether a discovery is of a large or small field is plausibly determined by chance. The results are given in Figure 6 below.

Note that for both subsets of the treatment group, pre-existing trends are minimal. While short and long-run treatment effects are positive for both groups, fifty years after discovery population density is estimated to be roughly 90% greater in counties with high endowments and just 30% greater in counties with lower endowments. Even if one attributed the estimated effect of small discoveries purely to endogeneity, the "additional" effect of large discoveries is significant (90%-30%=60%) and should be considered a lower bound on the effect of discovering oil. Figures B.7 and B.8 give the heterogeneous effects by favorability for large and small endowments and largely reinforce the idea that effects are larger for low amenity, high trans-

[23]Field size is estimated by summing cumulative production and estimated remaining reserves for fields from the Oil and Gas Journal Databook. Total endowments are then calculated by summing together fields within a county. When the same field is assigned to multiple counties, it is assumed that endowments are divided proportionally by county area.

Figure 6: Population Dynamics for High and Low Oil Endowments



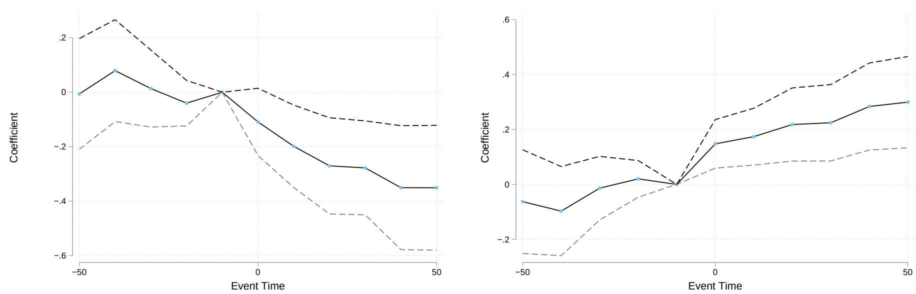
(a) Small Endowments (b) Large Endowments

Estimated average effects on log population density and 95% confidence intervals disaggregated by high and low endowments. "Low" endowments are those below the median of the treatment group and "High" endowments are above median.

portation cost places. For example, fifty years after discovering a large oil field, population density in low amenity counties is 350% greater than the pre-discovery counterfactual (this number is closer to 64% for small discoveries). There is no long-run effect of large or small discoveries in high amenity places.

Another possible concern with our finding that less favorable locations see larger effects is the possibility that oil discoveries have similar population effects in terms of raw numbers of people in all locations, but that low-favorability places have smaller initial populations, and so experience larger percentage increases following discovery. While this scenario would not change the overall conclusion that low-favorability places see larger relative impacts, it would imply that this is a result of low starting population rather than amenities or transport costs directly (note that this is less of a concern when examining mineral discoveries because the west was largely unpopulated with non-indigenous peoples prior to the first mineral discoveries in the mid-1800s). We attempt to separately identify how treatment effects vary by favorability and initial population by additionally controlling for the interactions of event-time indicators and log population density in 1870. For parsimony, in these regressions rather than separating by bins we include interactions of event-time indicators by the Z-scores of our amenity index or transportation cost measures (along with non-interacted event-time indicators). These heterogeneous effects are given in Figure 7 and show that counties with above-average amenity value (or below average transportation cost) experience smaller population gains resulting from oil discoveries, even after controlling for treatment interacted with starting population. These results again reinforce our baseline conclusions and quell concerns that our results are mechanical and simply reflect variation in initial population density across high and low amenity places.

Figure 7: Interaction of Event Year and Z-Score of Favorability



(a) Amenity Value (b) Transport Cost

The graphs show the coefficients for event-time interacted with the z-scores for amenity index and transport cost, conditional on the interaction of event-time and population density in the year 1870 and the non-interacted event-time indicators. Because amenity value and transport cost have been transformed to z-scores, these estimates are interpreted as the additional effect of discovery given one additional standard deviation in amenity index or transportation cost.

To the extent that regional resource discoveries generate state-wide economic growth and development, one can argue that including state-by-year fixed effects is overly restrictive as it exploits only within-state variation in outcomes. We therefore estimate an additional model that substitutes state-by-year fixed effects with year fixed effects. These results are nearly identical to before and are provided in Figure B.2.

We further examine the robustness of the heterogeneous effects by favorability by alternatively defining "high" and "low" favorability using 90% and 10% thresholds rather than 80% and 20% as in the baseline specification. These results are given in B.3 and are quite similar to our baseline results. We again document large treatment effects for low amenity places, smaller effects for moderate favorability places, and no effect for high favorability places.

Turning to the results for mineral discoveries, recall that for the baseline specification treatment, cells are chosen as those within thirty miles of a mineral discovery. Here, we let this radius vary from five, to ten, to fifty miles. Using a five-mile radius and the full sample, Figure B.4 shows that the average treatment effect is greater than in the baseline specification which uses a 30-mile radius. This makes some intuitive sense—if mineral discoveries attract populations, and populations are more heavily concentrated near the mine, one would expect larger treatment effects for smaller radii, though there are examples of boomtowns locating farther away in the nearest flat land. However, in this particular case, note that the estimated treatment effects for favorable locations, while somewhat erratic and imprecisely estimated, are actually larger than those for unfavorable locations. Similar results are found using a ten mile radius (see Figure B.5). The results using a fifty-mile radius are more in line with the baseline specification (see Figure B.6).

8 Conclusion

A large literature is devoted to understanding the determinants of where people live. Researchers generally agree that natural amenities matter; all else equal, people are more likely to live in places that provide the optimal set of environmental amenities like sunshine, productive soil, mountain views, and coastal access. But all else is not equal: labor market productivity, real wages, and consumption also matter. According to the regional economics literature, people "vote with their feet" and move to the place that maximizes the joint benefit of amenity and consumption. The importance of real wages in determining migration decisions is perhaps made most salient in the resource-development literature, in which real wages are exogenously shocked by commodity price swings or natural-resource discoveries. However, this literature almost entirely focuses on estimating the average effect of a resource shock, concealing significant heterogeneities and raising questions of external validity.

This paper explores the effect of permanent and temporary productivity shocks brought about by the discovery of large oil and mineral deposits in the United States. We estimate heterogeneous effects based on exogenous, naturally-occurring environmental and geographic features of the discovery site, using novel measures of amenities and transportation costs. More specifically, we estimate the effect of resource discoveries in places with varying levels of environmental amenity value (based off of observed characteristics like sun shine, temperature, soil quality, and mountain viewsheds) and geographic isolation (the cost of traveling to the nearest market).

Our findings provide several new insights relevant to the literatures on natural resource booms and economic geography. We show limited effects of productivity shocks in favorable locations in the short, medium or long term. We interpret this finding as reflecting an air of inevitable development in favorable locations. Just as cognitively gifted people benefit less from an education (Murnane, Willett, and Tyler, 2000), places with favorable environmental and geographic characteristics benefit less from a positive shock to productivity. But natural-resource discoveries play a more important role in the development process of less desirable locations, where we estimate large and sustained effects on population.

Our findings help to explain some of the observed heterogeneity in the estimated effects of resource shocks and the shale boom in particular in which the effect of hydraulic fracturing depends on the characteristics of the extraction site. We also provide strong evidence for the importance of geographic properties, and somewhat weaker evidence of path-dependence mechanisms for more isolated locations. The latter comes in the form of mixed long-run effects of relatively short-lived mineral shocks in especially isolated places. We find that path-dependence may be partially driven by development of transportation infrastructure in mountainous regions that is developed due to the mineral discovery.

Beyond environmental amenity value and geographical characteristics, we think that other important heterogeneities in the effect of economic shocks may exist. For example, limited work has been done on the joint role of resource shocks and institutional quality (Mehlum, Moene, and Torvik (2006) is one notable exception), tax rates, environmental regulation, and public policy more generally. Understanding these sources of heterogeneity is important for designing optimal public policy and helps to inform why, when, and where communities develop.

References

Albouy, D. and Stuart, B. (2014). Urban population and amenities: The neoclassical model of location. (No. w19919). National Bureau of Economic Research.

Albouy, D., Graf, W., Kellogg, R., and Wolff, H. (2016). Climate amenities, climate change, and American quality of life. *Journal of the Association of Environmental and Resource Economists*, 3(1), 205-246.

Allcott, H., & Keniston, D. (2017). Dutch disease or agglomeration? The local economic effects of natural resource booms in modern America. The Review of Economic Studies, 85(2), 695-731.

Allen, T., & Arkolakis, C. (2014). Trade and the Topography of the Spatial Economy. *The Quarterly Journal of Economics*, 129(3), 1085-1140.

Arnott, R.J. and Stiglitz, J.E. (1979). Aggregate land rents, expenditure on public goods, and optimal city size. *The Quarterly Journal of Economics*, 93(4), 471-500.

Berger, T., & Enflo, K. (2017). Locomotives of local growth: The short-and long-term impact of railroads in Sweden. *Journal of Urban Economics*, 98, 124-138.

Berman, N., Couttenier, M., Rohner, D., & Thoenig, M. (2017). This mine is mine! How minerals fuel conflicts in Africa. *American Economic Review*, 107(6), 1564-1610.

Betz, M. R., Partridge, M. D., Farren, M., & Lobao, L. (2015). Coal mining, economic development, and the natural resources curse. *Energy Economics*, 50, 105-116.

Bieri, D. S., Kuminoff, N. V., & Pope, J. C. (2014). National expenditures on local amenities. Department of Economics, Arizona State University, Manuscript.

Bleakley, H. and Lin, J. (2012). Portage and path dependence. *The Quarterly Journal of Economics*, 127(2), 587-644.

Blomquist, G.C., Berger, M.C., and Hoehn, J.P. (1988). New estimates of quality of life in urban areas. *The American Economic Review*, 89-107.

Brown, J. P. (2014). Production of natural gas from shale in local economies: a resource blessing or curse?. *Economic Review-Federal Reserve Bank of Kansas City*, 5.

Corden, W.M. and Neary, J.P. (1982). Booming sector and de-industrialisation in a small open economy. *The Economic Journal*, 92(368), 825-848.

Chay, K.Y. and Greenstone, M. (2005). Does air quality matter? Evidence from the housing market. *Journal of Political Economy*, 113(2), 376-424.

Davis, G. (2011). The resource drag. *International Economics and Economic Policy*, 8(2), 155-176.

Davis, D. R., & Weinstein, D. E. (2002). Bones, bombs, and break points: the geography of economic activity. *American Economic Review*, 92(5), 1269-1289.

