Trade and Geography*

Stephen J. Redding†
Princeton University, NBER and CEPR

April 17, 2021

Abstract

This paper reviews recent research on geography and trade. One of the key empirical findings over the last decade has been the role of geography in shaping the distributional consequences of trade. One of the major theoretical advances has been the development of quantitative spatial models that incorporate both exogenous *first-nature* geography (natural endowments) and endogenous *second-nature* geography (the location choices of economic agents relative to one another) as determinants of the distribution of economic activity across space. These models are sufficiently rich to capture first-order features of the data, such as gravity equations for flows of goods and people. Yet they remain sufficiently tractable as to permit an analytical characterization of the properties of the general equilibrium and facilitate counterfactuals for realistic policy interventions. We distinguish between models of regions or systems of cities (where goods trade and migration take center stage) and models of the internal structure of cities (where commuting becomes relevant). We review some of key empirical predictions of both sets of theories and show that they have been remarkably successful in rationalizing the empirical findings from reduced-form research. Looking ahead, the combination of recent theoretical advances and novel geo-coded data on economic interactions at a fine spatial scale promises many interesting avenues for further research, including discriminating between alternative mechanisms for agglomeration, understanding the implications of new technologies for the organization of work, and assessing the causes, consequences and potential policy implications of spatial sorting.

J.E.L. CLASSIFICATION: F1, J4, R1, R4

KEYWORDS: trade, geography, local labor markets

*I am especially grateful to Elhanan Helpman for extremely helpful comments. I would also like to thank the other editors, my discussants Cecile Gaubert and Tony Venables, and the other participants at the Handbook of International Economics virtual conference in March 2021, for their terrific comments and suggestions. Thanks also to Costas Arkolakis, Jonathan Dingel, Dave Donaldson, Gordon Hanson, Nina Pavcnik and Matt Turner for extremely helpful comments. I would also like to thank Benny Kleinman for excellent research assistance. Responsibility for results, opinions and errors lies with the author alone.

†Dept. Economics and School of Public and International Affairs, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Tel: 1 609 258 4016. Email: reddings@princeton.edu.

1 Introduction

One of the most striking features of the economy is the extremely uneven distribution of production, trade and income across geographic space. This concentration of economic activity is most evident in the existence of cities. Roughly two thirds of the world's population is projected to live in cities by the year 2050, with the urban population increasing by around 2.5 billion people, and nearly 90 percent of this increase concentrated in Asia and Africa.[1] But this concentration is also evident for individual economic activities within cities. In 1940, Motor Vehicles and Motor Vehicle Equipment Industries made up 55 percent of the city of Detroit's manufacturing employment and 24 percent of its total employment, and the city of Detroit was responsible for 30 percent of total employment in this industry in the United States.[2] A natural first question is what explains these concentrations of economic activity: to what extent are they explained by uneven differences in exogenous fundamentals versus endogenous forces for the agglomeration of economic activity? A natural second question is what are the implications of these concentrations of economic activity for the way in which economies respond to external shocks? As technology and trade shocks occur over time, and some locations gain a comparative advantage in new sectors, while others lose that comparative advantage, how does the spatial economy adjust? How do people and economic activities sort across geographic space in response? What are the implications of new communication technologies, and shocks such as the COVID-19 pandemic, for the future of work in densely versus sparsely-populated locations? If some locations are "left-behind" by increased economic integration, to what extent and how should public policy respond?[3]

This chapter reviews recent theoretical and empirical research on the causes and consequences of the uneven distribution of economic activity across space. We begin by reviewing a largely reduced-form empirical literature that has shown that geography is a key dimension along which the distributional consequences of international trade occur. This idea that international trade creates winners and losers is not new and is rather central to neoclassical economics. In the Heckscher-Ohlin model, trade affects income inequality between different factors of production that are mobile across industries. In the specific-factors model, these effects of trade on income inequality occur because factors of production are immobile between industries, at least in the short-run. Nevertheless, in both of these neoclassical theories, these distributional consequences of trade occur at the national level.[4] The key contribution of recent empirical research has been to highlight how differences in industry composition across local labor markets within countries induce differences in their exposure to international trade shocks. In particular, Autor, Dorn, and Hanson (2013a) show that local labor markets more exposed to Chinese import growth experience larger reductions in manufacturing employment as a share of the working-age population, the employment to population ratio and mean log weekly earnings, as well as larger increases in per capita unemployment, disability, and income assistance transfer benefits. These findings have stimulated a useful dialogue between theory and empirics, as researchers have developed quantitative spatial models to rationalize these empirical findings.

We next turn to the role of geography in explaining the observed uneven distribution of economic activity across

[1] These figures are taken from the World Urbanization Prospects, United Nations (2018). For further evidence on urbanization in a historical context, see Michaels, Rauch, and Redding (2012) and Desmet and Henderson (2015).

[2] Figures based on the 1940 individual-level population census data (Ruggles, Flood, Goeken, Grover, Meyer, Pacas, and Sobek 2018).

[3] For further discussion of these "left-behind" places and the scope for place-based policies, see Kline and Moretti (2014b), Austin, Glaeser, and Summers (2018), Wuthnow (2019) and Gaubert, Kline, and Yagan (2020).

[4] Another mechanism for trade to affect wage inequality at a more disaggregate level is reallocations of resources across heterogeneous firm that pay different wages, as for example in Helpman, Itskhoki, and Redding (2010) and Helpman, Itskhoki, Muendler, and Redding (2017). For a review of the broader literature on trade and income inequality, see Helpman (2018).

space. We highlight a key distinction between what is termed "first-nature" and "second-nature" geography. First-nature geography is concerned with locational fundamentals, including the physical geography of coasts, mountains and natural endowments. In contrast, second-nature geography is concerned with the location of agents relative to one another in geographic space, and the role that this plays in understanding the agglomeration of economic activity. While first-nature geography is largely exogenous, second-nature geography is typically endogenous, and could be influenced, at least in principle, by public policy interventions.

Separating first and second-nature geography raises both theoretical and empirical challenges. From a theoretical point of view, a key challenge has been to incorporate asymmetries across locations (first-nature geography) into theoretical models of economic geography while preserving their analytical tractability. Traditionally, the theoretical literature on economic geography has abstracted from first-nature geography by considering stylized settings with a small number of symmetric regions.[5] However, a major theoretical advance in recent years has been the development of quantitative spatial models that incorporate both first and second-nature geography. These models are sufficiently rich to capture first-order features of the data, such as large numbers of locations with heterogeneous productivity, amenities, and land area, as well as trade and commuting costs. Yet these models remain sufficiently tractable as to permit an analytical characterization of the general equilibrium and to be used for realistic counterfactuals. Relative to earlier computable general equilibrium (CGE) models, they are also relatively parsimoniously specified, with a small number of key equilibrium relationships and structural parameters to be estimated, thereby permitting transparent interpretation of results. They thus provide a platform for evaluating the impact of a host of public policy interventions, including specific transport infrastructure improvements, such as the construction of a new link in the U.S. Interstate Highway System or a high-speed railway between the North and South of California.[6]

From an empirical point of view, a central challenge has been separately estimating the contributions of first and second-nature geography. Suppose that we observe a group of contiguous locations with high levels of economic activity in the data. On the one hand, this concentration of economic activity could reflect good access to one another's markets (second-nature geography). On the other hand, it could be explained by common locational fundamentals such as access to nearby natural resources (first-nature geography). This challenge is an example of the broader problem in the social sciences of separately identifying spillovers from correlated individual effects.[7] Empirical research in recent years has seen more careful attention to the use of exogenous sources of variation, including natural experiments from history, as part of a broader "credibility revolution" in applied econometrics. A growing body of empirical research using such exogenous variation has provided evidence on the causal effects of transport infrastructure investments and changes in market access on the location of economic activity.

We distinguish between models of regions or systems of cities (where goods trade and migration take center stage) and models of the internal structure of cities (where commuting becomes relevant). We review the empirical evidence on the predictions of both sets of models. Some of the key empirical findings concern the importance of transport infrastructure for the spatial distribution of economic activity within countries; the role of market access (the proximity of firms and consumers to one another) for location choices; the extent to which economic activity is path dependent (such that temporary shocks can have permanent effects); the degree of spatial sorting of high and low-skilled work-

[5] For a synthesis of this theoretical literature, see Fujita, Krugman, and Venables (1999). For reviews of this earlier economic geography literature, see Overman, Redding, and Venables (2003), Redding (2010) and Redding (2011).

[6] For a review of this recent literature on quantitative spatial models, see Redding and Rossi-Hansberg (2017).

[7] See, in particular, the discussion in Manski (1993).

ers across geographic space; and the role of agglomeration forces in understanding the concentration of economic activities across geographic space. We show that the counterfactual predictions of both sets of models have proved to be remarkably successful in explaining empirical findings from reduced-form research, such as the impacts of the division of Germany on city growth, the effect of the division of Berlin on land prices, the distributional consequences of Chinese import growth across local labor markets, and the impact of transport infrastructure investments on the location of economic activity.

The remainder of this paper is structured as follows. Section 2 reviews recent reduced-form evidence on the distributional consequences of international trade across local labor markets. Section 3 introduces a class of quantitative spatial models for understanding the uneven distribution of economic activity across regions or cities within countries. Section 4 presents empirical evidence on some of the key predictions of this class of models. Section 5 introduces a class of quantitative urban models for understanding the internal structure of economic activity within cities. Section 6 offers some concluding comments.

2 The Geographic Incidence of International Trade Shocks

A major contribution of recent reduced-form empirical research in international trade has been to quantify the importance of geography as a dimension along which the distributional effects of international trade occur. Following the influential work of Autor, Dorn, and Hanson (2013a) for the United States, a substantial component of this research has focused on the large-scale shock provided by China's emergence into the global economy. In Section 2.1, we review the empirical findings from this extensive body of research.[8] In Section 2.2, we discuss the interpretation of these reduced-form empirical findings, before returning in a later section to examine the extent to which these reduced-form empirical findings are consistent with the predictions of quantitative spatial models.

2.1 The China Shock

With its rapid economic growth following its market-orientated reforms of 1978, China has emerged as a major source of import competition for producers of manufactured goods in developed countries such as the United States. In 1991, low-income countries accounted for just 9 percent of U.S. manufacturing imports. In contrast, by 2000, the low-income-country share of U.S. imports reached 15 percent and climbed to 28 percent by 2007, with China accounting for 89 percent of this growth. This rise in import penetration is particularly rapid following China's admission to the World Trade Organization (WTO) in 2001 and coincides with a marked decline in U.S. manufacturing.[9] Between 1987 and 2007, the rise in China's share of U.S. imports from less than 1 percent to close to 5 percent is accompanied by a decline in the share of the U.S. working-age population employed in manufacturing of around one third, from 12.6 percent to 8.4 percent.[10] However, although this relationship is suggestive, it should be interpreted with caution, because these trends in import penetration and manufacturing employment potentially could be explained by omitted third variables, including in particular changes in technology.

To provide further evidence on this relationship, Autor, Dorn, and Hanson (2013a) construct a measure of the exposure of local labor markets in the United States to the China shock, which is based on changes in aggregate

[8] For surveys that focus largely on these local labor market effects of international trade, see Autor, Dorn, and Hanson (2016) and Pavcnik (2017).

[9] Admission to the WTO in 2001 removed the uncertainty associated with annual votes in the U.S. Congress over whether to maintain Permanent Normal Trade Relations (PNTR) with China, as studied in Pierce and Schott (2016) and Handley and Limão (2017).

[10] For early evidence on the impact of the China shock at the plant-level, see Bernard, Jensen, and Schott (2006).

industry imports and the concentration of these industries in each local labor market. This measure is derived as a first-order approximation to a gravity model of international trade. In particular, the change in imports per worker (ΔIPW_{uit}) in each local labor market i at time t in the United States u is constructed as: $\Delta IPW_{uit} = \sum_j \frac{L_{ijt}}{L_{ujt}} \frac{\Delta M_{ucjt}}{L_{it}}$, where Δ in the difference operator; L_{it} is total employment in the local labor market i at the start of the period (year t); ΔM_{ucjt} is the change in U.S. imports from China in industry j between the start and end of the period; L_{ijt}/L_{ujt} is the local labor market i's share of U.S. employment in industry j at the start of the period (year t).

To address the concern that U.S. imports from China are endogenous to shocks to U.S. import demand, and to capture the supply-side shock from China's own liberalization and economic growth, Autor, Dorn, and Hanson (2013a) construct an instrument using the growth of Chinese imports to eight other developed countries.[11] In particular, U.S. import exposure is instrumented using an analogous import exposure measure that is constructed using the imports from China of these eight other developed countries and lagging the US employment levels ten years to address the concern that employment at the start of the period could be endogenous to the subsequent China shock.

To explore the causal impact of the China shock across U.S. local labor markets, Autor, Dorn, and Hanson (2013a) consider the following second-stage regression, in which the main outcome of interest is the change in the manufacturing employment share of the working-age population (ΔL_{it}^m):

$$\Delta L_{it}^m = \gamma_t + \beta_1 \Delta IPW_{uit} + \boldsymbol{X}_{it}' \beta_2 + e_{it}, \tag{1}$$

where γ_t is a time fixed effect; ΔIPW_{uit} is the U.S. import exposure measure discussed above, which is instrumented in a first-stage with the import exposure measure constructed using the imports of other developed countries; $\boldsymbol{X}_{it}'$ is a matrix of controls; and e_{it} is the regression error.