Deaton, B. J., & Niman, E. (2012). An empirical examination of the relationship between mining employment and poverty in the Appalachian region. *Applied Economics*, 44(3), 303-312.

Donaldson, D., & Hornbeck, R. (2016). Railroads and American economic growth: A "market access" approach. *The Quarterly Journal of Economics*, 131(2), 799-858.

Englin, J. (1996). Estimating the amenity value of rainfall. *The Annals of Regional Science*, 30(3), 273-283.

Fang, Y.,& Jawitz, J. W. (2018). High-resolution reconstruction of the United States human population distribution, 1790 to 2010. *Scientific Data*, 5.

Felkner, J. S., & Townsend, R. M. (2011). The geographic concentration of enterprise in developing countries. *The quarterly journal of economics*, 126(4), 2005-2061.

Feyrer, J., Mansur, E.T., and Sacerdote, B. (2017). Geographic dispersion of economic shocks: Evidence from the fracking revolution. *American Economic Review*, 107(4), 1313-34.

Groothuis, P. A., Groothuis, J.D. and Whitehead, J.C. (2007). The willingness to pay to remove billboards and improve scenic amenities. *Journal of Environmental Economics and Management*, 85(4), 1094-1100.

Henderson, J. V., Squires, T., Storeygard, A., & Weil, D. (2017). The global distribution of economic activity: Nature, history, and the role of trade. *The Quarterly Journal of Economics*, 133(1), 357-406.

Jacobsen, G. D. and Dominic P. Parker. (2016). The economic aftermath of resource booms: evidence from boom towns in the American West. *The Economic Journal*, 126(593), 1092-1128.]

James, A. (2016). The long-run vanity of Prudhoe Bay. *Resources Policy*, 50, 270-275.

James, A. & Smith, B. (2017). There will be blood: crime rates in shale-rich U.S. counties. *Journal of Environmental Economics and Management*, 84, 125-152.

James, A. & Smith, B. (2018). The Geographic Dispersion of Economic Shocks: Evidence from the Fracking Revolution: Comment. *American Economic Review* (No. 2018-02).

Jedwab, R., Kerby, E., & Moradi, A. (2017). History, path dependence and development: Evidence from colonial railways, settlers and cities in Kenya. *The Economic Journal*, 127(603), 1467-1494.

Kline, P., & Moretti, E. (2014). People, places, and public policy: Some simple welfare economics of local economic development programs.

Kuminoff, N. V., Smith, V. K., & Timmins, C. (2013). The new economics of equilibrium sorting and policy evaluation using housing markets. *Journal of Economic Literature*, 51(4), 1007-62.

Marchand, J., and Weber, J. (2018). Local labor markets and natural resources: A synthesis of the literature. *Journal of Economic Surveys*, 32(2), 469-490.

Matheis, M. (2016). Local economic impacts of coal mining in the United States 1870 to 1970. *The Journal of Economic History*, 76(4), 1152-1181.

Maurer, S. E. (2018). Public education benefited from oil booms in the postbellum South. Working Paper. LSE Business Review.

McGregor, T., & Wills, S. (2017). Surfing a wave of economic growth. CAMA Working Paper No. 31/2017.

Mehlum, H. Moene, K., and Torvik, R. (2006). Institutions and the resource curse. *The Economic Journal*, 116(508), 1-20.

Michaels, G. (2011). the long-term consequences of resource-based specialization. *The Economic Journal*, 121(551), 31-57.

Michaels, G., & Rauch, F. (2017). Resetting the urban network: 117-2012. *The Economic Journal*, 128(608), 378-412.

Mideksa, T. K. (2013). The economic impact of natural resources. *Journal of Environmental Economics and Management*, 65(2), 277-289.

Miguel, E., & Roland, G. (2011). The long-run impact of bombing Vietnam. *Journal of development Economics*, 96(1), 1-15.

Motamed, M. J., Florax, R. J., & Masters, W. A. (2014). Agriculture, transportation and the timing of urbanization: Global analysis at the grid cell level. *Journal of Economic Growth*, 19(3), 339-368.

Muehlenbachs, L. Spiller, E., and Timmins, C. (2015). The housing market impacts of shale gas development. *The American Economic Review*, 105(12), 3633-3659.

Murnane, R.J., Willett, J.B., and Tyler, J.H. (2000). Who benefits from obtaining a GED? Evidence from high school and beyond. *Review of Economics and Statistics*, 82(1), 23-37.

Oil & Gas Journal Data Book. (2000). PennWell Books. Tulsa, OK.

Peach, J., & Starbuck, C. M. (2011). Oil and gas production and economic growth in New Mexico. *Journal of Economic Issues*, 45(2), 511-526.

Redding, S. and Venables, A. (2004). Economic geography and international inequality. *Journal of International Economics*, 62(1), 53-82.

Richter, J. Salanguit, A., James, A. (2018). The (Uneven) Spatial Distribution of the Bakken Oil Boom. *Land Economics*, 94(4), 577-592.

Roback, J. (1982). Wages, rents, and the quality of life. *Journal of Political Economy*, 90(6), 1257-1278.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55.

Rosen, S. (1979). Wages-based indexes of urban quality of life. In *Current issues in urban economics*, ed. P. Mieszkowski and M. Straszheim. Baltimore: Johns Hopkins University Press.

Sachs, J. (2001). Tropical underdevelopment. (No w8119). National Bureau of Economic Research.

Schaetzl, R. J., Krist Jr, F. J., & Miller, B. A. (2012). A taxonomically based ordinal estimate of soil productivity for landscape-scale analyses. *Soil Science*, 177(4), 288-299.

Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of Political Economy*, 64(5), 416-424.

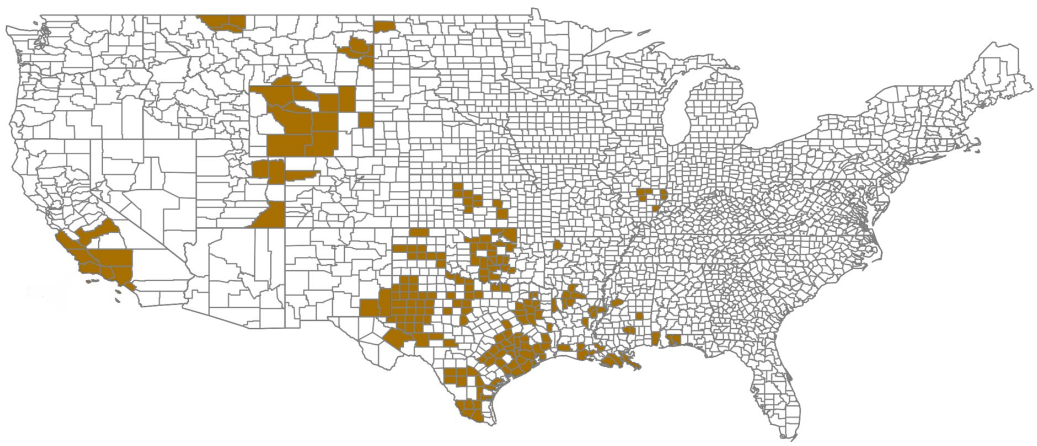
Waltert, F., Schläpfer, F. (2010). Landscape amenities and local development: A review of migration, regional economic and hedonic pricing studies. *Ecological Economics*, 70(2), 141-152.

Weber, J. G. (2014). A decade of natural gas development: The makings of a resource curse? *Resource and Energy Economics*, 37, 168-183.

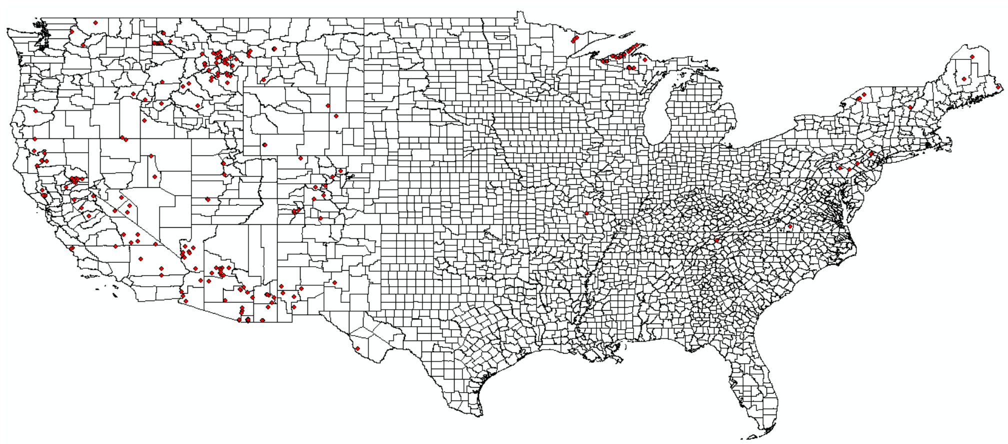
Weber, J. G. (2012). The effects of a natural gas boom on employment and income in Colorado, Texas and Wyoming, *Energy Economics*, 34(5), 1580-1588

A Additional Tables & Figures

Figure A.1: Mineral and Oil Discoveries



(a) Oil Discoveries



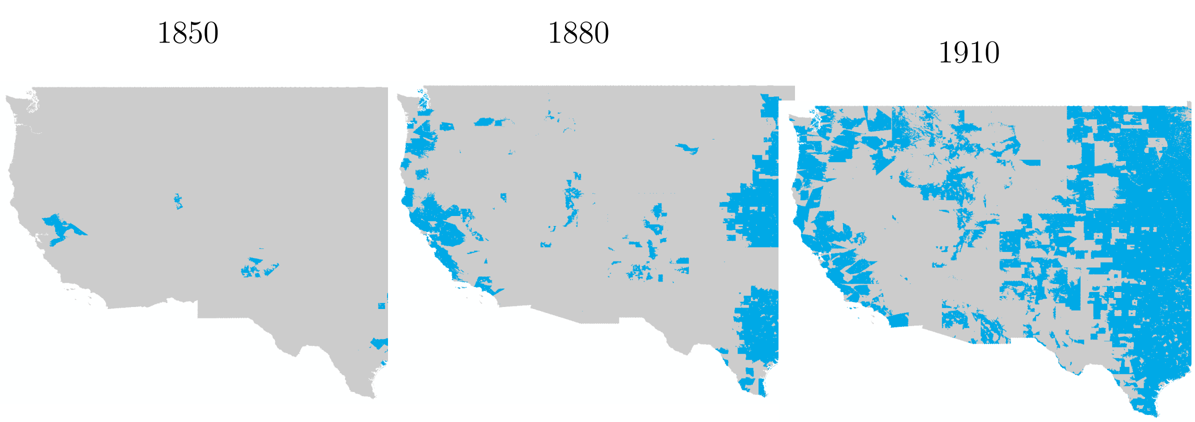
(b) Mineral Discoveries

Note: Oil discoveries are recorded at the county level whereas mineral discoveries are recorded by the latitude and longitude of the discovery site. Oil discovery locations come from the EIA. Mineral discoveries data come from USGS Mineral Resource Data System.

Figure A.2: Timing of Major Oil Field and Mineral Deposit Discoveries

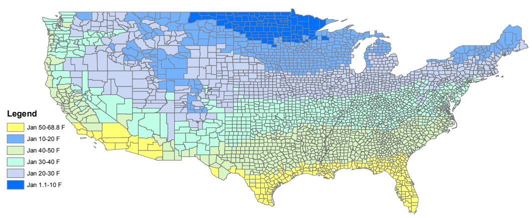
Note: Oil fields are major if they are larger than 100 million barrels. Mineral discoveries are major if they were qualitatively determined by the US Geological Survey to be of "world class significance". For mineral discoveries, if discovery date was missing, year of first production was used. Oil discovery dates come from the Oil and Gas Journal. Mineral discoveries data come from USGS Mineral Resource Data System.

Figure A.7: Population Dummy by Cell Over Time in Western US

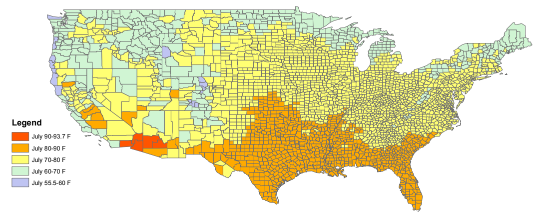


Note: 1km^2 cells are populated if they have more than one person living in them, according to the estimates by Fang & Jawitz (2018).

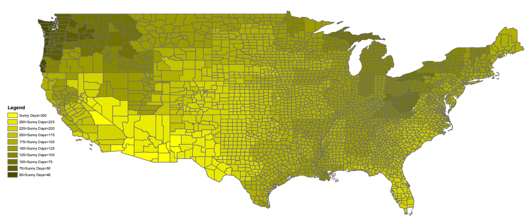
Figure A.3: Temperature, Sunshine, Humidity & Soil Productivity



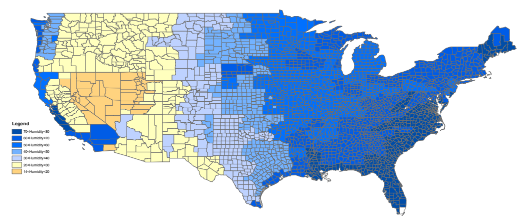
(a) Average January Temperature



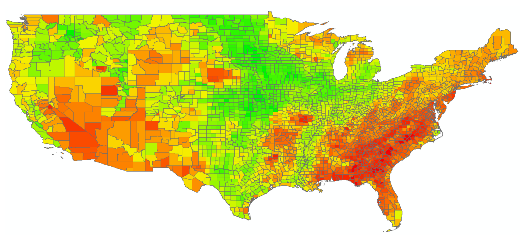
(b) Average July Temperature



(c) Average January Sunlight



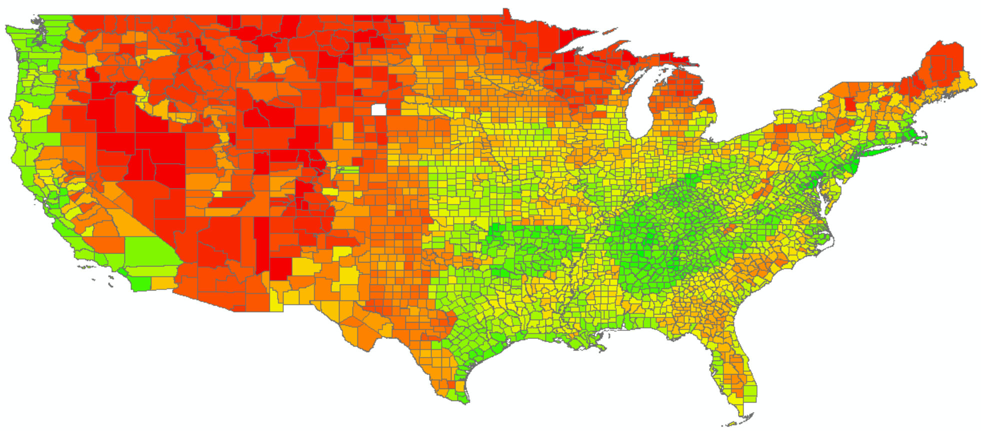
(d) Average Humidity



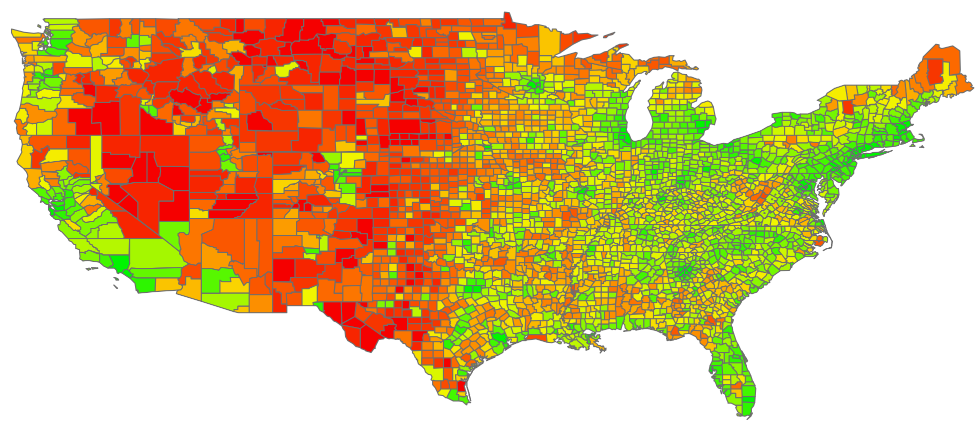
(e) Soil Productivity

Source: Figure from authors compiled from data from various sources. Data on temperature, humidity, and sunlight come from the USDA Natural Amenities Scale. Total January sunlight hours is measured as an average from 1941-1970. County-level soil quality data come from Schaetzl, Krist, and Miller (2012).

Figure A.4: Predicted Population (Amenity Index)



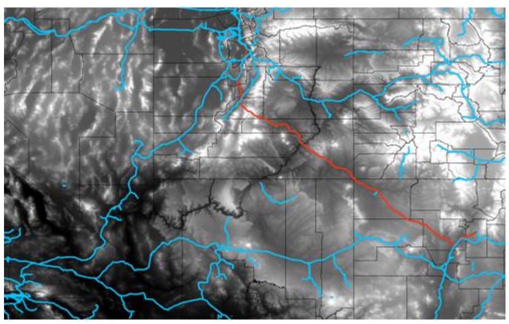
(a) Amenity Index (2000)



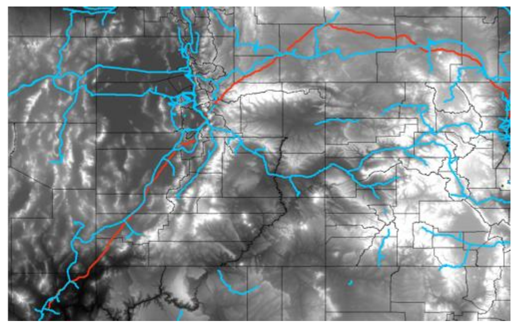
(b) Population Density (2000)

Note: Panel (a) gives predicted populated levels for each county, which serves as the proxy for exogenous environmental amenity. Panel (b) gives actual population density for the year 2000. For both figures, green indicates a large population density, yellow indicates medium population density, and red indicates low population density.

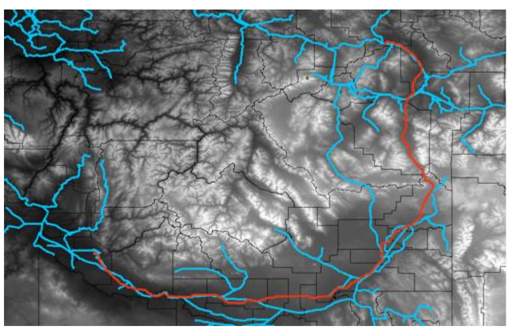
Figure A.5: Predicted v. Actual Rail Network in Four Case Studies