This regression (1) is a long-differences specification in which a fixed effect for time-invariant unobserved heterogeneity in the level of the manufacturing employment share has been differenced out. The key coefficient of interest on exposure to the China shock (β_1) has a "difference-in-differences" interpretation, in which the first difference is over time, and the second difference is across local labor markets with different levels of exposure to the China shock. In their baseline analysis, Autor, Dorn, and Hanson (2013a) consider two long differences from 1990-2000 and 2000-7, and a specification pooling both of these long differences from 1990-2007. In Placebo checks, they also consider two long differences pre-dating China's emergence as a major trading nation from 1970-80 and 1980-90, and a specification pooling both of these Placebo long differences from 1970-1990.

The instrumental variables estimation of equation (1) requires that the instrument is relevant (a good predictor of the endogenous variable) and valid (only affects the outcome variable of interest through the endogenous variable). Perhaps unsurprisingly, exposure measured using the imports of the eight other developed countries is a powerful predictor of U.S. import exposure, with first-stage F-statistics well above the conventional threshold of ten. Therefore, the more demanding of the two identifying assumptions is that import exposure constructed using the imports of the eight other developed countries (ΔIPW_{oit}) is uncorrelated with shocks to the manufacturing employment share of the working-age population (ΔL_{it}^m), as captured in the regression error (e_{it}).

To address concerns about omitted variables that could be correlated with both import exposure and changes in the manufacturing employment shares of the working-age population, thereby violating this exclusion restriction, Autor,

[11] The eight other high-income countries are those that have comparable trade data covering the full sample period: Australia, Denmark, Finland, Germany, Japan, New Zealand, Spain, and Switzerland.

Dorn, and Hanson (2013a) include a wide range of controls ($\boldsymbol{X}'_{it}$): (i) Percentage of employment in manufacturing at the start of the period; (ii) Percentage of college educated in the population at the start of the period; (iii) Percentage of foreign-born population at the start of the period; (iii) Percentage of employment among women at the start of the period; (iv) Percentage of employment in routine occupations at the start of the period; (v) a measure of the offshorability of occupational employment at the start of the period; and (v) Census-division-year fixed effects. The first control for the size of the manufacturing sector addresses the concern that the exposure measure just could be capturing the uneven distribution of manufacturing across local labor markets and its secular decline over time. The fourth control for the percentage of employment in routine occupations alleviates the concern that shocks to industry technology (such as computerization) could both directly affect the economic outcome of interest and be correlated with shocks to industry imports from China.[12] The fifth control for the census-division-year fixed effects mitigates the concern that the results could just capture a broader shift in the distribution of economic activity from the North-East and the Mid-West to the West and the South.

Even after conditioning on these controls, Autor, Dorn, and Hanson (2013a) find that local labor markets that were more exposed to the China shock experienced a larger decline in the manufacturing employment share of the working-age population. This relationship is both statistically significant and economically relevant. In the preferred specification, each $1,000 increase in import exposure per worker is predicted to reduce manufacturing employment as a share of the population by -0.596 percentage points. To put this into context, the observed difference in the ten-year equivalent growth in import exposure between the 75th and 25th percentiles from 2000-7 is 1.06. Therefore, these estimates imply that the share of manufacturing employees in the working-age population of a local labor market at the 75th percentile of import exposure declined by $-0.6466 = 1.06 \times -0.596$ percentage points more than in a local labor market at the 25th percentile over this period.

In addition to the manufacturing employment shares of the working-age population, exposure to the China shock is found to affect a wide range of other economic outcomes of interest. Comparing two local labor markets at the 75th and 25th percentiles of exposure to Chinese import growth from 2000-7, the more exposed local labor market experiences a 0.8 percentage point larger reduction in the employment to population rate, a 0.8 percent larger decline in mean log weekly earnings, and larger increases in per capita unemployment, disability, and income assistance transfer benefits on the order of 2 to 3.5 percent. These findings for transfers suggest that federally funded transfer programs, such as Social Security Disability Insurance (SSDI), play an important role in insuring U.S. workers against trade-related employment shocks. Notably, despite these negative effects on local labor market outcomes, Autor, Dorn, and Hanson (2013a) find little evidence of population movements in response to the China shock.

These findings have had a major impact on both academic research and the policy debate, being presented in the Oval Office during the administration of President Obama, and being featured in the publicity material for the 2016 election campaign of President Trump.[13] Subsequent research has explored the impact of the China shock in a number of different countries, as reviewed in Autor, Dorn, and Hanson (2016). Although research on the China shock initially concentrated on labor market outcomes, a proliferating literature has explored its broader effects on mortality (Pierce and Schott 2020), marriage outcomes (Autor, Dorn, and Hanson 2019), political polarization (Autor, Dorn, Hanson, and Majlesi 2020 and Che, Lu, Pierce, Schott, and Tao 2020), and innovation (Autor, Dorn, Hanson, Pisano, and Shu

[12] Autor, Dorn, and Hanson (2013b) show that there is relatively little correlation between the exposure of U.S. local labor markets to the trade shock from China and the technology shock from the automation of routine tasks, thereby permitting separate identification of these two forces.

[13] See https://assets.donaldjtrump.com/Trump_Economic_Plan.pdf.

2020 and Bloom, Draca, and Van Reenen 2016), among others.

In subsequent work, Autor, Dorn, Hanson, and Song (2014), use longitudinal linked employee-employer data from social security records to provide evidence on the consequences for individual workers of differential exposure to Chinese imports across industries. Comparing workers at the 75th and 25th percentiles of exposure to Chinese imports, the more exposed worker experiences a reduction of cumulative earnings by approximately one half of the initial annual wage. Most of this net reduction in cumulative earnings operates through a reduction in within-year earnings rather than from additional time with zero earnings. Cumulative earnings fall by around 50 percent more than cumulative employment at both the initial employer and in the initial industry. Earnings gains in other manufacturing industries are only half as large as the losses incurred with the original employer and industry. This finding of substantial income losses from job displacement is consistent with a broader literature in labor economics, including Jacobson, LaLonde, and Sullivan (1993) and Neal (1995). Understanding the mechanisms underlying these earnings losses from job displacement, and the extent to which they reflect the loss of firm and industry-specific human capital remains an area where more work is needed.

Another group of studies has provided evidence of uneven geographical effects of trade liberalization reforms. In an early important contribution, Topalova (2010) finds that rural districts in India that were more exposed to India's import liberalization experienced slower declines in poverty and lower consumption growth. McCaig (2011) find that Vietnamese provinces more exposed to export liberalization from the 2001 U.S.-Vietnam Bilateral Trade Agreement (BTA) saw faster reductions in poverty from 2002-4.[14] Consistent with the predictions of a multi-region specific-factors model, Kovak (2013) finds that Brazilian local labor markets that were more exposed to Brazil's tariff reductions exhibited larger wage declines. A particularly striking finding in Dix-Carneiro and Kovak (2017) is that Brazil's trade liberalization induced persistent dynamic differences between regions, with the impact of tariff changes on regional earnings 20 years after liberalization equal to around three times the effect after 10 years. Two potential mechanisms for these persistent dynamic differences between regions are slow downward adjustment in complementary capital stocks and agglomeration economies, both of which receive some support from the data. Understanding these persistent local labor market effects, the muted population response, and the importance of labor force participation as an adjustment margin are important areas for further research looking ahead.[15]

2.2 Interpretation

An enduring legacy of these reduced-form empirical findings has been to focus trade economists on geography as a neglected dimension along which the distributional effects of international trade occur. In thinking about these distributional effects, traditional trade theories such as the Heckscher-Ohlin model had concentrated on national labor markets. However, if labor is imperfectly mobile across space, worker outcomes are heavily influenced by the local labor market. In a world in which industries are geographically concentrated, shocks to local labor demand for different types of workers can be considerably larger and more concentrated than the corresponding shocks to national demand for these types of workers. In highlighting these uneven geographical effects, Autor, Dorn, and Hanson (2013a) has opened up a wide range of areas for further research in interpreting and elaborating these empirical

[14] For evidence on the impact of this 2001 BTA on labor allocation between the formal and informal sector, see McCaig and Pavcnik (2018)

[15] Amior and Manning (2018) finds persistent differences in the employment-population ratios across U.S. local labor markets, despite migration responses. These results are rationalized by serially-correlated labor demand shocks, which generate a "race" between employment declines and population responses. In this race, population lags behind employment, thereby generating persistent differences in employment-population ratios. Using data for an extended sample period, Autor, Dorn, and Hanson (2020) finds persistent impacts of the China shock on U.S. local labor markets.

results. In the remainder of this section, we discuss some of the areas for further debate and outstanding issues where additional research is needed.

Relative Versus Aggregate Effects As discussed in Autor, Dorn, and Hanson (2013a), equation (1) is a "difference-in-difference" regression specification that identifies *relative* effects between local labor markets with different levels of import exposure, and does not identify aggregate level or general equilibrium effects, which are captured along with other aggregate shocks in the regression constant γ_t. As part of the analysis in that paper, a quantitative exercise is reported in which rising exposure to Chinese import competition is found to explain 44 percent of the U.S. manufacturing employment decline for the full sample period from 1990-2007. However, this quantitative exercise uses an additional assumption that there is one local labor market in which the China shock has zero effect on the manufacturing employment shares of the working-age population, in which case the effects relative to that one local labor market can be used to compute an aggregate effect.

More generally, separating aggregate effects from other contemporaneous economy-wide shocks is extremely challenging from the point of view of empirical identification. In order to abstract from other economy-wide shocks, as well as to compute unobserved model-based objects such as welfare, one typically requires a general equilibrium model. One promising approach is to use the reduced-form "differences-in-differences" estimates from Autor, Dorn, and Hanson (2013a) as a targeted moment for the model in either a calibration or a simulated method of moments (SMM) / indirect inference estimation procedure, and then use the structure of the model to evaluate the implied aggregate or general equilibrium effect, as discussed further in Section 4.8 below.[16] In general, subsequent research using quantitative spatial models has suggested that the aggregate welfare effects from the China shock are small relative to the distributional consequences across local labor markets.

Mechanisms and Measurement A second set of issues concerns the underlying economic mechanisms through which trade shocks affect the economy and the appropriate measurement of these trade shocks. Autor, Dorn, and Hanson (2013a) provides compelling evidence of worse labor market outcomes in local labor markets more exposed to the China shock. However, other research suggests that international trade also affects welfare through the separate mechanism of the price of tradeable consumption goods, including Amiti, Dai, Feenstra, and Romalis (2017), Borusyak and Jaravel (2018) and Fajgelbaum and Khandelwal (2016). Furthermore, some of the more negative effects of the China shock in more exposed local labor markets may be offset by adjustments in the prices of local goods and services, including the price of local fixed factors, such as durable building structures and land.[17] From the perspective of welfare, what matters is the effect of international trade on real income, after taking all of these mechanisms into account, including both effects in the labor market through nominal income, and effects in the product market through the prices of tradeable consumption goods and local fixed factors.

The fact that there are multiple mechanisms through which trade affects the economy also raises the question of other possible measures of trade shocks. First, should one focus on imports from one country (e.g. China), from a group of countries (e.g. low-wage countries), or from all countries (including both high and low-wage countries)? Second, while imports of final goods increase domestic market competition and reduce domestic labor market demand, those of

[16] In a broader discussion of the macroeconomics literature, Nakamura and Steinsson (2018) refers to these targeted moments used to a discipline a model as "identified moments," in the sense that these moments permit credible empirical identification of model parameters, while the structure of the model enables one to evaluate other objects of interest for which credible empirical identification is hard to achieve.

[17] For theory and evidence on the importance of durable housing for asymmetries in urban growth and decline, see Glaeser and Gyourko (2005).

intermediate goods reduce domestic production costs and increase domestic labor market demand. Third, a key insight from neoclassical theory is that trade involves both imports and exports, which raises the challenge of quantifying the positive effects of increased labor market demand for the export market in additional to any negative effects of reduced labor market demand in the domestic market. Fourth, in thinking about the impact of trade on domestic welfare, further general equilibrium effects occur through third markets, as when increased import competition from China in European markets reduces the demand for American goods in those markets.[18]

All of these mechanisms are considered in Autor, Dorn, and Hanson (2013a). Additional empirical specifications are reported to incorporate effects through U.S. exports to China, imports of intermediate inputs, and third-market effects. Nevertheless, simultaneously incorporating all of these mechanisms and quantifying their relative importance is a challenging endeavor, and this remains an interesting area for further research. Subsequent research has provided further evidence on the labor market effects of U.S. exports (e.g. Feenstra, Ma, and Xu 2019) and input-output linkages (e.g. Wang, Wei, Yu, and Zhu 2018). Another interesting area for further research is the extent to which there could be heterogeneity in the local labor market effects of trade shocks. One potential dimension of heterogeneity is the magnitude of the trade deficit. Standard neoclassical theory suggests that a country's overall trade deficit is determined by macroeconomic considerations of saving and investment and struggles to find much relevance for bilateral trade deficits. Notwithstanding these caveats, Dix-Carneiro, Pessoa, Reyes-Heroles, and Traiberman (2020) develop a dynamic model of endogenous trade deficits, in which the magnitude of the overall trade deficit influences the severity of labor market adjustment to a trade shock. Another potential dimension of heterogeneity is the depth of comparative advantage. Eriksson, Russ, Shambaugh, and Xu (2019) develop a product cycle model, in which the effects of a trade shock on local labor markets are especially severe when industries are in a late-stage of the the product cycle, and argue that this insight is of relevance for the effects of the China shock on U.S. local labor markets.

Econometric Specification A third set of issues relates to econometric specification and interpretation. The difference-in-differences specification in equation (1) has an interpretation as a shift-share research design following Bartik (1991), in which one studies the impact of a set of aggregate shocks (or "shifters") on units differentially exposed to them, with the exposure measured by a set of disaggregate weights (or "shares"). A key issue in interpreting such shift-share specifications is whether one believes that the exogenous source of variation for identification is the aggregate shifters (changes in sectoral import exposure) or the disaggregate shares (shares of sectors in employment within each local labor market). Adopting the first of these two perspectives, Adão, Kolesár, and Morales (2019) shows that the presence of these unobserved shift-share components introduces a spatial dependence into the regression residuals. Using the identifying assumption that the shifters are as good as randomly assigned conditional on the regression controls and the shares, the paper develops methods for making statistical inference that take this spatial dependence into account. These methods essentially form clusters of cross-section units based on sectoral composition, which have a similar variance of a weighted sum of the regression residuals. Using both Monte Carlo evidence and two popular applications of shift-share regressions, the paper shows that computing standard errors in the correct way matters in practice.

Which of these two perspectives one takes is important not only for inference but also for identification. The

[18]For a sufficient statistics approach that highlights each of these mechanisms in the class of international trade models characterized by a constant trade elasticity, see Kleinman, Liu, and Redding (2020).

validity of shift-share instrument variables (SSIV) regressions hinges on the orthogonality between the shift-share instrument and the residual across the exposed units (e.g. local labor markets). Again taking the perspective that the aggregate shocks are the source of identifying variation, Borusyak, Hull, and Jaravel (2019) show that this orthogonality condition is equivalent to the orthogonality of the aggregate shocks and an aggregate residual (e.g. at the sector level). Therefore, the SSIV regression coefficients can be obtained from an equivalent IV regression estimated at the level of the aggregate shocks, in which the outcome and treatment variables are first averaged, using the exposure shares as weights. This equivalence result is used to show that SSIV estimates are consistent if either the aggregate shocks are as good as randomly assigned (as if arising from a natural experiment) or a law of large numbers applies at the aggregate level of the shocks (the instrument incorporates many sufficiently independent shocks, each with sufficiently small average exposure). Importantly, these conditions allow the disaggregate shares themselves to be endogenous, as is likely to be the case for the shares of sectors in employment within each local labor market, which could be affected by other common industry shocks to for example tastes or technology.

Adopting the second perspective that the disaggregate shares are the source of identifying variation, Goldsmith-Pinkham, Sorkin, and Swift (2020) show that the shift-share instrumental variables (SSIV) estimator is equivalent to an overidentified generalized method of moments (GMM) estimator using the disaggregate shares as instruments and a weight matrix constructed from the aggregate shifters. An important implication of this result is that the Bartik estimator can be decomposed into a weighted sum of just-identified instrumental variable estimators that use each disaggregate share as a separate instrument. Recovering these weights informs a researcher about which disaggregate shares are driving the estimated Bartik coefficient, and hence how sensitive this estimate is to misspecification (i.e. endogeneity) in any of these instruments.

More broadly, there remain other potential concerns about the validity of the exclusion restriction in any Bartik research design. A classic issue in the econometric literature is the so-called reflection problem, in which the aggregate shifters are influenced by correlated individual shocks across the disaggregate units. To alleviate this concern, many Bartik empirical applications use a "leave-out" version of the instrument, in which the aggregate shifter for each disaggregate unit is computed leaving out that disaggregate unit. As discussed above, to address the concern that U.S. import exposure could be endogenous to changes in the share of manufacturing in employment in U.S. local labor markets, Autor, Dorn, and Hanson (2013a) use import exposure constructed using the aggregate imports of other developed countries as an instrument. In principle, one could still raise identification concerns, such as a change in technology that leads to a common loss of comparative advantage in both the U.S. and other developed countries, and hence is responsible for both a decline in the share of manufacturing employment and increased imports from China. To address this concern, Autor, Dorn, and Hanson (2013a) report a number of specification checks, including an alternative measure of import exposure based on estimating a gravity equation for international trade. In subsequent joint tests of the orthogonality of the sector shocks and sector residual, Goldsmith-Pinkham, Sorkin, and Swift (2020) are unable to reject the null hypothesis of orthogonality.

Further support for substantial local labor markets effects of international trade comes from the robustness of these findings from a wide range of different empirical contexts, such as the findings from Brazil's trade liberalization discussed above. In this vein, using the trade shock from the grain invasion that followed the 1846 Repeal of the Corn Laws and an instrument based on the suitability of agroclimatic conditions for the cultivation of wheat, Heblich, Redding, and Zylberberg (2020) find substantial distributional consequences of trade across locations within England

and Wales.

Theoretical Foundations A fourth set of issues relates to the theoretical foundations for the reduced-form regression specifications estimated across local labor markets. In Autor, Dorn, and Hanson (2013a), the reduced-form regression (1) is derived from a first-order approximation to a gravity equation model of international trade. In Kovak (2013), a related specification is derived from a first-order approximation to the specific-factors model of international trade, in which labor is assumed to immobile across regions but mobile across sectors, whereas the specific factors are immobile across both regions and sectors.

More broadly, the finding that local labor markets more exposed to trade shocks have substantially and statistically significantly worse labor market outcomes, with little evidence of population movements, raises the question of how these empirical results can be consistent with spatial equilibrium.[19] To the extent that population is mobile across local labor markets, and some are more adversely affected by the trade shock than others, we would expect to observe population movements across these local labor markets. Furthermore, looking further back in history, we have examples in which changes in economic and political opportunities led to large-scale population movements, including for example the migration of African-Americans from the South to the North, as examined in Wilkerson (2011) and Platt Boustan (2020). A broader puzzle for the theoretical and empirical literature is understanding the reasons why geographical mobility (and other measures of mobility) have declined in the U.S. economy in recent decades.[20] Despite this decline in geographical mobility, some evidence has begun to emerge of relatively small population responses to the China shock, as for example in Greenland, Lopresti, and McHenry (2019).

One response to these empirical findings could be that not enough time has elapsed yet, and the long-run population response ultimately could be larger than the short-run population response. This suggests frictions and adjustment costs, as in Artuç, Chaudhuri, and McLaren (2010) and Caliendo, Dvorkin, and Parro (2019), and discussed further in Section 4.8 below. Another response could be that many of the labor market outcomes are in nominal terms, and changes in nominal wages need not be fully informative about changes in real wages, if there are offsetting changes in the prices of local goods and services, including for example housing. Both of these responses push in the direction that the empirical findings from these reduced-form regression provide useful moments to discipline theoretical models. Nevertheless, the relative absence of a population response, and the relative prominence of adjustments through labor force participation, provide challenges for standard neoclassical theoretical frameworks and call out for further theoretical and empirical research.

Another issue for consideration is that these reduced-form regressions treat local labor markets as independent units in a cross-section regression, whereas in reality they are linked through goods trade, commuting and migration flows, complicating inference. Although choosing commuting zones as the measure of the local labor market minimizes commuting flows between different local labor markets, wherever one draws the boundaries of these commuting zones, there are typically some commuting flows across these boundaries. The resulting spatial correlation introduced by linkages in goods and factor markets could help to explain the findings of muted population responses to local labor market shocks. For workers with skills specific to an industry, the ability to relocate geographically may not insulate them from a trade shock, if all locations specializing in that industry are hit by the same shock.

[19] For evidence on departures from factor price equalization across U.S. local labor markets, see Bernard, Redding, and Schott (2013).

[20] For evidence on the fall in geographical mobility, see Kaplan and Schulhofer-Wohl (2017). For findings of a decline in business dynamism, see Decker, Haltiwanger, Jarmin, and Miranda (2014).

An emerging literature is starting to explore the implications of these spatial linkages for interpreting reduced-form specifications, including Monte, Redding, and Rossi-Hansberg (2018) and Adão, Arkolakis, and Esposito (2019).

The interpretation of the effects of trade shocks can be further complicated by reallocations of resources between locations within firms. Large U.S. manufacturing firms typically operate many plants, only some of which are in the manufacturing sector, while other plants undertake research and development (R&D), perform headquarter services, or operate in the wholesale and retail sectors. Although there is strong evidence of increased exit and a contraction of employment at U.S. manufacturing plants more exposed to Chinese imports, there is also evidence that firms operating those plants simultaneously expand production at their non-manufacturing plants, as in Magyari (2017), Bloom, Handley, Kurman, and Luck (2020) and Ding, Fort, Redding, and Schott (2019).

3 Modeling Economic Activity Between Cities and Regions

While the research discussed in the previous section uses the uneven distribution of economic activity within countries to quantify the distributional consequences of international trade, we now turn to research that seeks to explain this uneven distribution of economic activity. Traditionally, research on economic geography considered stylized settings with a small number of symmetric locations.[21] More recent research has sought to develop quantitative spatial models that are sufficiently rich as to connect directly with the observed data on many asymmetric locations. These frameworks incorporate differences across locations in both first-nature geography (locational endowments, such as access to the coast) and second-nature geography (the proximity of economic agents relative to one another in geographic space). Typically, these models have the property that they can be inverted to recover unobserved location characteristics (e.g. productivity, amenities and trade costs) that exactly rationalize the observed data on the endogenous variables of the model as an equilibrium outcome. Therefore, these frameworks can be used to quantify the roles of first and second-nature geography in explaining the observed distribution of economic activity. Nevertheless, these frameworks remain sufficiently tractable as to permit an analytical characterization of the existence and uniqueness of the general equilibrium and to be used for realistic counterfactuals.

In this section, we examine quantitative spatial models of the distribution of economic activity across regions or systems of cities, where geography matters because of goods trade costs and migration frictions between locations. In a later section, we consider quantitative urban models of the internal structure of cities, which introduces another dimension of geography in the form of commuting costs between locations. To connect with the prior theoretical literature on economic geography across regions and cities, we consider a multi-region version of the Helpman (1998) model, in which there is a single production sector and the costs of trading goods are the only friction between locations. Nevertheless, a key insight of recent research on economic geography is that there is an entire class of quantitative spatial models with a constant trade elasticity that are isomorphic in terms of predictions for the spatial distribution of activity, as shown in Section C of the online appendix. This class of models includes frameworks of perfect competition and external economies of scale, such as a version of the Eaton and Kortum (2002) model (as examined in Redding 2016) and a version of the Armington (1969) model (as considered in Allen and Arkolakis 2014).

In this class of quantitative spatial models, the key exogenous differences in first-nature geography across locations are productivity (A_n), amenities such as climate and scenic views (B_n), the supply of floor space (H_n), and

[21] See the classic theoretical papers by Krugman (1991c) and Krugman and Venables (1995).

bilateral trade costs between locations (d_{ni}). The spatial distribution of economic activity is determined by the interaction between these exogenous features of first-nature geography and endogenous second-nature geography from agglomeration and dispersion forces. In the Helpman (1998) model, agglomeration forces arise from the combination of love of variety, increasing returns to scale and transport costs, while dispersion forces take the form of a perfectly inelastic supply of floor space. In the wider class of quantitative spatial models with a constant trade elasticity, alternative sources of agglomeration forces can be introduced, such as endogenous components of productivity (A_n) and amenities (B_n), because of external economies of scale.

For most of this section, we focus for simplicity on a static model of worker and firm locations, such that we are concerned with the comparative statics of steady-state distributions of economic activity. Later in this section, we extend the analysis to incorporate multiple sectors and migration frictions, in order to connect with the reduced-form empirical literature on local labor markets considered in the previous section. In some of these extensions, worker and firm location decisions become dynamic, and we are concerned with both the comparative statics of steady-state distributions of economic activity and the transition dynamics between steady-states.

We now introduce the basic model structure. We consider an economy consisting of a set of locations indexed by $n \in N$. Each location is endowed with an exogenous supply of floor space (H_i).[22] The economy as a whole is endowed with a measure $\bar{L}$ of workers, where each worker has one unit of labor that is supplied inelastically with zero disutility.[23] Workers are perfectly geographically mobile and hence in equilibrium amenity-adjusted real wages are equalized across all populated locations. Locations are connected by a bilateral transport network that can be used to ship goods subject to symmetric iceberg trade costs, such that $d_{ni} = d_{in} \geq 1$ units must be shipped from region i in order for one unit to arrive in region n, where $d_{ni} > 1$ for $n \neq i$ and $d_{nn} = 1$.

3.1 Consumer Preferences

Preferences are defined over goods consumption and residential floor space use. We assume that these preferences take the Cobb-Douglas form, such that indirect utility for a worker in location n is given by:[24]

$$U_n = \frac{B_n v_n}{P_n^{\alpha} Q_n^{1-\alpha}}, \qquad 0 < \alpha < 1, \tag{2}$$

where v_n is worker income; P_n is the consumption goods price index; Q_n is the price of floor space; and recall that B_n denotes amenities.[25] The consumption goods price index is assumed to take the constant elasticity of substitution (CES) form:

$$P_n = \left[\sum_{i \in N} \int_0^{M_i} p_{ni} \left(\psi \right)^{1-\sigma} d\psi \right]^{\frac{1}{1-\sigma}}, \tag{3}$$

where M_i is the endogenous measure of varieties produced in each location; $p_{ni}(\psi)$ is the cost to a consumer in location n of a variety ψ from location i; and we assume that varieties are substitutes ($\sigma > 1$).

[22] Allowing for an endogenous supply of floor space that is a constant elasticity function of its price following Saiz (2010) is straightforward. Hsieh and Moretti (2019) show that heterogeneity in this floor space elasticity that is correlated with location productivity can have important aggregate implications for economy-wide productivity.