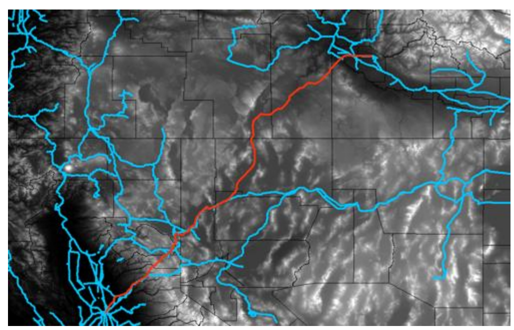
(a) Santa Fe to Salt Lake



(b) Denver to Las Vegas



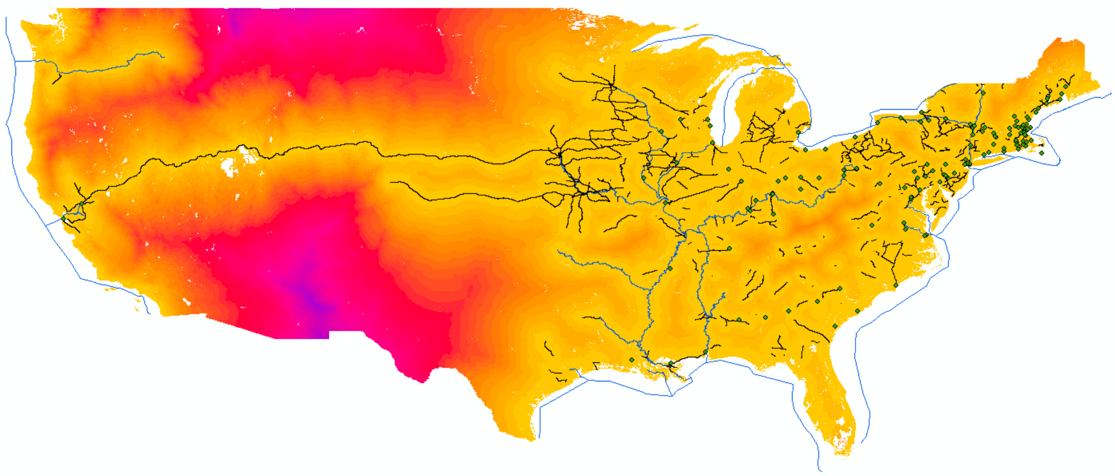
(c) Boise to Helena



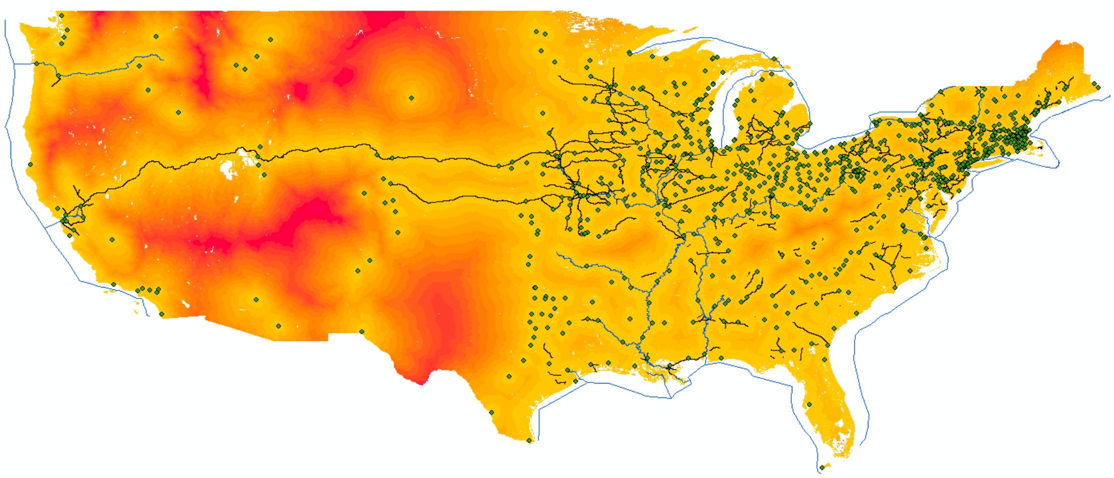
(d) Boise to Sacramento

Note: The current actual rail network is depicted in blue. The predicted paths based on cell-level measures of terrain ruggedness and water coverage are given in red. These are overlaid on a topographical map. These cases show a high-fidelity to the actual rail network from our method in panels (b) and (c), but our method also rightly ignores the endogenous city development and predicts more direct routes in panels (a) and (c). See Section 4.2 for details.

Figure A.6: Estimated Transport Cost



(a) Transport Cost by Cell (1850 cities)



(b) Transport Cost by Cell (1900 cities)

Note: Estimated transportation cost from each 1km^2 cell to the nearest national or international market location. Warmer colors indicate higher costs to reach markets. Markets are defined by either a pre-existing city of 5000 or more population, a railroad built by 1870, or a navigable waterway. Cities which reached 5000 people by 1850 or 1900 are denoted by points. Railroads in 1870 and marine highways (major, navigable waterways) are given by black and blue lines, respectively.

Table A.1: Summary Statistics for Oil-Rich and Poor Counties

		Population		Median Income	Manufacturing	
	N	1890	2000	2000	1890	2000
All Counties	3105	19.98	88.24	35.19	2.94	13.35
Had Oil Discovery						
Yes	204	8.71	140.71	32.66	1.08	8.10
No	2849	20.90	84.89	35.37	3.04	13.73
Oil-Poor with Amenity:						
High	562	30.21	122.29	35.72	2.74	16.29
Medium	1685	22.98	87.24	35.67	3.15	14.55
Low	602	6.40	43.37	34.22	2.98	8.10
Oil-Poor with Transportation Cost:						
High	540	4.48	37.75	31.00	1.06	7.64
Medium	1719	18.49	70.90	35.34	2.78	15.06
Low	590	42.98	168.77	39.45	4.95	14.42
Oil-Rich with Amenity:						
High	14	16.37	459.18	34.81	1.46	11.17
Medium	117	12.29	163.84	33.02	1.37	9.69
Low	73	1.51	42.57	31.66	0.35	4.44
Oil-Rich with Transportation Cost:						
High	73	1.98	42.29	31.62	0.38	5.09
Medium	113	12.37	207.71	32.85	1.37	9.66
Low	18	13.01	119.27	35.67	1.66	8.60

Summary statistics for outcome variables in 2000 and 1890 (income data are not available in 1890) are presented for all counties, counties with and without oil discoveries, and counties with and without oil discoveries by amenity and transportation cost favorability. N is the number of counties that fall into a particular category. First Discovery is the average first year that an oil or mineral discovery is made for counties in a group. Population is the average population in 1000s of persons. Median income is measured in 1,000s of USD. Manufacturing is the percentage share of employment in manufacturing.

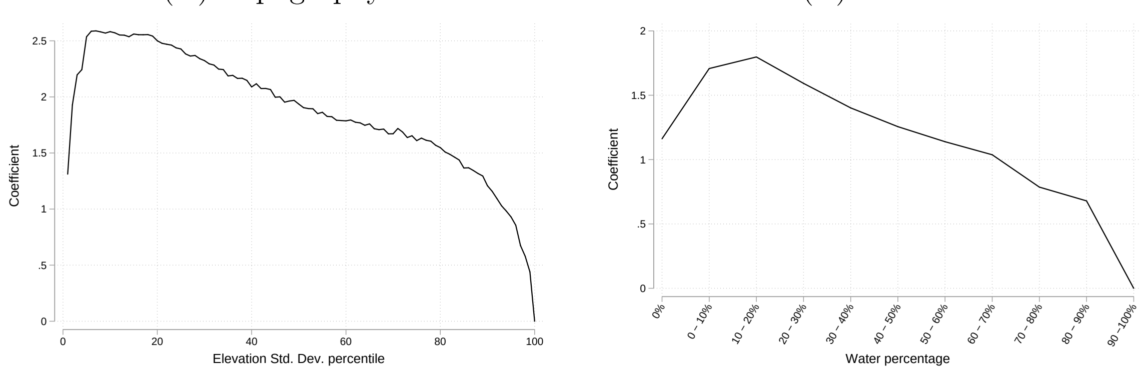
Table A.2: Summary Statistics for Cells Near Mineral Discoveries

	N (1850)	Population (1850)	Population (1950)
Had Mineral Discovery			
Yes	291,471	.024	.339
No	2,388,223	.009	.332
Mineral-Poor with Amenity:			
High	484,715	.020	.531
Medium	1,419,191	.006	.316
Low	484,317	.011	.178
Mineral-Poor with Transportation Cost:			
High	435,772	.005	.282
Medium	1,439,652	.008	.294
Low	512,799	.017	.483
Mineral-Rich with Amenity:			
High	55,298	.079	.438
Medium	171,661	.016	.357
Low	64,512	0	.206
Mineral-Rich with Transportation Cost:			
High	81,943	0	.337
Medium	177,498	.014	.310
Low	32,030	.146	.510

Note: N is the number of cells that fall into a particular category. Population is the percent of cells within a category with a population greater than zero. Mineral rich cells are those within thirty miles of a mineral discovery.

Figure A.8: Effects of Terrain on Rail Probability

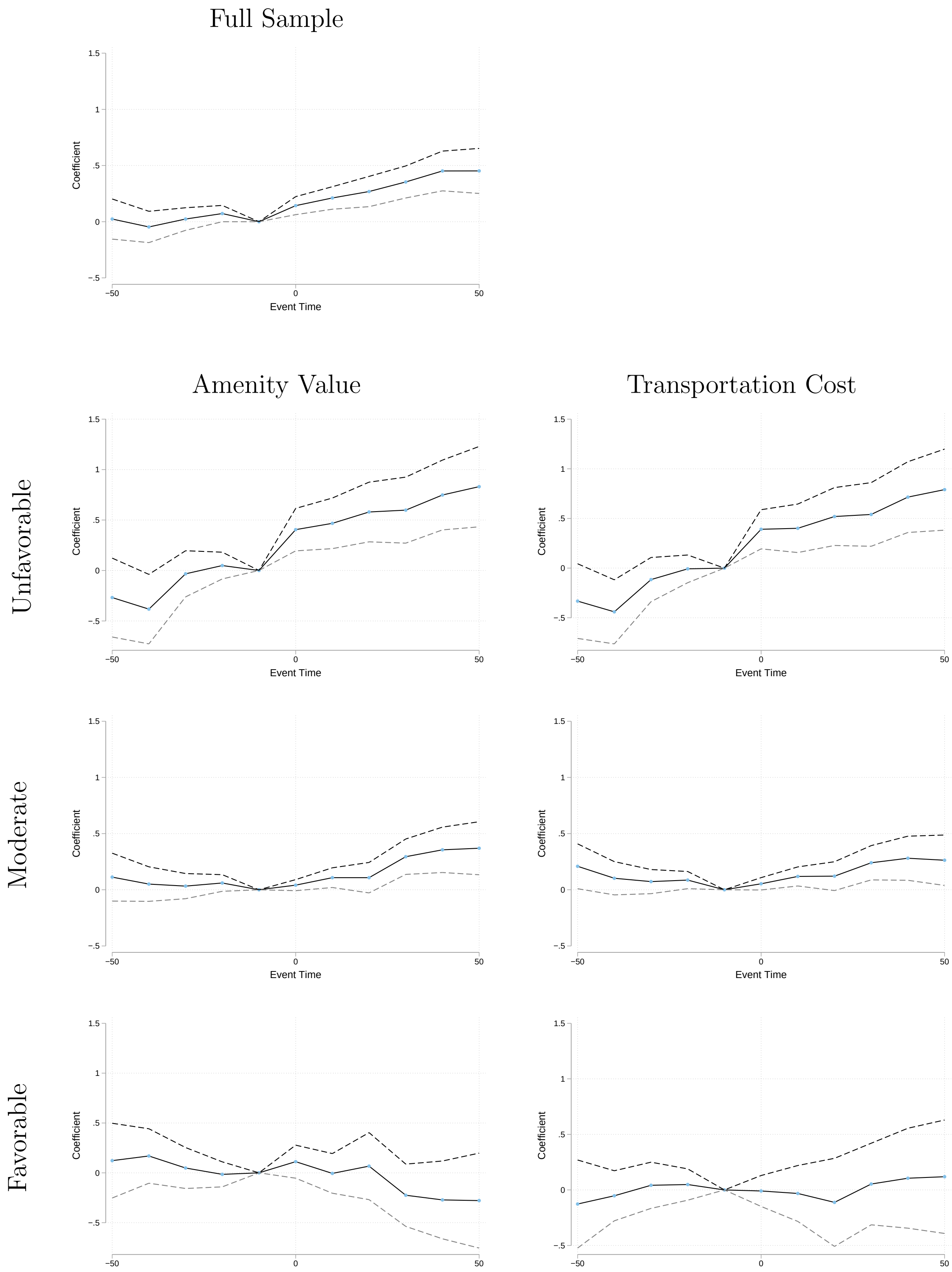
(A) Topography (B) Water



(A) Rail Probability Effect by Elevation SD Percentile. (B) Rail Probability Effect by Water Percentage.