[23] Although for simplicity we assume that all workers are *ex ante* identical with a single type of labor, an interesting area of active research is the spatial sorting of workers who are heterogeneous in terms of human capital or skills across space, including Moretti (2013), Diamond (2016), Fajgelbaum and Gaubert (2020), Davis and Dingel (2020) and Rossi-Hansberg, Sarte, and Schwartzman (2020).

[24] For empirical evidence using U.S. data in support of the constant housing expenditure share implied by the Cobb-Douglas functional form, see Davis and Ortalo-Magné (2011).

[25] A straightforward generalization is to allow workers to have idiosyncratic amenity draws for each location from an extreme value distribution, as in Redding (2016). In this generalization, each location faces an upward-sloping supply curve for labor, with an elasticity determined by the shape parameter of this extreme value distribution.

3.2 Production

Varieties are produced under conditions of monopolistic competition and increasing returns to scale. To produce a variety in a location, a firm must incur a fixed cost of F units of labor and a constant variable cost in terms of labor that depends on that location's productivity (A_i).[26] Therefore the total amount of labor ($l_i(\psi)$) required to produce $x_i(\psi)$ units of a variety ψ in location i is:

$$l_i(\psi) = F + \frac{x_i(\psi)}{A_i}. \tag{4}$$

Profit maximization implies that equilibrium prices are a constant markup over marginal cost of supplying a variety to a market. Therefore, the cost to a consumer in market n of consuming variety ψ from location i is given by:

$$p_{ni}(\psi) = p_{ni} = \left(\frac{\sigma}{\sigma - 1}\right) d_{ni} \frac{w_i}{A_i}, \tag{5}$$

where w_i is the wage. With free entry and exit of varieties, equilibrium profits are zero. Using this equilibrium pricing rule (5) in the requirement that profits are zero, equilibrium output of each variety is equal to a constant that depends on location productivity:

$$x_i(\psi) = \bar{x}_i = A_i(\sigma - 1)F. \tag{6}$$

Using this solution for equilibrium output (6) in the production technology (4), equilibrium employment for each variety is the same for all locations: $l_i(\psi) = \bar{l} = \sigma F$. Given this constant equilibrium employment for each variety, labor market clearing implies that the total measure of varieties supplied by each location is proportional to the endogenous supply of workers choosing to locate there:

$$M_i = \frac{L_i}{\sigma F}. \tag{7}$$

3.3 Price Indices and Expenditure Shares

Using equilibrium prices (5) and labor market clearing (7), the price index (3) can be expressed in terms of wages and employment in each location:

$$P_n = \frac{\sigma}{\sigma - 1} \left(\frac{1}{\sigma F}\right)^{\frac{1}{1-\sigma}} \left[\sum_{i \in N} L_i \left(d_{ni} w_i / A_i\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}, \tag{8}$$

where the presence of employment in this expression reflects the fact that the measure of varieties produced in each location is endogenous to the measure of workers that choose to live in that location.[27]

Using the CES expenditure function, equilibrium prices (5) and labor market clearing (7), the share of location n's expenditure on goods produced in location i is:

$$\pi_{ni} = \frac{M_i p_{ni}^{1-\sigma}}{\sum_{k \in N} M_k p_{nk}^{1-\sigma}} = \frac{L_i \left(d_{ni} w_i / A_i\right)^{1-\sigma}}{\sum_{k \in N} L_k \left(d_{nk} w_k / A_k\right)^{1-\sigma}}. \tag{9}$$

The model therefore implies a "gravity equation" for goods trade, where the bilateral trade between locations n and i depends on both "bilateral resistance" (bilateral trade costs d_{ni}) and "multilateral resistance" (trade costs to all other

[26] Although for simplicity we assume a representative firm in each location, a straightforward generalization is to introduce firm heterogeneity with an untruncated Pareto productivity distribution following Melitz (2003). Similarly, the production technology can be generalized to include intermediate inputs and input-output linkages following Caliendo and Parro (2015) and Caliendo, Parro, Rossi-Hansberg, and Sarte (2018). Finally, another extension allows firms to have heterogeneous productivities across several locations, and choose in which of these locations to produce, depending for example on tax incentives, as in Serrato and Zidar (2016) and Fajgelbaum, Morales, Suárez Serrato, and Zidar (2019).

[27] Although we focus for simplicity on a single consumption goods sector and the overall level of economic activity, geography can also influence structural transformation across sectors and the composition of economic activity, as in Fajgelbaum and Redding (2018).

locations d_{nk}), as in Anderson and van Wincoop (2003). A key parameter in this gravity equation is the partial elasticity of bilateral trade with respect to trade costs ($d\ln(\pi_{ni}/\pi_{nn})/d\ln d_{ni} = -(\sigma-1)$), which is determined here by the elasticity of substitution between varieties (σ). Together (8) and (9) imply that each location's price index can be again written in terms of its trade share with itself, such that:

$$P_n = \frac{\sigma}{\sigma-1}\left(\frac{L_n}{\sigma F \pi_{nn}}\right)^{\frac{1}{1-\sigma}}\frac{w_n}{A_n}. \tag{10}$$

3.4 Market Clearing

Expenditure on floor space in each location is assumed to be redistributed lump sum to the workers residing in that location.[28] Combining this lump-sum redistribution with the implication of Cobb-Douglas utility that expenditure on floor space is a constant share of income, we find that income in each location ($v_n L_n$) is a constant multiple of labor income in each location ($w_n L_n$):

$$v_n L_n = w_n L_n + (1-\alpha) v_n L_n = \frac{w_n L_n}{\alpha}. \tag{11}$$

Goods market clearing implies that revenue in each location equals expenditure on goods produced in that location. We assume that trade in balanced, such that expenditure equals income in each location.[29] Using zero profits, which implies that revenue equals labor income, and utility maximization, which implies that expenditure on goods is a constant share of income, this goods market clearing condition can be written as:

$$w_i L_i = \sum_{n \in N} \alpha \pi_{ni} v_n L_n = \sum_{n \in N} \pi_{ni} w_n L_n. \tag{12}$$

Land market clearing implies that the supply of floor space equals the demand for floor space. Using utility maximization in this land market clearing condition, the price of floor space (Q_n) is given by:

$$Q_n = \frac{(1-\alpha) v_n L_n}{H_n} = \frac{1-\alpha}{\alpha}\frac{w_n L_n}{H_n}. \tag{13}$$

3.5 Population Mobility

Population mobility implies that workers receive the same real income in all populated locations:

$$U_n = \frac{B_n v_n}{P_n^{\alpha} Q_n^{1-\alpha}} = \bar{U}, \tag{14}$$

Using the price index (10), the equality between expenditure and income in each location (11), and land market clearing (13) in indirect utility (2), we can write this population mobility condition as follows:

$$\bar{U} = \frac{A_n^{\alpha} B_n H_n^{1-\alpha} \pi_{nn}^{-\alpha/(\sigma-1)} L_n^{-\frac{\sigma(1-\alpha)-1}{\sigma-1}}}{\alpha\left(\frac{\sigma}{\sigma-1}\right)^{\alpha}\left(\frac{1}{\sigma F}\right)^{\frac{\alpha}{1-\sigma}}\left(\frac{1-\alpha}{\alpha}\right)^{1-\alpha}}. \tag{15}$$

Re-arranging this population mobility condition to take population (L_n) over to the left-hand side, and dividing this expression by its sum across all locations, we obtain the following result that the population share of each location

[28] An alternative assumption is that expenditure on floor space is paid into an economy-wide portfolio, in which workers in each location hold shares, as in Caliendo, Parro, Rossi-Hansberg, and Sarte (2018). Although these different assumptions are useful simplifications, an interesting area for further research is developing quantitative spatial models with distributions of asset ownership across heterogeneous agents.

[29] Following Dekle, Eaton, and Kortum (2007) in the quantitative international trade literature, it is straightforward to introduce exogenous trade deficits, but a more satisfactory approach is to endogenize these deficits as in Eaton, Kortum, and Neiman (2016) and Reyes-Heroles (2016).

($\lambda_n \equiv L_n/\bar{L}$) depends on its productivity (A_n), amenities (B_n), supply of floor space (H_n) and domestic trade share (π_{nn}) relative to those of all other locations:

$$\lambda_n = \frac{L_n}{\bar{L}} = \frac{\left[A_n^{\alpha} B_n H_n^{1-\alpha} \pi_{nn}^{-\alpha/(\sigma-1)}\right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}{\sum_{k \in N}\left[A_k^{\alpha} B_k H_k^{1-\alpha} \pi_{kk}^{-\alpha/(\sigma-1)}\right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}. \tag{16}$$

Intuitively, locations with higher productivity (A_n), amenities (B_n), supply of floor space (H_n), and market access (lower own trade shares π_{nn}) relative to other locations have higher equilibrium population shares (λ_n).

3.6 General Equilibrium

The general equilibrium of the model can be again represented by the share of workers in each location ($\lambda_n = L_n/\bar{L}$), the share of each location's expenditure on goods produced by other locations (π_{ni}) and the wage in each location (w_n). Using goods market clearing (12), the trade share (9), and population mobility (16), this equilibrium triple $\{\lambda_n, \pi_{ni}, w_n\}$ solves the following system of equations for all $i, n \in N$:

$$w_i \lambda_i = \sum_{n \in N} \pi_{ni} w_n \lambda_n, \tag{17}$$

$$\pi_{ni} = \frac{\lambda_i \left(d_{ni} w_i / A_i\right)^{1-\sigma}}{\sum_{k \in N} \lambda_k \left(d_{nk} w_k / A_k\right)^{1-\sigma}}, \tag{18}$$

$$\lambda_n = \frac{\left[A_n^{\alpha} B_n H_n^{1-\alpha} \pi_{nn}^{-\alpha/(\sigma-1)}\right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}{\sum_{k \in N}\left[A_k^{\alpha} B_k H_k^{1-\alpha} \pi_{kk}^{-\alpha/(\sigma-1)}\right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}. \tag{19}$$

In contrast to international trade models, in which country labor endowments are exogenous, the population of each population is endogenously determined in this system of equations (17)-(19). This population mobility, together with love of variety, increasing returns to scale and transport costs, gives rise to agglomeration forces from forward and backward linkages between firms and consumers (see Hirschman 1958 and Krugman 1991a). The forward linkage runs downstream from firms to consumers: love of variety implies that consumers demand all varieties and transport costs imply that they want to locate close to these varieties. The backward linkage runs upstream from consumers to firms: increasing returns to scale imply that firms want to concentrate production of their variety in a single location and transport costs imply that they want this location to be close to markets. Together, these forward and backward linkages engender a virtuous circle of cumulative causation, which acts an agglomeration force: consumers want to locate close to firms and firms want to locate close to consumers. Working against this agglomeration force is a congestion force from an inelastic supply of floor space: as more economic activity concentrates in a location, this bids up the price of floor space, making that location less attractive relative to other locations.

The general equilibrium distribution of economic activity that solves the system of equations (17)-(19) reflects the interaction between these agglomeration and dispersion forces (second-nature geography) and the exogenous differences in productivity, amenities, floor space supply and transport costs across locations (first-nature geography). If the agglomeration forces are sufficiently strong relative to the dispersion forces in the model, there is the potential for multiple equilibria, and we now turn in the next section to a formal characterization of the conditions under which there exists a unique equilibrium versus multiple equilibria.

3.7 Existence and Uniqueness

The properties of the general equilibrium of the model can be characterized by combining the gravity structure of international trade with the population mobility condition. Assuming symmetric trade costs ($d_{ni} = d_{in}$), and following the arguments in Allen and Arkolakis (2014), we can reduce the general equilibrium of the model to the following system of N equations that determine the N populations of each location in terms of the exogenous location characteristics (A_n, B_n, H_n, d_{ni}) and parameters (σ, α, F):[30]

$$L_n^{\tilde{\sigma}\gamma_1} A_n^{-\frac{(\sigma-1)(\sigma-1)}{2\sigma-1}} B_n^{-\frac{\sigma(\sigma-1)}{\alpha(2\sigma-1)}} H_n^{-\frac{\sigma(\sigma-1)(1-\alpha)}{\alpha(2\sigma-1)}} = \frac{\bar{W}^{1-\sigma}}{\sigma F} \sum_{i \in N} \left(\frac{\sigma}{\sigma - 1} d_{ni} \right)^{1-\sigma} \left(L_i^{\tilde{\sigma}\gamma_1} \right)^{\frac{\gamma_2}{\gamma_1}} A_i^{\frac{\sigma(\sigma-1)}{2\sigma-1}} B_i^{\frac{(\sigma-1)(\sigma-1)}{\alpha(2\sigma-1)}} H_i^{\frac{(\sigma-1)(\sigma-1)(1-\alpha)}{\alpha(2\sigma-1)}}, \quad (20)$$

where the scalar $\bar{W}$ is determined by the requirement that the labor market clear ($\sum_{n \in N} L_n = \bar{L}$) and

$$\tilde{\sigma} \equiv \frac{\sigma - 1}{2\sigma - 1}, \qquad \gamma_1 \equiv \frac{\sigma(1-\alpha)}{\alpha},$$

$$\gamma_2 \equiv 1 + \frac{\sigma}{\sigma - 1} - \frac{(\sigma - 1)(1-\alpha)}{\alpha},$$

as shown in the online appendix. This system of equations (20) summarizes how the population of each location (L_n) is influenced by first-nature geography (A_n, B_n, H_n, d_{ni}) and second-nature geography (the population of all other locations). We are now in a position to formally state the conditions under which there exists a unique equilibrium.

Proposition 1 *Assume $\sigma(1-\alpha) > 1$. Given the productivities, amenities and floor space in each location $\{A_n, B_n, H_n\}$, and symmetric bilateral trade frictions $\{d_{ni} = d_{in}\}$ between all pairs of locations $n, i \in N$, there exist unique equilibrium populations in each location (L_n^*) that solve the system of equations (20).*

Proof. Under the assumption $\sigma(1-\alpha) > 1$, we have $\gamma_2/\gamma_1 < 1$, which implies that there exists a unique solution to this system of equations (20), as shown in Fujimoto and Krause (1985) and Allen and Arkolakis (2014). ■

When the parameter restriction $\sigma(1-\alpha) > 1$ is not satisfied, the model can have multiple equilibria, such that the spatial distribution of economic activity is not uniquely pinned down by exogenous location characteristics (A_n, B_n, H_n, d_{ni}). This restriction on parameters for a unique equilibrium ($\sigma(1-\alpha) > 1$) has an intuitive interpretation and corresponds to the assumption that agglomeration forces are not too strong relative to dispersion forces. A higher elasticity of substitution (σ) reduces consumer love of variety, which weakens agglomeration forces, because consumers care less about locating close to large numbers of varieties. A higher share of expenditure on floor space ($1-\alpha$) strengthens dispersion forces in the model, because as economic activity concentrates in a given location and bids up the price of floor space, this now has a larger impact on consumers' cost of living.

Although we have derived Proposition 1 for the Helpman (1998) economic geography model, analogous results hold for the entire class of quantitative spatial models characterized by a constant trade elasticity, as shown for a version of the Eaton and Kortum (2002) model from Redding (2016) and a version of the Armington (1969) model from Allen and Arkolakis (2014) in Section C of the online appendix. As data are typically observed for discrete regions, we have focused throughout the exposition on this case of discrete regions rather than on continuous space. Nonetheless,

[30] For a more general characterization of the conditions for the existence and uniqueness of the equilibrium, without assuming symmetric trade costs, see Allen, Arkolakis, and Takahashi (2019).

we can instead consider the case of continuous space, in which case equation (20) corresponds to an integral equation, with an integral rather than a sum on the right-hand side. For this case of continuous space, Allen and Arkolakis (2014) show that similar results for existence and uniqueness apply using Theorem 2.19 in Zabreyko, Koshelev, Krasnoselskii, Mikhlin, Rakovshchik, and Stetísenko (1975).

This characterization of the conditions for existence and uniqueness in this class of quantitative spatial models with a constant trade elasticity is important for a number of reasons. First, for parameter values for which $\sigma(1-\alpha) > 1$, this characterization yields an algorithm for solving for the unique fixed point of this system of equations, in which one starts with an initial guess for the vector of equilibrium populations (L_n), and then updates that guess using the solution of the system. Second, for these parameter values, this characterization ensures that counterfactual changes in location characteristics (e.g. in trade costs d_{ni} as a result of transport infrastructure improvements) will have determinate effects on the spatial equilibrium. Therefore, we can use this class of quantitative spatial models to assess the general equilibrium effects of a wide range of counterfactual interventions.

For these reasons, the literature on quantitative spatial models has largely concentrated on the range of parameter values for which there exists a unique spatial equilibrium distribution of economic activity. This focus creates a tension with the earlier theoretical literature on economic geography, as synthesized in Fujita, Krugman, and Venables (1999), which was often motivated by the idea that multiple equilibrium spatial distributions of economic activity could emerge from a "featureless plain" of symmetric space. At a more conceptual level, whether a model is characterized by multiple equilibria may depend on its level of abstraction. If we omit from our model the relevant idiosyncratic factors that in reality determined one allocation rather than another, our model can exhibit multiple equilibria. Once we include more of these idiosyncratic factors into the model, the equilibrium distribution of economic activity can become unique. Nevertheless, no model can include all such idiosyncratic factors, since otherwise it would cease to be a model, and would instead become a description of reality. Therefore, whether one uses a model with a single equilibrium or multiple equilibria could be viewed in part as a practical question of what is the most useful way to think about the world, which in turn could depend on the question at hand and the data available.

3.8 Recovering Locational Fundamentals

As quantitative spatial models incorporate both first-nature geography (productivity, amenities, floor space supply and trade costs) and second-nature geography (the endogenous location of agents relative to one another), they provide frameworks for assessing the relative importance of these two sets of determinants of the spatial distribution of economic activity. Typically, these models have an invertibility property, such that given the observed data on the endogenous variables and values for the model parameters, one can solve for the unique unobserved exogenous location characteristics that exactly rationalize the observed data as an equilibrium outcome. In the remainder of this subsection, we illustrate this invertibility property using the Helpman (1998) model, assuming that the model parameters (σ, α) are known and that we observe wages (w_n), populations (L_n) and bilateral trade shares (π_{ni}) between locations.

Under our assumption that bilateral trade costs are symmetric ($d_{ni} = d_{in}$) and there are no internal trade costs ($d_{nn} = d_{ii} = 1$), we can recover bilateral trade costs from the observed bilateral trade shares (π_{ni}) using the Head

and Ries (2001) index. Using these assumptions and equation (9), we have:

$$d_{ni}^{1-\sigma} = \left(\frac{d_{ni}d_{in}}{d_{nn}d_{ii}}\right)^{\frac{1-\sigma}{2}} = \left(\frac{\pi_{ni}\pi_{in}}{\pi_{nn}\pi_{ii}}\right)^{\frac{1}{2}}. \tag{21}$$

Using these solutions for bilateral trade costs (d_{ni}) and the trade share (9), we can solve for productivities from the goods market clearing condition (12):

$$w_i L_i = \sum_{n \in N} \frac{L_i \left(d_{ni} w_i / A_i\right)^{1-\sigma}}{\sum_{k \in N} L_k \left(d_{nk} w_k / A_k\right)^{1-\sigma}} w_n L_n, \tag{22}$$

where all variables in this equation are either observed (w_n, L_n) or already have been solved for (d_{ni}) except for productivity (A_i). As the fraction on the right-hand side of this equation is homogenous of degree zero in productivity, these productivities only can be determined up to a normalization or choice of units in which to measure them. Given this normalization (e.g. setting productivity in one location equal to one), equation (22) determines a unique vector of productivities.

Using these solutions for bilateral trade costs ($d_{ni}^{1-\sigma}$) and productivities (A_i), we can recover a composite of amenities and floor space supply from the location choice probabilities (16):

$$\lambda_n = \frac{L_n}{\bar{L}} = \frac{\left[A_n^{\alpha} \mathbb{B}_n \pi_{nn}^{-\alpha/(\sigma-1)}\right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}{\sum_{k \in N} \left[A_k^{\alpha} \mathbb{B}_k \pi_{kk}^{-\alpha/(\sigma-1)}\right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}, \tag{23}$$

where composite amenities are defined as $\mathbb{B}_n \equiv B_n H_n^{1-\alpha}$. Again all variables in this equation are either observed (L_n, π_{ni}) or already have been solved for (A_n). As the fraction on the right-hand side of this equation is homogenous of degree zero in composite amenities, they again can be determined up to a normalization or choice of units in which to measure them. Given this normalization (e.g. setting composite amenities in one location equal to one), equation (23) determines a unique vector of composite amenities, up to this normalization.

Under our model assumptions, these unobserved location characteristics (d_{ni}, A_n, $\mathbb{B}_n$) exactly rationalize the observed data (π_{ni}, w_n, L_n) as an equilibrium outcome. Having recovered these unobserved location characteristics (d_{ni}, A_n, $\mathbb{B}_n$), we can undertake model-based decompositions to assess the relative importance of these different determinants of the observed spatial distribution of economic activity, as discussed further in the next section.

Although model inversion here takes a particularly simple form, this example illustrates broader properties of quantitative spatial models. Whether invertibility is satisfied depends on both the model and the data. On the one hand, even in a model with multiple equilibria, invertibility can be satisfied if sufficient data are observed. Conditional on these observed data, it may be possible to use the equilibrium conditions of the model recover the unique unobserved location characteristics that exactly rationalize this observed equilibrium, even if another (unobserved) equilibrium could have occurred. On the other hand, even in a model with a single equilibrium, invertibility may not be satisfied if some data are unobserved. Conditional on the observed equilibrium, there may not be enough information available to uniquely determine unobserved location characteristics, or only composites of these unobserved location characteristics may be uniquely determined (as here for composite amenities $\mathbb{B}_n$).

3.9 Counterfactuals

A key feature of quantitative spatial models is that they remain sufficiently tractable to be used for counterfactuals to evaluate empirically-relevant public policy interventions, such as the construction of a particular link in the U.S.

Interstate Highway System. These counterfactuals can be undertaken using the exact-hat methodology introduced into the quantitative international trade literature by Dekle, Eaton, and Kortum (2007). We start with the counterfactual equilibrium conditions of the model. We next rewrite these counterfactual equilibrium conditions in terms of the observed values of the endogenous variables in the initial equilibrium and the relative changes in these endogenous variables between the two equilibria. In doing so, we use a prime to denote a counterfactual value of a variable and a hat to denote the relative change of a variable between the two equilibria, such that $\hat{x} \equiv x'/x$. Adopting this exact-hat algebra approach, we can use these equilibrium conditions to solve for the counterfactual changes in the endogenous variables, given only the observed values of the endogenous variables in the initial equilibrium, and without having to solve for the high-dimensional unobserved location fundamentals. For example, instead of having to make additional assumptions to parametrize unobserved bilateral trade costs, we use the observed values of bilateral trade shares in the initial equilibrium to capture these unobserved bilateral trade costs.[31]

To undertake these counterfactuals, we require four key inputs. The first is the observed values of the endogenous variables in the initial equilibrium in the data, which are here wages (w_i), population (L_i), and bilateral trade shares (π_{ni}). The second are values for the model's structural parameters, which here are the elasticity of trade with respect to trade costs ($\sigma - 1$) and the share of floor space in expenditure ($1 - \alpha$). The third is an assumed comparative static, such as an assumed relative change in bilateral trade costs ($\hat{d}_{ni} \neq 1$ for some n, i), as a result of the construction of new transport infrastructure. The fourth is an assumption about what location characteristics remain constant in response to this comparative static, such as an assumption that productivity, amenities and the supply of floor space remain unchanged ($\hat{A}_n = 1$, $\hat{B}_n = 1$ and $\hat{H}_n = 1$ for all n).

Given these four inputs, we now re-write the system of equations for the counterfactual equilibrium (17)-(19) in terms of relative changes and the observed values of the endogenous variables in the initial equilibrium:

$$\hat{w}_i \hat{\lambda}_i w_i \lambda_i = \sum_{n \in N} \hat{\pi}_{ni} \hat{w}_n \hat{\lambda}_n \pi_{ni} w_n \lambda_n, \tag{24}$$

$$\hat{\pi}_{ni} \pi_{ni} = \frac{\pi_{ni} \hat{\lambda}_i \left(\hat{d}_{ni} \hat{w}_i / \hat{A}_i \right)^{1-\sigma}}{\sum_{k \in N} \pi_{nk} \hat{\lambda}_k \left(\hat{d}_{nk} \hat{w}_k / \hat{A}_k \right)^{1-\sigma}}, \tag{25}$$

$$\hat{\lambda}_n \lambda_n = \frac{\lambda_n \left[\hat{A}_n^{\alpha} \hat{B}_n \hat{H}_n^{1-\alpha} \hat{\pi}_{nn}^{-\alpha/(\sigma-1)} \right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}{\sum_{k \in N} \lambda_k \left[\hat{A}_k^{\alpha} \hat{B}_k \hat{H}_k^{1-\alpha} \hat{\pi}_{kk}^{-\alpha/(\sigma-1)} \right]^{\frac{\sigma-1}{\sigma(1-\alpha)-1}}}, \tag{26}$$

where recall that for our example of a change in transport infrastructure, we assume $\hat{d}_{ni} \neq 1$ for some n, i and $\hat{A}_n = 1$, $\hat{B}_n = 1$ and $\hat{H}_n = 1$ for all n.

Given the observed endogenous variables in the initial equilibrium (w_n, L_n, π_{ni}), we can solve this system of equations (24)-(26) for the unique counterfactual changes in the endogenous variables ($\hat{w}_n$, $\hat{L}_n$, $\hat{\pi}_{ni}$). For parameter values for which the model has a unique equilibrium ($\sigma(1-\alpha) > 1$), these counterfactual changes in the endogenous variables are unique, and can be recovered with a similar fixed point algorithm to that discussed above for solving for general equilibrium. Key advantages of this model-based approach are that we can evaluate counterfactual interventions that have not yet occurred; we can incorporate general equilibrium effects; and we can evaluate the

[31] Alternatively, counterfactuals can be undertaken in levels with the unobserved fundamentals recovered from the observed data, using the procedure discussed in the previous section.

effects of these interventions on model-based objects such as welfare. Furthermore, compared to the earlier literature on computable general equilibrium (CGE) models, there are relatively few general equilibrium relationships, which are derived explicitly from microfoundations, and which feature a small number of structural parameters to be estimated.[32] This parsimonious theoretical structure facilitates a transparent interpretation of counterfactual predictions in terms of the underlying economic mechanisms of the model.

From the population mobility condition (15), the welfare effects of a transport infrastructure improvement that changes bilateral trade costs (d_{ni}) can be evaluated using the two sufficient statistics of the changes in a location's trade share with itself ($\hat{\pi}_{nn}$) and its population ($\hat{L}_n$):

$$\hat{\bar{U}} = \hat{\pi}_{nn}^{-\alpha/(\sigma-1)} \hat{L}_n^{-\frac{\sigma(1-\alpha)-1}{\sigma-1}}. \tag{27}$$

Therefore, having recovered the counterfactual changes in own trade shares ($\hat{\pi}_{nn}$) and populations ($\hat{L}_n$) from solving the system of equations (24)-(26), we can compute the counterfactual changes in welfare using equation (27).