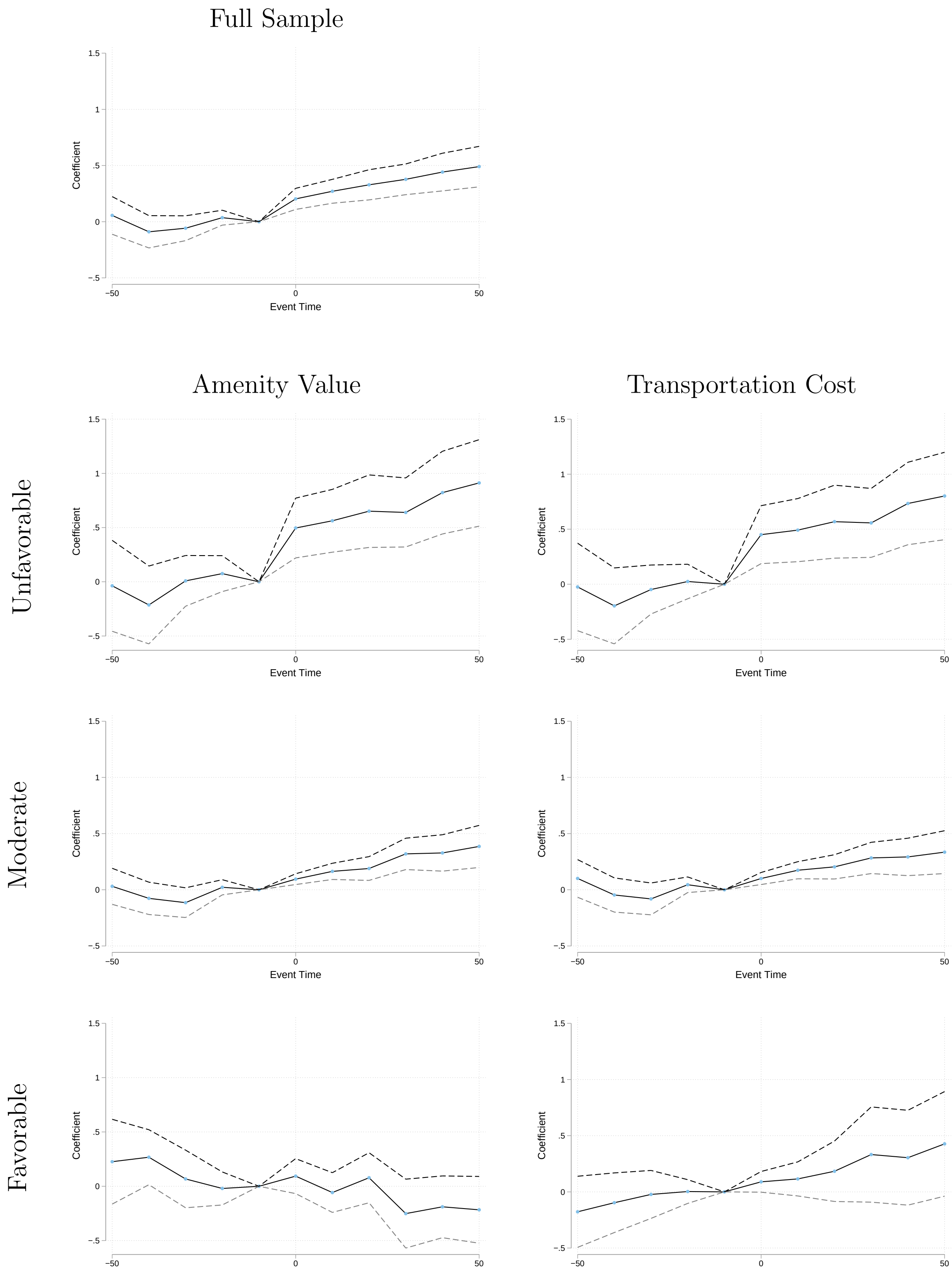
B Robustness

Figure B.1: Oil Discovery Results, Bordering Counties Excluded



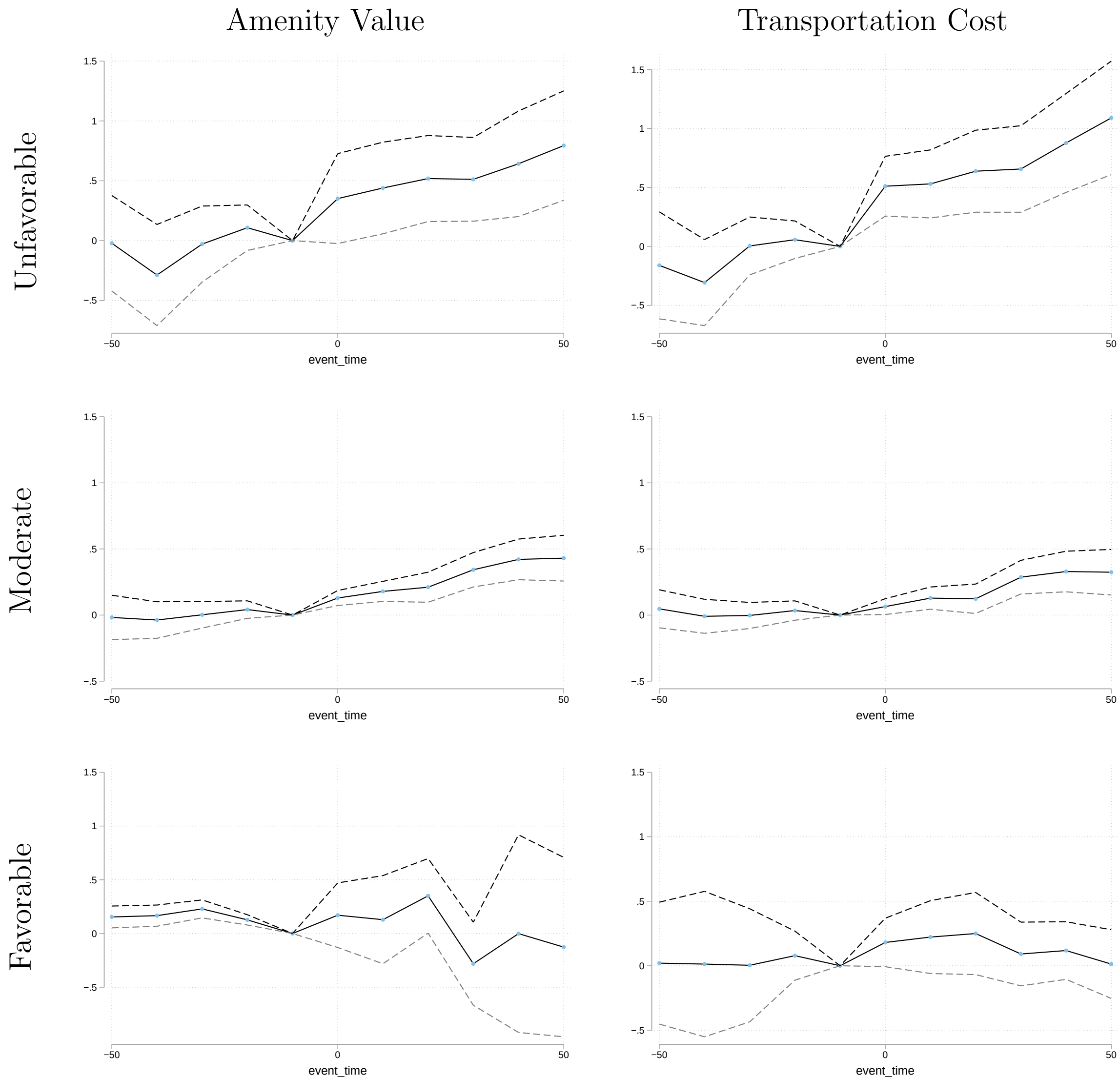
The graphs plot the average effects of oil discovery on log population density over time estimated from Equation 5, with control counties bordering treatment counties excluded from the sample. 95% confidence intervals are included.

Figure B.2: Oil Discovery Results, Year FEs



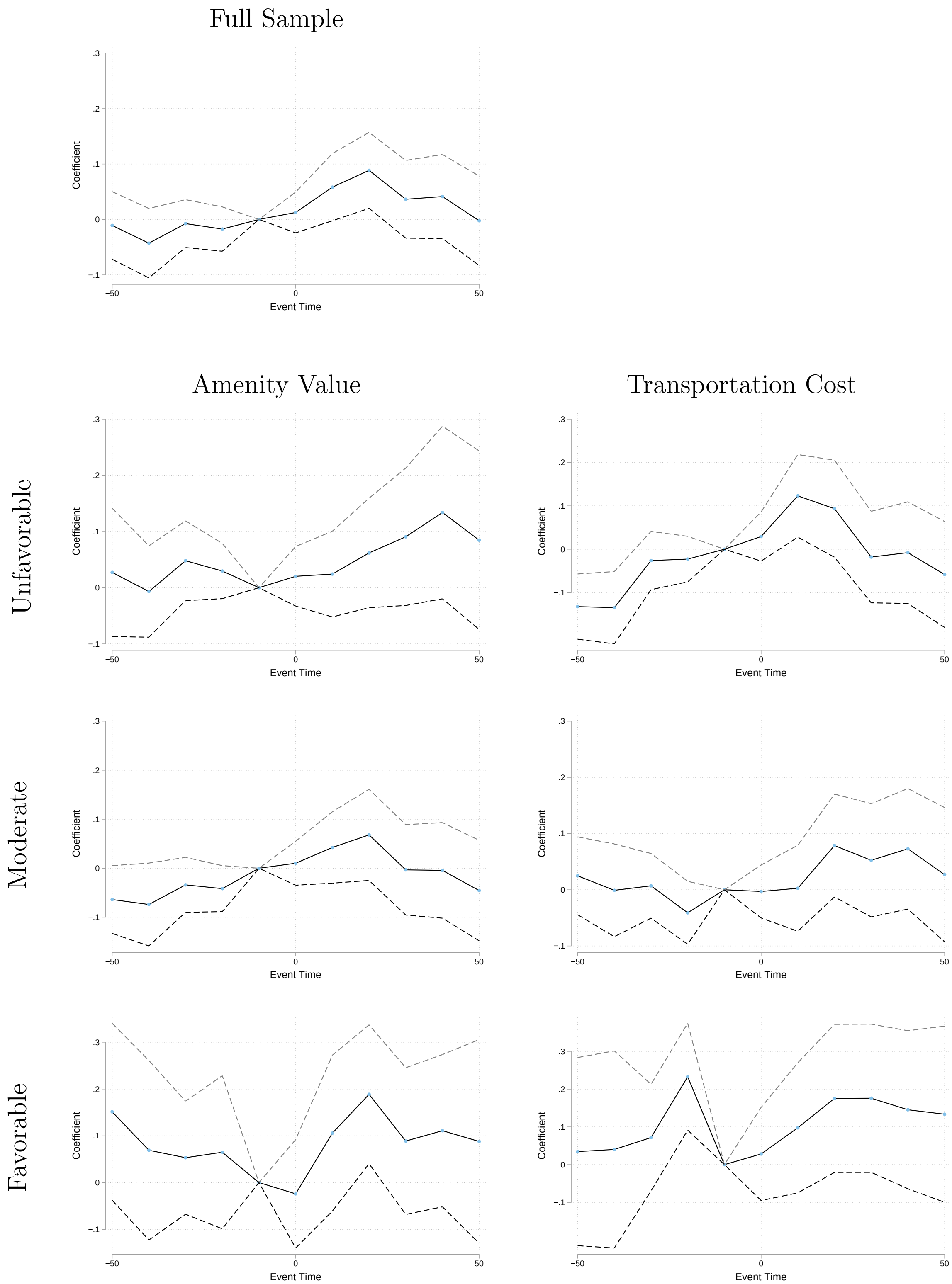
The graphs plot the average effects of oil discovery on log population density over time estimated from Equation 5, with year fixed effects included instead of state-by-year fixed effects. 95% confidence intervals are included.