In international trade models with a constant trade elasticity, labor is immobile between countries, and hence the change in each country's domestic trade share ($\hat{\pi}_{nn}$) is a sufficient statistic for the change in its welfare in response to a change in trade costs, as shown in Arkolakis, Costinot, and Rodríguez-Clare (2012). Intuitively, holding all domestic characteristics constant, if a country's share of expenditure on its own goods falls ($\hat{\pi}_{nn} < 1$), this implies that foreign prices must have fallen relative to domestic prices, thereby improving a country's terms of trade and raising its welfare. In contrast, in this economic geography model with labor mobility, both the change in a location's domestic trade share ($\hat{\pi}_{nn}$) and the change in its population ($\hat{L}_n$) must be taken into account, as shown in Redding (2016). Intuitively, if one location experiences a larger fall in its domestic trade share (lower $\hat{\pi}_{nn}$) and hence a larger increase in its real income than another location, this attracts a population inflow (higher $\hat{L}_n$), until the price of floor space adjusts, such that the change in real income is the same across all populated locations.

4 Empirical Evidence Between Cities and Regions

In this section, we examine the empirical evidence on the predictions of this class of quantitative spatial models for the distribution of economic activity between cities and regions. In Subsection 4.1, we review recent reduced-form evidence on the impact of transport infrastructure on the location of economic activity. In Subsection 4.2, we show that these quantitative spatial models provide microfoundations for empirical measures of market access that incorporate the effects of changes in transport infrastructure. In Subsection 4.3, we characterize the theoretical properties of these measures of market access. In Subsection 4.4, we show how these measures of market access can be estimated from observed data on bilateral trade between locations. In Subsection 4.5, we discuss the measurement of bilateral trade costs in these measures of market access. In Subsection 4.6, we examine the empirical evidence on the role of market access in determining wages, the price of floor space and population.

4.1 Transport Infrastructure

A key implication of the class of quantitative models introduced in the previous section is that trade costs shape the distribution of economic activity across locations within countries. An important determinant of these trade costs

[32] For a review of the earlier literature on computable general equilibrium models, see Shoven and Whalley (1992).

is the network of transport infrastructure, in the form of canals, ports, railways and highways. With the increasing availability of geographic information systems (GIS) data on transport networks and spatially-disaggregated data on economic activity within countries, a growing body of empirical research has provided reduced-form evidence on the impact of transport infrastructure on the location of economic activity.[33]

One of the main challenges in analyzing the relationship between transport infrastructure and the location of economic activity is that investments in transport infrastructure are unlikely to be randomly assigned. On the one hand, these transport investments are often made by private-sector companies, whose search for profits could lead them to select regions that otherwise would have grown more rapidly even in the absence of these investments. On the other hand, federal, state and local governments often promote investments in transport infrastructure in lagging regions, which could target locations that otherwise would have grown less rapidly. As part of the broader credibility revolution in applied econometrics discussed in Angrist and Pischke (2010), one of the key contributions of recent empirical research has been to use quasi-experimental sources of variation to provide credible evidence on the causal impacts of transport infrastructure improvements.

Most of this reduced-form empirical research on transport infrastructure and the location of economic activity has estimated cross-section regression specifications of the following form:

$$\Delta \ln y_{it} = A_0 + A_1 \Delta r_{it} + x_{it} B_0 + \epsilon_{it}, \tag{28}$$

where $\Delta \ln y_{it}$ is the log change in an economic outcome of interest (e.g. population); Δr_{it} is a measure of the intensity of treatment with transport infrastructure, such as the number of new highway rays; x_{it} is a matrix of controls for other determinants of the economic outcome of interest; and ϵ_{it} is a stochastic error.

This regression specification (28) has a "difference-in-differences" interpretation, where the first difference is over time, and the second difference is between locations that receive different intensities of treatment with transport infrastructure. By taking differences over time, we difference out any fixed effect in the level of the economic outcome of interest. With a single cross-section of data, the regression constant (A_0) captures any common macro shocks that affect the economic outcome of interest across all locations.

To address the endogenous placement of transport infrastructure, this empirical literature has followed three main instrumental variables (IV) strategies, which we categorize using the taxonomy introduced in Redding and Turner (2015). The first approach, the *planned route IV*, is an instrumental variables strategy that uses planning maps and documents as a source of quasi-experimental variation in transport infrastructure. The second strategy, the *historical route IV*, uses historical exploration and transportation routes as sources of such variation. The third method, the *inconsequential place approach*, relies on choosing a sample of spatial units that are inconsequential in the sense that the characteristics of these units did not affect the placement of the transport infrastructure.

Planned Route IV This approach was pioneered by Baum-Snow (2007), which uses a 1947 plan of the U.S. interstate highway network as an instrument for the actual network. In particular, Baum-Snow (2007) counts the number of radial highways entering a metropolitan area on the 1947 plan and uses this variable to predict the actual number of radial highways.[34] The idea behind this instrument is that the 1947 plan was developed for military purposes and

[33] For reviews focusing on transport costs and the location of economic activity, see Donaldson (2015) and Redding and Turner (2015).

[34] Other papers using planned route instrumental variables include Michaels (2008), Duranton and Turner (2011), Duranton and Turner (2012), Duranton, Morrow, and Turner (2014), Michaels, Rauch, and Redding (2019), and Baum-Snow (2019).

the identifying assumption is that these military purposes are orthogonal to economic considerations. A variation on this approach is to use transport infrastructure that was planned but not constructed in Placebo specifications, as in Donaldson (2018). To the extent that economic activity is unrelated to these planned but not constructed links, this provides evidence that the planning process was not selecting routes based on economic considerations.

Historical Route IV This approach was introduced in a sequence of papers by Duranton and Turner (2011), Duranton and Turner (2012) and Duranton, Morrow, and Turner (2014), which use the U.S. railroad network around 1898 and the routes of historical exploration expeditions between 1535 and 1850 as sources of quasi-experimental variation for the U.S. interstate highway network at the end of the 20th century.[35] Conditional on the control variables included in the regression, the identifying assumption is that the factors that affected these historical routes do not directly affect patterns of economic activity at the end of the 20th century. The inclusion of these controls is important, because they can be used to capture other channels through which historical factors could have long-lived effects. For this reason, most studies include controls for initial levels of economic activity at the beginning of the sample period.

Inconsequential Units Approach This approach was developed in Chandra and Thompson (2000) in their analysis of the impact of access to the interstate highway system on rural counties in the U.S.. This idea behind this instrument is that if transport infrastructure is built to connect urban areas, and follows a convenient route between those cities, the characteristics of rural locations along the way are inconsequential for the route chosen. Under this identifying assumption, the transport infrastructure is as good as randomly assigned to the rural locations along the way. A variation on this approach involves constructing hypothetical transport networks as instruments, such as connecting historical treaty ports in China to major interior trading centers in Banerjee, Duflo, and Qian (2020). Another variation on this approach is to construct least-cost path networks between cities, which take into account the costs of traversing different types of terrain (e.g. hills versus valleys) in constructing highways, as in the analysis of the impact of the construction of China's National Trunk Highway System in Faber (2014).

Discussion Although any identifying assumption can be questioned, we now have a growing body of credible evidence on the causal impact of transport infrastructure investments from these studies using quasi-experimental sources of variation. A number of the most convincing studies use multiple instruments from these different approaches, including Duranton and Turner (2011), Duranton and Turner (2012), and Duranton, Morrow, and Turner (2014). With more instruments than endogenous variables, the regression specification can be estimated using either all instruments together or subsets of these instruments. Additionally, over-identification tests can be used as specification checks under the identifying assumption that one of the instruments is valid. Given that the instruments from these approaches use quite different sources of variation, if one obtains similar results using the different instruments, this strengthens the evidence in support of a causal interpretation of the results.

Main Findings In principle, the effects of transport infrastructure investment could be different for different transport technologies (e.g. roads versus rail) and could depend on how a particular investment changes the entire transport network. For example, as argued in Glaeser and Kohlhase (2004), rail transport is relatively infrastructure-heavy,

[35] Other studies using historical route instrumental variables include Baum-Snow, Brandt, Henderson, Turner, and Zhang (2017) (using Chinese road and rail networks from 1962), Garcia-López, Holl, and Viladecans-Marsal (2015) (using 18th-century postal routes and Roman roads for Spain), Hsu and Zhang (2014) (using historical Japanese railroad networks), and Martincus, Carballo, and Cusolito (2017) (using the Inca roads for Peru).

which favored a hub-and-spoke structure that reduced travel times into central locations. In contrast, road transport is relatively infrastructure-light, with dense networks of lateral connections, which are likely to have reduced travel times between outlying locations. Of these transport technologies, highways have received somewhat more attention than railways, although several empirical studies have considered both.

We focus in this section on evidence from more reduced-form approaches, before discussing more structural approaches in Section 4.6 below. Perhaps the most robust finding from this existing reduced-form literature is that improvements in transport infrastructure lead to a decentralization of population, in the sense of a decline in the share of the central city in the total population of the metropolitan area (see in particular Baum-Snow 2007). A number of studies also find evidence of employment decentralization, in the analogous sense of a decline in the share of the central city in the total employment of the metropolitan area (see for example Baum-Snow 2019).

Using U.S. data from 1983-2003, Duranton and Turner (2011) find positive effects on interstate highways on urban growth, with a 10 percent increase in a city's initial stock of highways causing about a 1.5 percent increase in its employment over this twenty-year period. Using U.S. data on interstate highways in metropolitan areas, Duranton and Turner (2012) find that vehicle kilometers travelled increase one for one with interstate highways, confirming what has been termed the "fundamental law of highway congestion," and thus suggesting that increased highway provision is unlikely to relieve congestion. Again using data on U.S. cities and the interstate highway network, Duranton, Morrow, and Turner (2014) find that more highways within cities raise the weight of exports but have no effect on the value of exports, where as highways between cities raise both the weight and value of trade. One interpretation of these results within cities is that more highways within cities change the composition of economic activity towards sectors with low value to weight ratios, highlighting the potential for transport infrastructure to affect economic activity through changes in comparative advantage across sectors.

In a study of China's National Trunk Highway System, which was built to connect provincial capitals and cities with an urban population above 500,000, Faber (2014) finds that these network connections reduce GDP growth among non-targeted peripheral counties. This pattern of results is confirmed in Baum-Snow, Brandt, Henderson, Turner, and Zhang (2017). These findings connect with the theoretical literature on economic geography, in which reductions in transport costs can increase the concentration of economic activity in existing centers through home market effects (see in particular Krugman 1991c). These findings also highlight that transport infrastructure improvements need not be a panacea for left-behind regions: transport connections run in both directions and it is quite possible that such transport infrastructure improvements accelerate rather than retard the decline of backward regions.

In a study of fifteen sub-Saharan African cities whose largest city is a port, Storeygard (2016) uses changes in the price of oil and variation in the distance of peripheral cities from this main port to identify the impact of transport costs on economic activity. An oil price increase of the magnitude experienced between 2002 and 2008 is found to raise the income of cities near that port by 7 percent relative to otherwise identical cities 500 kilometers farther away. Consistent with transport costs varying across different types of transport infrastructure, paved and unpaved roads have systematically different effects.

Interpretation Although we now have a large body of credible reduced-form evidence on the causal effects of transport infrastructure investments, there remain a number of areas for further debate and elaboration. As for the empirical evidence on the local labor market effects of international trade, most of the empirical evidence is from

"difference-in-difference" regression specifications, which identify relative effects on locations that receive more versus less transport infrastructure, but do not identify aggregate or general equilibrium effects that are common across all locations. Therefore, these specifications cannot distinguish the reallocation of existing economic activity from the creation of new economic activity, and cannot be used to identify welfare effects.

Furthermore, these reduced-form specification focus on the direct treatment effect of transport investments on locations receiving those investments. But the class of quantitative spatial models developed above suggests that there are likely to be indirect or general equilibrium effects that vary across locations. For example, a road built between locations A and B that increases economic activity in those locations may affect the nearby location C more than it affects the further away location D. Relatedly, the IV estimates of these reduced-form specifications have an interpretation as a local average treatment effect (LATE). But the class of quantitative spatial models developed above suggests that transport improvements are likely to have heterogeneous treatment effects, depending on how they affect the entire transport network and relative levels of market access across locations.

Finally, one intriguing finding from this reduced-form literature is that the IV estimates are often larger than the ordinary least squares (OLS) estimates. If transport infrastructure was selectively placed in locations that otherwise would have grown more rapidly even in the absence of those investments, one would expect the opposite pattern of results, with IV estimates that are smaller than the OLS estimates. One interpretation could be that transport infrastructure is selectively placed in locations that otherwise would have grown less rapidly for political economy reasons. Further research on the political economy of transport investments, and comparing actual with alternative transport investments is an interesting area for further research, as discussed in Section 4.5 below.

4.2 Microfounding Market Access

A large reduced-form empirical literature in regional science has examined the relationship between a variety of economic outcomes and *ad hoc* measures of access to surrounding economic activity. Following Harris (1954), market potential was typically defined in this literature as the distance-weighted sum of surrounding population:

$$MP_{nt} = \sum_{i \in N} \frac{L_{it}}{\text{dist}_{ni}}. \tag{29}$$

Although these market potential measures were typically found to be both statistically significant and quantitatively relevant for a range of economic outcomes, the economic mechanisms underlying these reduced-form correlations were largely unexplored. The class of quantitative spatial models developed above provides microfoundations for such empirical measures of access to surrounding economic activity and highlights potential underlying economic mechanisms. We now use the structure of the Helpman (1998) model to derive a theoretically-founded measure of market access. From profit maximization and zero profits, equilibrium output is a constant ($\bar{x}_i$) that depends only on location productivity (A_i) and parameters in equation (6). Using this result together with CES demand and market clearing, it follows that equilibrium prices of each variety in location i must be such as to sell exactly this constant amount ($\bar{x}_i$), given demand in all markets:

$$\bar{x}_i = p_i^{-\sigma} \sum_{n \in N} d_{ni}^{1-\sigma} \left(w_n L_n\right) \left(P_n\right)^{\sigma-1}. \tag{30}$$

But profit maximization also implies that equilibrium prices are a constant mark-up over marginal cost in equation (5). Using this result in equation (30), and re-arranging terms, we obtain a key prediction of this class of quantitative

spatial models that the equilibrium wage depends on a measure of *firm market access* (FMA):

$$w_i = \left(\frac{\sigma-1}{\sigma}\right)^{\frac{\sigma-1}{\sigma}} A_i^{\frac{\sigma-1}{\sigma}} \left(\bar{l}\right)^{-\frac{1}{\sigma}} \left(\text{FMA}_i\right)^{\frac{1}{\sigma}} . \tag{31}$$

where firm market access (FMA) corresponds to a measure of trade-cost weighted access to markets and is defined as:

$$\text{FMA}_i \equiv \sum_{n \in N} d_{ni}^{1-\sigma} \left(w_n L_n\right) \left(P_n\right)^{\sigma-1} . \tag{32}$$

The equilibrium wage in each location (w_i) in equation (31) is increasing in firm market access (FMA_i) and productivity (A_i). Other things equal, locations with low trade costs (d_{ni}) to surrounding markets have high firm market access (FMA_i). Intuitively, the lower trade costs to these surrounding markets, the greater the revenue left over after incurring these trade costs to remunerate domestic factors of production, and hence the higher wages. Both wages (w_i) and firm market access (FMA_i) are endogenous variables. Therefore, the wage equation (31) corresponds to an equilibrium relationship between these endogenous variables, which presents econometric challenges in estimating the relationship between wages and market access, as discussed further below.

Although this class of quantitative spatial models provides microfoundations for measures of market access, the theoretically-correct measure of market access in equation (32) differs from the reduced-form measure of market potential in equation (29) in several ways. First, demand in each market in the theoretically-correct measure depends not only on population, but also on wages and price indexes. Second, trade costs in the theoretically-correct measure can be modeled as a power function of distance ($d_{ni}^{1-\sigma} = \text{dist}_{ni}^{\phi(1-\sigma)}$), but the exponent on distance need not equal minus one, and is instead equal to the elasticity of trade flows with respect to distance ($\phi(1-\sigma)$).

In the class of quantitative spatial models developed above, market access also matters for the consumption goods price index, which depends on consumers' access to tradeable varieties from surrounding locations. We summarize this access to tradeable varieties using *consumer market access* (CMA_n):

$$P_n = \left(\text{CMA}_n\right)^{\frac{1}{1-\sigma}} , \tag{33}$$

$$\text{CMA}_n \equiv \sum_{i \in N} M_i \left(d_{ni} p_i\right)^{1-\sigma} . \tag{34}$$

Recalling that varieties are substitutes ($\sigma > 1$), consumer market access in equation (34) is increasing in the mass of firms in each location (M_i) and decreasing in the price of varieties in each location (p_i) and trade costs to other locations (d_{ni}). Intuitively, the lower trade costs to these surrounding sources of supply, the lower the cost of sourcing varieties from those locations, and hence the lower the cost of living.

As market access affects both wages in equation (31) and the consumption goods price index in equation (33), it also plays an important shaping the spatial distribution of population. Using the relationship between income and expenditure (11) and land market clearing (13), we can write the population mobility condition (14) in terms of population, wages and the consumption goods price index. Using the relationship between wages and market access (31) and the relationship between the price index and consumer market access (33), we can further re-write this population mobility condition to express equilibrium population in each location in terms of firm and consumer market access, exogenous location characteristics and parameters:

$$L_n = \chi A_n^{\left(\frac{\alpha}{1-\alpha}\frac{\sigma-1}{\sigma}\right)} B_n^{\frac{1}{1-\alpha}} H_n (\text{FMA}_n)^{\frac{\alpha}{(1-\alpha)\sigma}} \left(\text{CMA}_n\right)^{\frac{\alpha}{(1-\alpha)(\sigma-1)}} , \tag{35}$$

where χ is a scalar that includes the common level of utility across all locations.

Therefore, equilibrium population is increasing in productivity (A_n), amenities (B_n), the supply of floor space (H_n), firm market access (FMA_n) and consumer market access (CMA_n). Intuitively, locations with greater firm and consumer market access have higher wages and lower price indexes, which attracts population until the price of the immobile factor (floor space) is bid up, such that real income is the same in all populated locations. Again both population and firm and consumer market access are endogenous. From the characterization of the existence and uniqueness of the equilibrium in Section 3.7 above, equilibrium population in each location can be expressed solely in terms of the exogenous characteristics of all locations in the system of equations (20).

These theoretically-consistent measures of market access incorporate the effects of changes in transport infrastructure, through bilateral trade costs (d_{ni}), and have a number of advantages over the reduced-form specifications considered in the previous subsection. First, these measures of market access are explicitly derived from an underlying theoretical model, which can be used to examine general equilibrium effects that are common across all locations (including model-based objects such as welfare). Second, these measures of market access capture not only the direct effects of transport infrastructure investments on the locations receiving those investments, but also indirect effects on nearby locations, through the trade-cost weighted sums across locations. Third, these measures of market access capture heterogeneous treatment effects of transport infrastructure investments, because the effect of a reduction in bilateral trade costs (d_{ni}) between a pair of locations depends on levels of economic activity in those locations and surrounding locations (as captured by wages, population and price indexes).

4.3 Firm and Consumer Market Access

In general, firm and consumer market access are likely to be closely related to one another. If bilateral trade costs are symmetric ($d_{ni} = d_{in}$), they are in fact proportional to one another in this class of quantitative spatial models, as shown in Donaldson and Hornbeck (2016). To demonstrate this result, we start by using the definitions of firm market access (32) and consumer market access (34) to obtain a first relationship between these variables:

$$\text{FMA}_i \equiv \sum_{n \in N} d_{ni}^{1-\sigma} \left(w_n L_n\right) \text{CMA}_n^{-1}. \tag{36}$$

We next derive a second relationship between these variables using the gravity structure of trade and market clearing. From the price index (8), the definition of consumer market access (34), the gravity equation (9) and the definition of firm market access (32), we obtain the following second relationship between these variables:

$$\text{CMA}_n \equiv \sum_{i \in N} d_{ni}^{1-\sigma} \left(w_i L_i\right) \text{FMA}_i^{-1}, \tag{37}$$

as shown in the online appendix. Under the assumption of symmetric trade costs ($d_{ni} = d_{in}$), the eigenvector that solves this system of two equations (36) and (37) satisfies: $\text{FMA}_i = \psi \text{CMA}_i$, where ψ is a scalar. We thus obtain the following recursive solution for firm market access:

$$\text{FMA}_i \equiv \sum_{n \in N} d_{ni}^{1-\sigma} \left(w_n L_n\right) \psi \text{FMA}_n^{-1}. \tag{38}$$

Therefore, assuming symmetric bilateral trade costs ($d_{ni} = d_{in}$), firm and consumer market access are perfectly correlated with one another and can be recovered (up to a scalar or normalization) from the system of equations (38).

4.4 Measuring Market Access

We now show how the gravity structure of international trade can be used to estimate these theoretically-correct measures of firm and consumer market access, following Redding and Venables (2004). From the gravity equation (9), the aggregate value of bilateral trade from location i to location n can be re-written as follows:

$$X_{ni} = M_i p_i^{1-\sigma} d_{ni}^{1-\sigma} X_n P_n^{\sigma-1}. \tag{39}$$

Collecting terms, aggregate bilateral trade depends on a measure of exporter supply capacity (s_i), a measure of importer market capacity (m_n) and bilateral trade costs (d_{ni})

$$X_{ni} = s_i d_{ni}^{1-\sigma} m_n, \tag{40}$$

$$s_i \equiv M_i p_i^{1-\sigma}, \qquad m_n \equiv X_n P_n^{\sigma-1}, \tag{41}$$

Given data on aggregate bilateral trade (X_{ni}) and proxies for bilateral trade costs (d_{ni}), supply capacity (s_i) and market capacity (m_n) can be estimated as the exporter and importer fixed effects in this gravity equation (up to a normalization or choice of units in which to measure these fixed effects). Using the resulting estimates of bilateral trade costs (d_{ni}), supply capacity (s_i) and market capacity (m_n), the theoretically-correct measures of firm market access (FMA_n) and consumer market access (CMA_n) can be computed as follows:

$$FMA_i = \sum_{n \in N} d_{ni}^{1-\sigma} m_n, \qquad CMA_n = \sum_{i \in N} s_i d_{ni}^{1-\sigma},$$

where we have again used our assumption of symmetric trade costs ($d_{ni} = d_{in}$).

Consistent with the theoretical result in the previous section, Redding and Venables (2004) finds that firm and consumer market access estimated from international trade data are extremely highly correlated with one another. The fact that this correlation is not perfect is in line with the idea that in practice bilateral trade costs need not be perfectly symmetric, as found for example in Waugh (2010).

4.5 Measuring Trade Costs

One important issue for estimating market access in the previous section is the measurement of bilateral trade costs (d_{ni}). In the empirical trade literature using international trade data between countries, the traditional approach is to compute the Great-circle distance between countries' capital cities or the population-weighted average of the distance between countries' major cities.[36] Particularly at fine spatial scales within countries, this measure has obvious limitations, such as not taking into account the transport network. With the increasing availability of geographic information systems (GIS) data on transport networks and spatially-disaggregated data within countries, research in spatial economics increasingly uses least-cost path measures of bilateral trade costs that take into account both the structure of the transport network and the relative costs of alternative modes of transport.

In a path-breaking paper along many dimensions, Donaldson (2018) uses a measure of lowest-cost route effective distance, in which bilateral trade costs ($d_{ni}(\boldsymbol{R}, \boldsymbol{\delta})$) are modeled using graph theory as depending on a set of nodes and arcs ($\boldsymbol{R}$) and the cost of traveling along each arc ($\boldsymbol{\delta}$). In empirical applications, nodes are typically finely-distributed

[36] Great-circle distance is the shortest distance between two points on the surface of a sphere and can be computed from latitude and longitude coordinates using the Haversine formula. Both Great-circle distance measures are available in CEPII's GEODIST dataset (Mayer and Zignago 2011).

points in space and arcs are the available means of transportation between the nodes (e.g. a rail, road, river or coast connection). The cost of traveling along each arc is a vector ($\boldsymbol{\delta} = \left(\delta^{rail}, \delta^{road}, \delta^{river}, \delta^{coast}\right)$) that summarizes the per unit distance cost of using each mode of transport. One of these per unit distance costs is normalized to one (e.g. $\delta^{rail} = 1$) and the others are typically either assumed based on relative travel speeds (as in Ahlfeldt, Redding, Sturm, and Wolf 2015 and Heblich, Redding, and Sturm 2020) or estimated using for example observed data on price differences (as in Donaldson 2018). Given the vector of per unit distance costs ($\boldsymbol{\delta}$) and the transport network ($\boldsymbol{R}$), the lowest-cost route effective distance between any pair of locations n and i ($d_{ni}(\boldsymbol{R}, \boldsymbol{\delta})$) is assumed to equal the cost of traveling along the least-cost path between those locations using the available transport network. For any discrete set of nodes and arcs, this lowest-cost route effective distance can be computed efficiently using Dijkstra's shortest-path algorithm (Ahuja, Magnanti, and Orlin 1993).

In another path-breaking contribution along many fronts, Allen and Arkolakis (2014) uses analogous methods for constructing bilateral trade costs in continuous space. Suppose that geographic space (S) is a finite-dimensional compact manifold in $\mathbb{R}^N$. Let $\tau : S \to \mathbb{R}_+$ be a continuous function where $\tau(i)$ gives the "instantaneous" trade cost incurred by crossing point $i \in S$. Define $t(n, i)$ as the solution to the following least-cost path minimization problem:

$$t(n,i) = \inf_{g \in \Gamma(n,i)} \int_0^1 \tau\left(g(t)\right) \left\| \frac{dg(t)}{dt} \right\| dt, \tag{42}$$

where $g : [0,1] \to S$ is a path and $\Gamma(n,i) \left\{ g \in C^1 | g(0) = n, g(1) = i \right\}$ is the set of all possible continuous and once-differentiable paths that lead from location n to location i; and $\|\cdot\|$ represents the Euclidean norm.

Geographic bilateral trade costs are defined in Allen and Arkolakis (2014) as the assumption that the bilateral trade cost function is such that for all $n, i \in S$, $d(n,i) = f(t(n,i))$, for some monotonically increasing function $f : \mathbb{R}_+ \to [1, \infty)$ with $f(0) = 1$. Under this assumption, there exists a unique mapping from the instantaneous trade cost function τ to the bilateral trade costs d. Two implications of this assumption are that geographic bilateral trade costs are symmetric ($T(n,i) = T(i,n)$), and that nearby locations face similar trade costs to all other destinations, because the topography of the surface is smooth.