Figure B.3: Oil Discovery Results, Alternate Bins



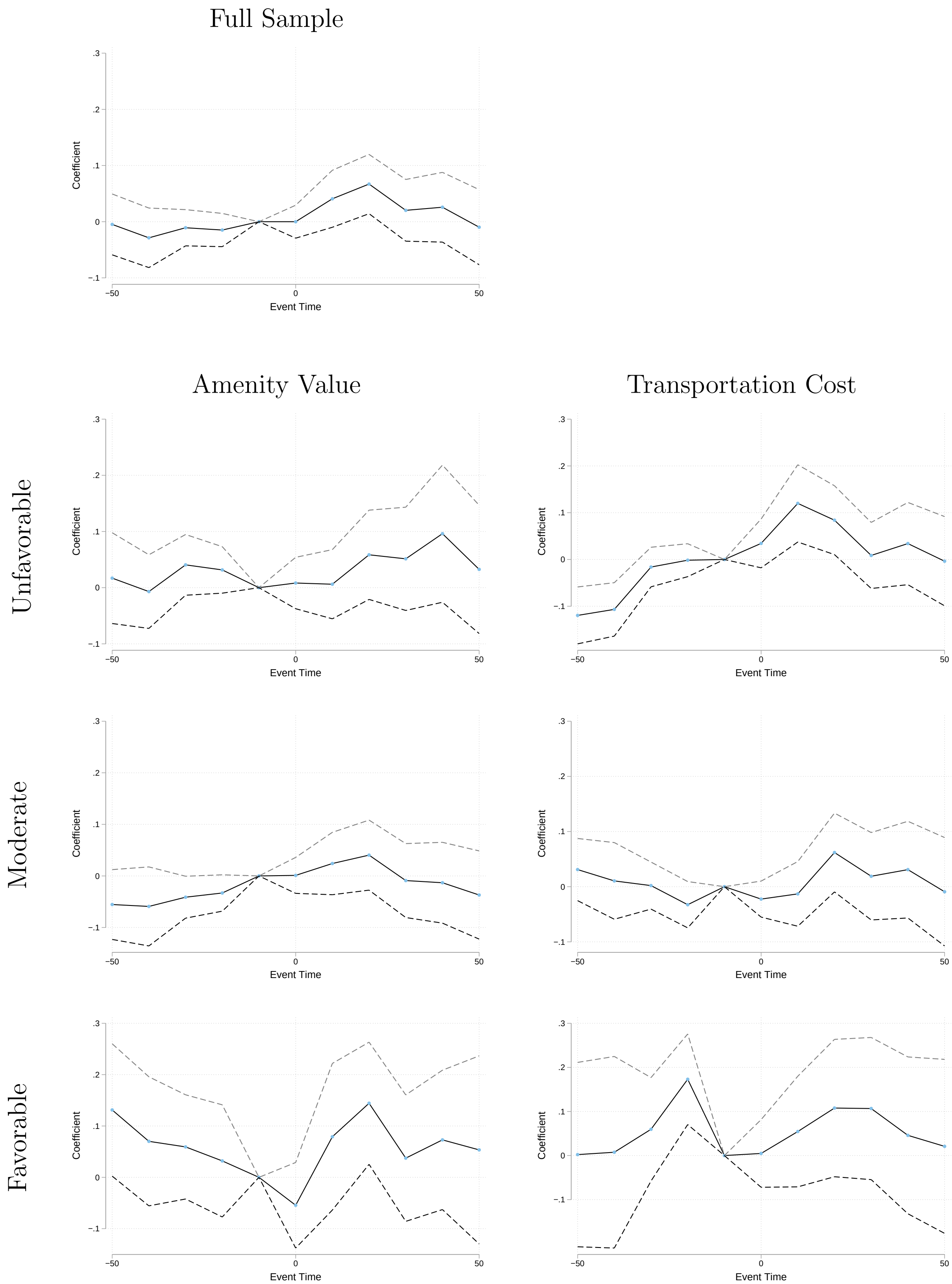
The graphs plot the average effects of oil discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. The data is partitioned based on the 10th and 90th percentile cutoffs into high, medium or low amenity and transportation groups, rather than 20th and 80th percentile cutoffs.

Figure B.4: Mine Discovery Results, Using a 5-mile Treatment Threshold



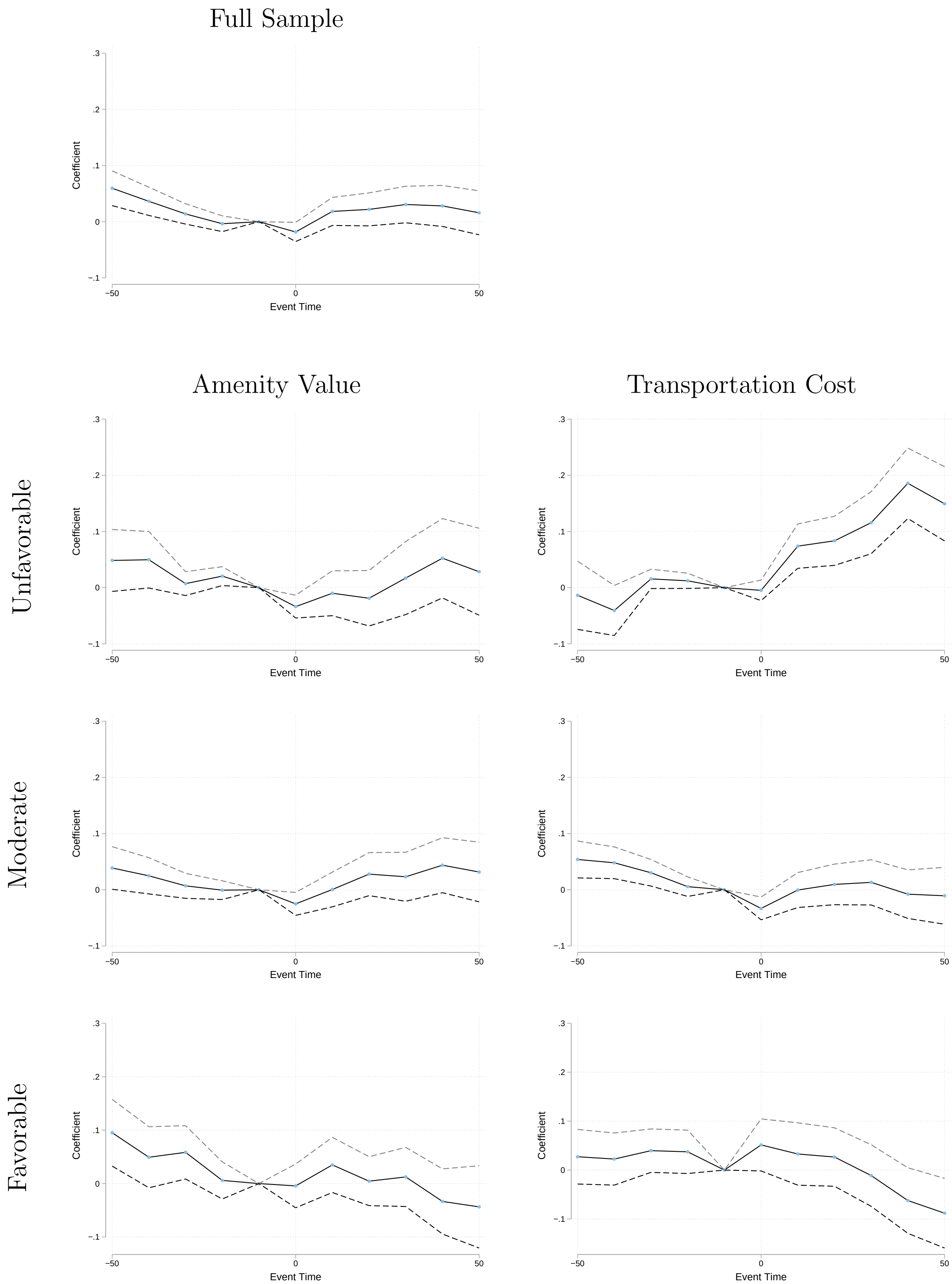
The graphs plot the average effects of mineral discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. The data is partitioned based on the 20th and 80th percentile cutoffs into high, medium or low amenity and transportation groups. Treatment cells are those within 5 miles of a discovery rather than 30 miles as in the baseline specification.

Figure B.5: Mine Discovery Results, Using a 10-mile Treatment Threshold



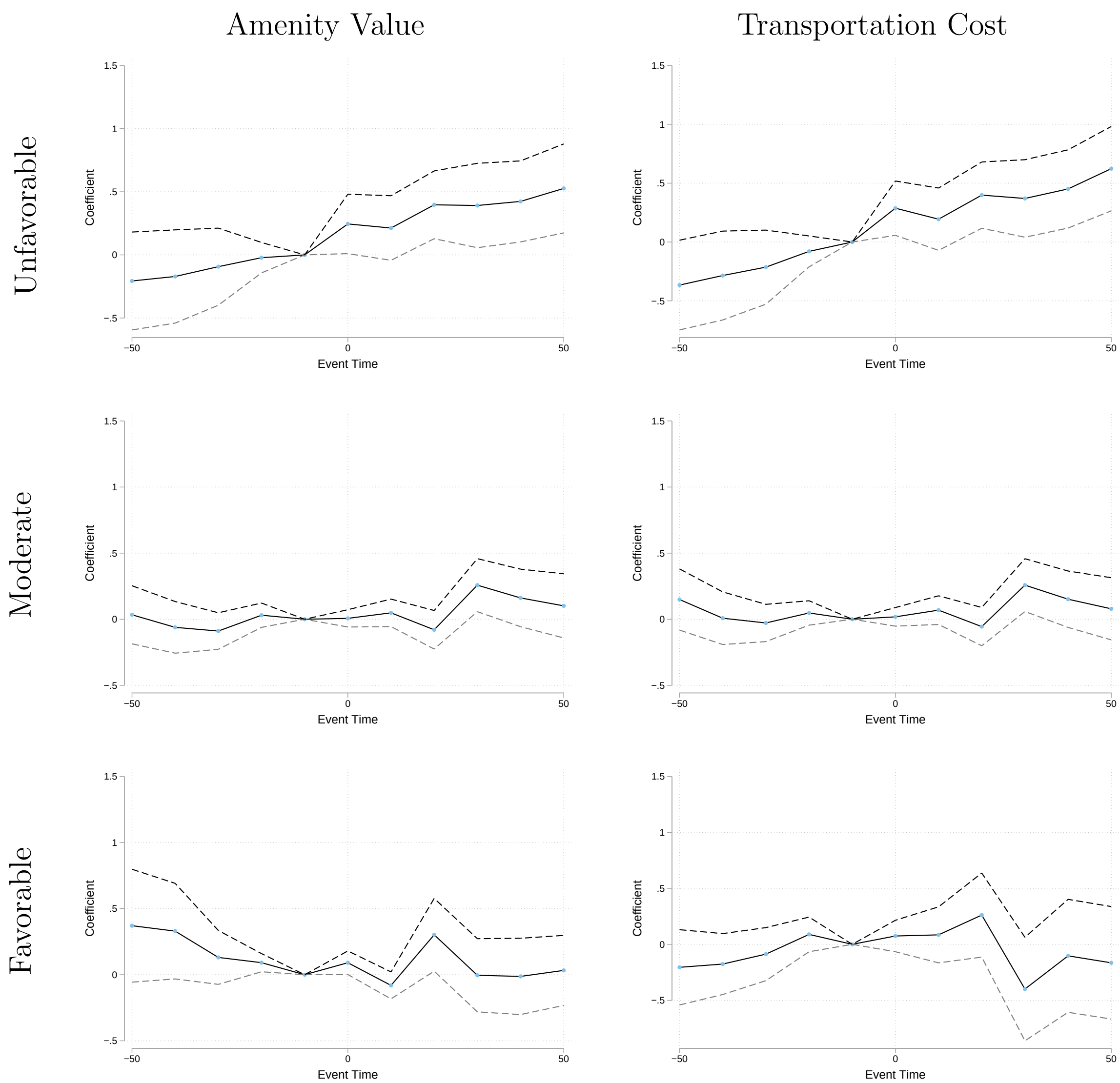
The graphs plot the average effects of mineral discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. The data is partitioned based on the 20th and 80th percentile cutoffs into high, medium or low amenity and transportation groups. Treatment cells are those within 10 miles of a discovery rather than 30 miles as in the baseline specification.

Figure B.6: Mine Discovery Results, Using a 50-mile Treatment Threshold



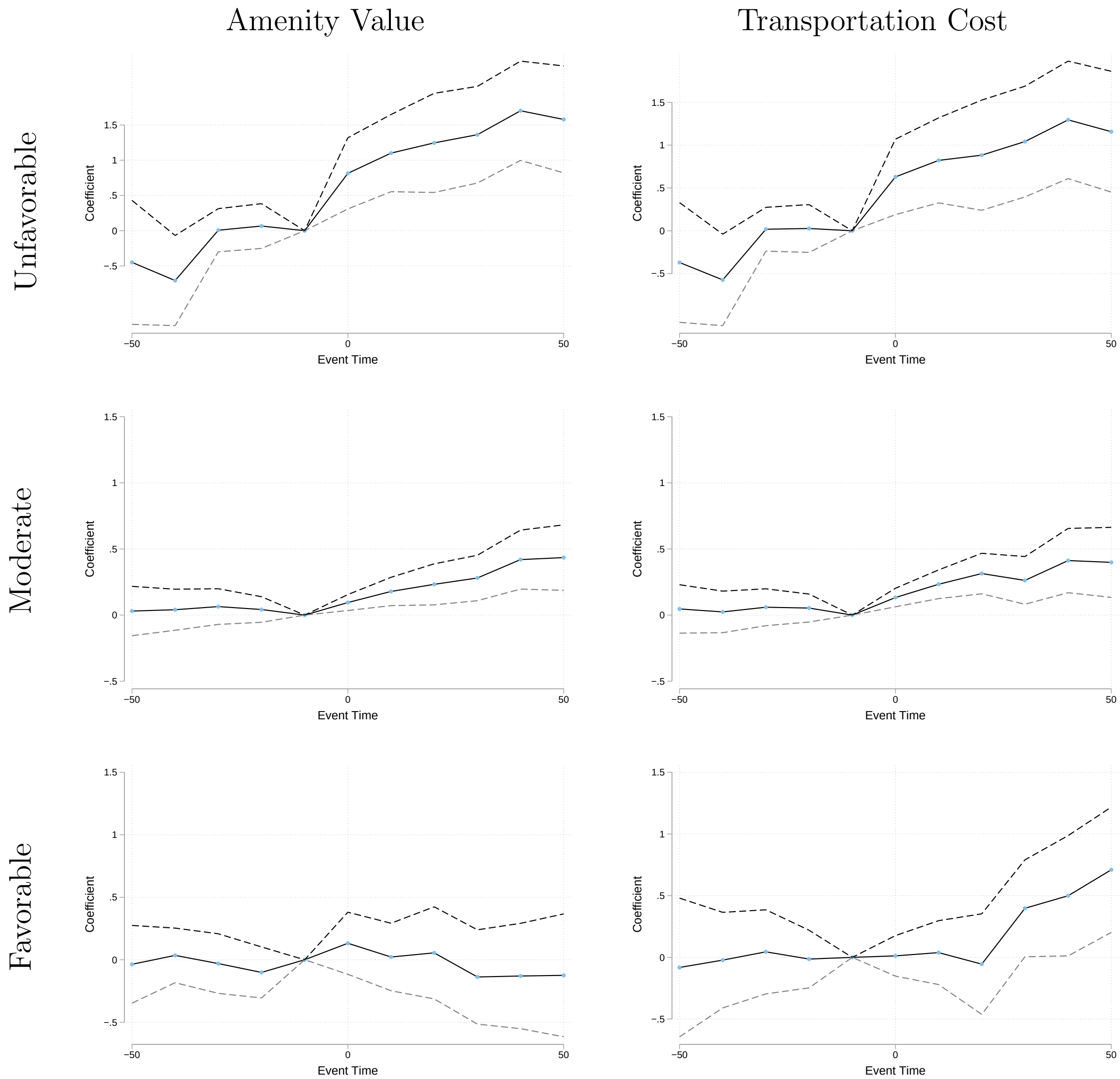
The graphs plot the average effects of mineral discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. The data is partitioned based on the 20th and 80th percentile cutoffs into high, medium or low amenity and transportation groups. Treatment cells are those within 50 miles of a discovery rather than 30 miles as in the baseline specification.

Figure B.7: Oil Discovery Results, Low Reserves Treatments



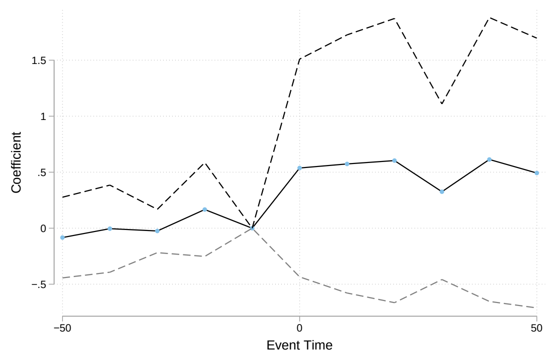
The graphs plot the average effects of oil discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. Only below median oil deposits are used in the estimation. The data is partitioned based on the 20th and 80th percentile cutoffs into high, medium or low amenity and transportation groups.

Figure B.8: Oil Discovery Results, High Reserves Treatments



The graphs plot the average effects of oil discovery on log population density over time estimated from Equation 5. 95% confidence intervals are included. Only above median oil deposits are used in the estimation. The data is partitioned based on the 20th and 80th percentile cutoffs into high, medium or low amenity and transportation groups.

Figure B.9: Event Study with 8 Boom-Ending Treatments Only



Event study of oil discovery on population using a subsample of 8 counties which did not see employment booms during the 1970's oil-boom period. For the time frame where employment data are available by sector (1960-2000), these counties saw precipitous declines in oil-sector employment, indicating their local boom was transient.

C Estimation of Transport Cost

C.1 Monte Carlo Simulation

This section develops a simulation to demonstrate key features of the proposed transportation cost model. We show sensitivity of the method to: a) the relative values $c^{terrain}$ to those for distance cost c^{km} and unobserved cost u, b) the number of nodes used to descretize the United States, i.e. the resolution of the fishnet, c) the assumed probability to transportation cost transformation $\hat{c}_{ij} = 1/\hat{x}_{ij}$, d) spatial correlation of $terrain$ and u. We show our methodology outperforms two alternative measures of recovering the intrinsic portion of transportation cost, namely the crow-flies distance and the actual rail network (which can be based on non-topographic features).

C.1.1 Simulation Methodology

The first step of each iteration is generate a grid network of size n x n, where n is a number of arcs that make up the grid's width. We evaluate three cases of increasing grid size n={10, 25, 100} to explore the consequences of finer versus coarser resolutions. Each arc x_{ij} of the network is populated with two randomly generated values, a value of terrain $terrain_{ij}$ and an unobserved determinant of transportation cost u_{ij}. This unobservable component might represent, for example, interspersed existing cities between the origin and destination that affect the placement (cost) of connecting rail lines. Because in practice, values of $terrain_{ij}$ and u_{ij} are likely autocorrelated across space, in alternative runs, we generate $terrain_{ij}$ and u_{ij} with some degree of spatial autocorrelation. Given the variables $terrain_{ij}$ and u_{ij}, we calculate the cost of using each arc in the network using Eq. 9:

$$Cost_{ij} = a + b * terrain_{ij} + u_{ij} \tag{9}$$

parameters a and b are defined as c^{km} and $c^{terrain}$ before. For simplicity, we hold fixed $a = 1$ across all cases of our simulations, instead varying b at levels 1e-5, 1, 1e+5. This allows us to understand the distance versus terrain tradeoff, by varying their relative cost. For example, does the estimate perform similarly when it is relatively less expensive to travel over mountains (ie when b is small) than it is to travel around them?

With the network populated with values of $cost_{ij}$, we solve the shortest path problem, choosing the origin and destination points as the opposite corners of the square grid. The optimal path over this simulated network represents the "actual" rail network that we observe contemporaneously. This path is a function of distance, observed terrain costs, and unobservable costs. As our goal is to estimate what transportation cost would be if unobservable costs were zero, we solve for the shortest path again with $u_{ij} = 0$. This second route represents the "true" route without interference from factors such as interspersed cities. We also solve for a third route setting $Cost_{ij} = 1$ for all arcs. The optimal route over this network, which ignores the actual cost variation of traveling a given arc, represents the crow-flies distance route. This route serves as a benchmark for our method relative to an alternative approximation that is used frequently in the literature. The relationship between the existence of actual rail on an arc and observed terrain values for that arc can be approximated using our reduced form logistic regression specification. Using the fitted model, we calculate the predicted values for the probability that rail exists on all arcs on the network given a value of observable terrain, $\Pr = \Pr(x_{ij} = 1 | terrain_{ij})$. We transform these predicted probabilities to our cost approximation using the form: $\hat{c}_{ij} = 1/\mathrm{Pr}_{ij}$. We compare this approximation to a linear transformation, $\hat{c}_{ij} = 1 - \mathrm{Pr}_{ij}$, as a check for robustness. The shortest path problem is re-solved for cost values $\hat{c}$. The solution path gives us the least cost route from the origin to destination accounting only for the observed portion of transportation cost, terrain.

We use several summary measures which are calculated for each iteration. For each route (actual, naive, true, and our model), we calculate the total cost of the entire route ($\sum c_{ij} * x_{ij}$). We also calculate our approximated total cost for the route, $\sum \hat{c}_{ij} * x_{ij}$. We also calculate the percentage of arcs from the true route that overlap with the arcs used in the actual path, the naive path, and our approximated route.

We simulate 1,000 networks for each combination of parameter values (b, n), presense of autocorrelation, as well as the transformation functions used to translate probability of rail to approximated cost.

C.1.2 Results

In Figure B.10, we show the relationship between the true total cost (i.e. $\sum b * terrain * x$) and the approximated total cost derived from rail probability (i.e. $\sum \hat{c} * x$) is monotonic across the various parameter values. In fact, for some parameter values of b and n, this is linear or near linear.

Figure B.10: Relationship between total approximated cost and total observed cost

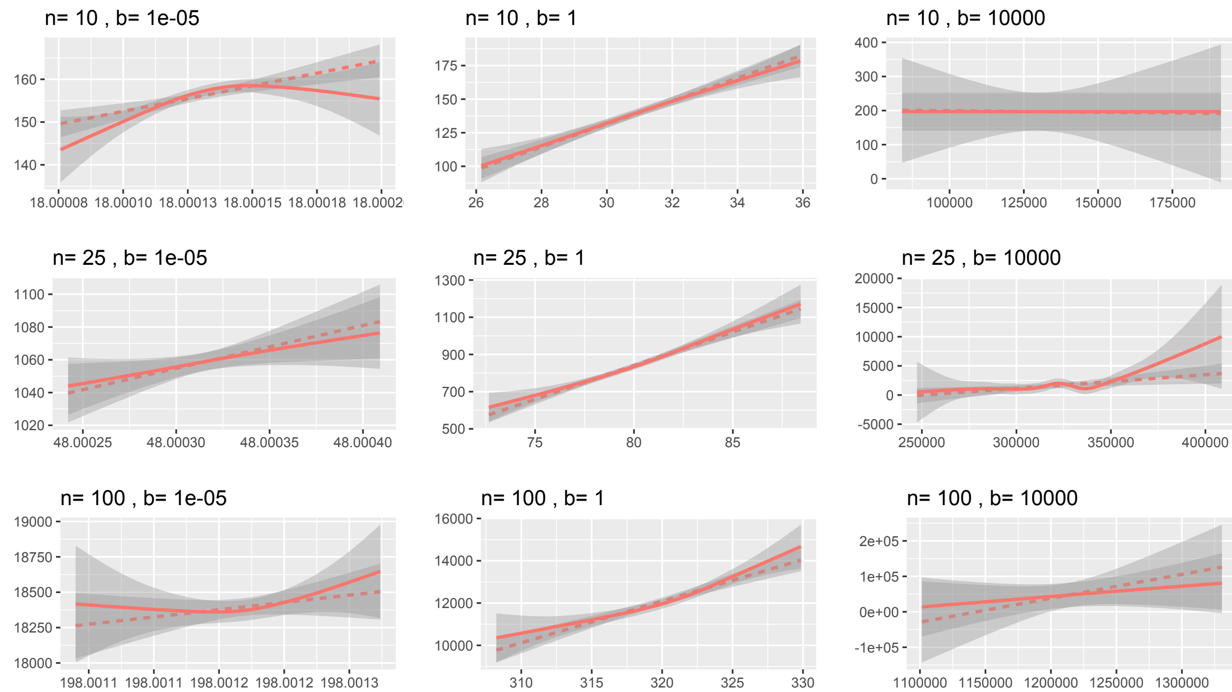
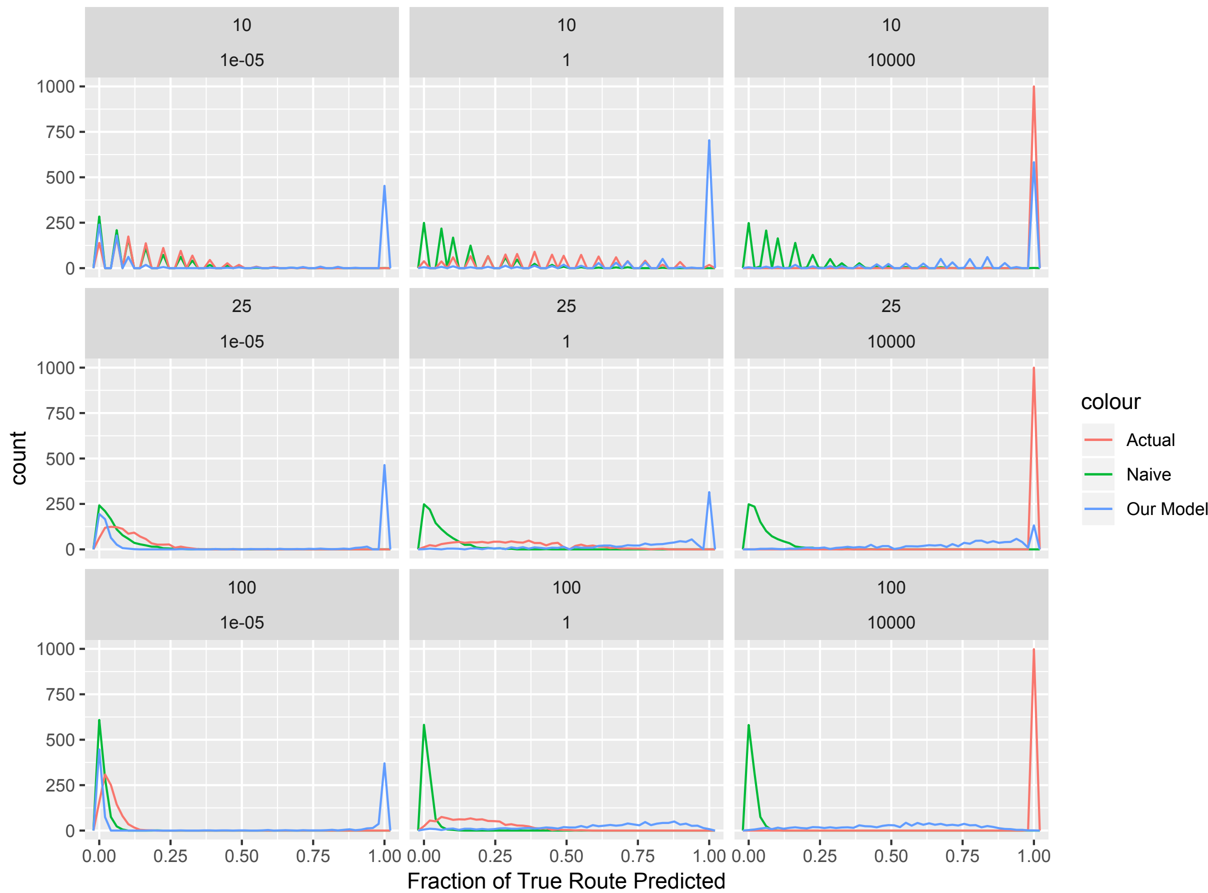


Figure B.11 shows one performance measure of our model of transportation costs compared to both the actual network and a naive (crow-flies distance) route. The measure is the fraction of the true route (based on terrain alone) shared by the given alternative route. Intuitively, as the cost of terrain relative to the cost of the unobservable component increases from the left to right panels of the figure, the actual route and the true route converge. When the cost of terrain relative to the unobservable factors is small or modest (b=1e-5 or b=1), our model predicts the true route at a much higher rate than the naive or actual route.

Figure B.11: Histogram of fraction of true route intersected by actual, naive, and predicted routes



While the results from Figure B.11 show adequate performance of our method for predicting the true route, our goal is not to identify the true route, but rather the costs associated with the true least-cost route. In Figure B.12 we plot the distribution of total transportation costs for the various parameter values and routes. Consistent with the pattern in Figure B.11, the actual and true route costs converge as the importance of terrain increases relative to the unobserved costs. Also consistent with Figure B.11 the naive model predicts the true transportation cost poorly. However, our model performs well across all values of terrain costs, providing a close approximation to the total cost of the true route.

Figure B.12: Histogram of total cost of true, actual, naive, and predicted routes