The solution to the problem (42) satisfies an eikonal partial differential equation whose solution can be characterized using the Fast Marching Method (FMM) of Sethian (1996). Intuitively, starting at any initial point $i \in S$, the FMM constructs iso-trade cost contours around that point. As the instantaneous trade costs are positive everywhere, bilateral trade costs always increase as one "marches" outward from any iso-cost contour. As a result, contours further from each point can be constructed using only the immediately previous contour, thereby bringing substantial gains in computational efficiency. This FMM can be interpreted as a generalization of Dijkstra's algorithm to continuous space: bilateral trade costs can be determined by approximating a surface with a grid (i.e. a graph of nodes and arcs) and taking the appropriate weighted average over different paths along the grid (see Tsitsiklis 1995). Nevertheless, this discretization will be subject to a digitization bias, because any chosen grid necessarily restricts the possible directions of travel (see Mitchell and Keirsey 1984).

Empirical research on trade and geography now commonly uses rich data on the transport network and either the Dijikstra algorithm (for applications in discrete space) or the Fast Marching algorithm (for applications in continuous space) to compute bilateral trade costs between locations. In principle, both methods can accommodate asymmetries in bilateral trade costs ($T(n,i) \neq T(i,n)$). Depending on the spatial scale of the application, one consideration that can be important in practice is that some modes of transports can only be joined at particular points in geographic space

(stations for railways or ramps for limited-access highways). Additionally, there can be costs incurred in changing modes of transport (the costs loading goods on or off railway cars or the costs of waiting for a bus or a train), which can be consequential for least-cost paths and bilateral trade costs. For example, in an influential analysis of the impact of the construction of the U.S. railroad network during the 19th century on the value of agricultural land, Donaldson and Hornbeck (2016) allows for transshipment costs of 50 cents per ton whenever transferring goods to/from a railroad car, river boat, canal barge, or ocean liner.

In most existing research on trade and geography, the transport network itself is taken as exogenously given, whereas in reality investments in transport infrastructure are likely to be endogenous to economic incentives. A small number of studies have creatively exploited natural experiments from history to examine the endogenous response of trade routes. Pascali (2017) finds that the invention of the steamship had a powerful effects on bilateral trade and development by weakening the dependence of trade routes on wind patterns. Feyrer (2009) uses the closure of the Suez Canal from 1967-75 following the outbreak of conflict in the Middle East as an exogenous shock to trading distances to estimate the relationships between trade and distance and income and trade.

Another interesting area for further research is microfounding bilateral trade costs and tracing the general equilibrium implications of these microfoundations. Brancaccio, Kaloupsidi, and Papageorgiou (2020) develop an explicit model of the transport sector, in which ships and exporters match with one another, and embed this specification of the matching process in a general equilibrium model of trade. The presence of matching frictions attenuates differences in comparative advantage, introduces externalities such that trade costs between any pair of countries depends on economic activity in their neighbors, and shapes the way in which trade shocks propagate through the trading network. Other research has explored the implications of containerization and deep-port technologies for the spatial distribution of economic activity, including Bernhofen, El-Sahli, and Kneller (2016), Brooks, Gendron-Carrier, and Rua (2019) and Ducruet, Juhász, Nagy, and Steinwender (2020)

To develop an approach for evaluating the welfare effects of transport infrastructure improvements, Allen and Arkolakis (2017) embed a specification of endogenous route choice in the spatial general equilibrium model of Allen and Arkolakis (2014). Individual traders experience extreme-value distributed idiosyncratic shocks to trade costs along each route and choose the least-cost route taking into account these idiosyncratic shocks. A key implication of this framework is that the welfare effects of a small improvement in a transport link is equal to the percentage cost savings achieved multiplied by the initial value of trade along that link. Although this result is derived for particular functional forms, this implication is closely related to the celebrated result of Hulten (1978) that a sufficient statistic for the welfare effect of a small productivity shock in an efficient economy can be summarized by the appropriate Domar weight. Implementing this framework for the U.S. interstate highway network, Allen and Arkolakis (2017) find the highest ratios of benefit to costs for highway links on the North-East corridor close to New York. One area for further research is improving our understanding of the costs of these transport investments and their determinants. For the U.S. interstate highway network, these costs appear large relative to other countries and have risen substantially over time, as shown in Brooks and Liscow (2019) and Duranton, Nagpal, and Turner (2020).

In a fundamental contribution, Fajgelbaum and Schaal (2020) develop a framework for characterizing optimal transport networks in spatial equilibrium. This characterization is challenging, because the problem is high dimensional and can be non-convex. The paper shows that the problem of finding the optimal transport network can be transformed into the problem of finding the optimal flow in a network, which has been studied in the operations

research literature. The planner chooses the optimal amount to invest in each link in the transport network, where the trade costs for each link are assumed to be increasing in the volume of traffic on that link and decreasing the level of the investment for that link. This model of transport infrastructure investments is then embedded within the class of quantitative spatial models introduced above and used to evaluate the observed transport networks in a number of European countries. In counterfactuals comparing the optimal and observed transport networks, the average welfare gain is 2 percent and these welfare gains range from 0.1 to 7 percent. In the coming years, developing models of endogenous transport investments and evaluating the implications for the spatial equilibrium distribution of economic activity is likely to be an exciting area for further research.

4.6 Empirical Evidence on Market Access

Motivated by the microfoundations provided above, a substantial empirical literature examined the relationship between wages and market access. Using data on U.S. counties, Hanson (2005) structurally estimates the equation for nominal wages from the Helpman (1998) model. This estimation yields plausible estimates for the structural parameters of the model and the theory-based measure of market access from the model is found to have greater explanatory power than traditional *ad hoc* measures of market potential (such as inverse distance weighted GDP). Using data on Japanese regions and countries, Davis and Weinstein (1999, 2003) provide evidence in support of the home market effect prediction of new economic geography models that an increase in expenditure on a good leads to a more than proportionate increase in production of that good. Using data on regions of the European Union, and exploiting both cross-section and time-series variation, Breinlich (2006) and Head and Mayer (2006) provide further support for the empirical relationship between nominal wages and market access.[37]

While there is strong evidence of a clear association between wages and market access, a key challenge for the empirical literature has been to establish that this association is causal. In particular, it is difficult empirically to disentangle the effects of market access from other leading determinants of comparative economic development, such as locational fundamentals or institutions. For example, the prosperity of a group of neighboring regions could reflect good access to one another's markets, but it could equally well reflect common good institutions or common favorable natural endowments. To empirically disentangle market access from these other leading determinants of comparative economic development, one requires exogenous variation along at least one dimension. One approach is therefore to use instruments for market access, such as lagged population levels or growth rates. However, a challenge is that institutions and natural endowments are strongly persistent, raising questions about the identifying assumption that lagged population only affects economic activity solely through market access.[38]

An alternative approach is to use trade liberalizations as a source of variation in market access. In influential work, Hanson (1996, 1997) has used Mexico's trade liberalization of 1985 as a natural experiment that changes the relative market access of regions. Following liberalization, there is evidence of a re-orientation of economic activity within Mexico towards the U.S. border and a shift from domestic production to offshore assembly for foreign (largely U.S.) firms. Consistent with the predictions of new economic geography models, these changes in the location of production lead to a re-orientation of the strong regional wage gradient previously centered on Mexico City towards

[37] Although most empirical research on market access has focused on either wages or production structure, Redding and Schott (2003) and Dekle and Eaton (1999) consider human capital accumulation and land rents respectively.

[38] For other instrumental variables approaches using a range of identifying assumptions, see Redding and Venables (2004), Hanson (2005), Costinot, Donaldson, Kyle, and Williams (2019) and Bartelme (2020). More broadly, for evidence on the role of demand composition in influencing quality specialization, see Dingel (2017).

the U.S. border.[39] While evidence based on trade liberalizations has bolstered the case for a causal interpretation of the relationship between market access and wages, there remain potential concerns. In particular, a political economy literature models trade policy as itself an endogenous outcome that could be influenced by market access.

To provide further evidence of a causal role for market access, Redding and Sturm (2008) uses the division of Germany after the Second World War as a natural experiment that provides plausibly exogenous variation in market access. The division of Germany has a number of attractive features for isolating the role played by market access. First, in contrast to cross-country studies, there is no obvious variation in institutions across cities within West Germany. Second, there are no obvious changes in natural advantage, such as access to navigable rivers or coasts, climatic conditions or the disease environment. Third, the change in market access following German division is much larger than typically observed in other contexts and the effects can be observed over a long period of time. Fourth, the drawing of the border dividing Germany into East and West Germany was based on military considerations that are unlikely to be correlated with pre-division characteristics of cities.

The population mobility condition (35) implies that a reduction in relative market access in some locations leads to a population outflow to other locations until the price of floor space adjusts to restore real wage equalization. In line with these predictions, Redding and Sturm (2008) finds that the imposition of the East-West German border leads to a sharp decline in population growth of West German cities close to the East-West border relative to other West German cities. Over the forty-year period of division, the East-West border cities experience a decline in their annualized rate of population growth of 0.75 percentage points, implying a cumulative reduction in their relative size of around one third.[40] A variety of additional pieces of evidence are provided in support of a market-access-based explanation and against other potential explanations, such as differences in industrial structure, war-related disruption, fear of further armed conflict, and Western European integration.[41]

In the class of quantitative spatial models developed above, the treatment effect of division on border cities is shaped by two parameter combinations that capture (a) the strength of agglomeration and dispersion forces and (b) the elasticity of trade with respect to distance. Redding and Sturm (2008) undertake a quantitative analysis of the model and show that for plausible values of these parameter combinations, the model can account quantitatively for both the average treatment effect of division and the larger treatment effect of division for small cities than for large cities. In the model, smaller cities experience this larger treatment effect, because they have smaller own markets, and hence are more dependent on markets in other cities.

Combining a general equilibrium trade model with archival data from colonial India, Donaldson (2018) evaluates the impact of India's vast railroad network. The empirical analysis is structured around an extension of Eaton and Kortum (2002) to incorporate multiple agricultural commodities, which shares many features with the class of quantitative spatial models developed above. This model delivers four key theoretical predictions that are taken to the data. First, for goods that are traded between regions, price differences between those regions can be used to measure bilateral trade costs. Second, the model yields a gravity equation for bilateral trade flows that can be used to estimate

[39] Other studies using trade liberalization as a source of variation in market access include Overman and Winters (2006) for the United Kingdom, Tirado, Paluzie, and Pons (2002) for early-twentieth century Spain, and Wolf (2007) for early-twentieth century Poland.

[40] Using detailed data on whether West German municipalities qualified for the Zonenrandgebiet (ZRG) place-based policy, Ehrlich and Seidel (2018) find even larger effects of market access at the municipality level after conditioning on qualification for this policy.

[41] Using the opening of Central and Eastern European markets after the fall of the Iron Curtain in 1990, Brülhart, Carrère, and Trionfetti (2012) find substantial increases in both wages and employment for Austrian municipalities within 50 kilometers of the former Iron Curtain. Using the economic separation of Japan and Korea after the Second World War and implementing the same empirical specification as in Redding and Sturm (2008), Nakajima (2008) finds a similar pattern of market access effects.

the response of trade flows to trade costs. Third, railroads increase real income levels, as measured by the real value of land income per unit area. Fourth, as in the theoretical framework developed above, each location's trade share with itself is a sufficient statistic for welfare. Consistent with these predictions of the model, there is a strong and statistically significant estimated effect of railroads on real income levels, but this effect becomes statistically insignificant after controlling for the model's sufficient statistic of a region's own trade share. These results provide evidence that the estimated effects of railroads are capturing the goods trade mechanism emphasized in the model.

Using data for the U.S. during the 19th century, Donaldson and Hornbeck (2016) investigate the impact of the expansion of the railroad network on the agricultural sector. In contrast to the reduced-form studies discussed above, the analysis captures not only the direct effect of a railroad connection but also the indirect effects through changes in economic activity in neighboring locations. In particular, motivated by the class of quantitative spatial models discussed above, the analysis uses measures of market access that capture each location's access to surrounding economic activity. In constructing market access, measures of bilateral trade costs are used that take into account the transport network of railroads and waterways, and compute lowest-cost county-to-county freight routes. County agricultural land values are found to increase substantially with increases in county market access, as the railroad network expanded from 1870 to 1890. Removing all railroads in 1890 is estimated to decrease the total value of U.S. agricultural land by 60 percent, with limited potential for mitigating these losses through feasible extensions to the canal network or improvements to country roads.[42] This reduction in agricultural land values generates annual income losses equal to 3.22 percent of gross national product (GNP), which is somewhat larger than the estimate of 2.7 percent based on a social savings approach in Fogel (1964).

Finally, using data on 39 countries from Sub-Saharan Africa over 50 years from 1960-2010, Jedwab and Storeygard (2020) provides evidence in support of another of the key implications of measures of market access, namely the heterogeneous effects of transport infrastructure investments.

4.7 Multiple Equilibria and Path Dependence

As discussed above, recent quantitative research on economic geography has focused on region of the parameter space for which there exists a unique equilibrium, thereby ensuring that counterfactuals for transport improvements and other interventions have determinate predictions for the spatial distribution of economic activity. In contrast, a key implication of early theoretical research on economic geography was the potential for multiple equilibria. In the presence of such multiple equilibria, temporary shocks can have permanent effects on the spatial distribution of economic activity, if they shift the economy between these multiple equilibria.

Motivated by this theoretical prediction, a number of empirical studies have sought to provide empirical evidence on the extent to which economic activity is path dependent in the sense that temporary shocks can have permanent effects on the spatial distribution of economic activity. In two path-breaking papers, Davis and Weinstein (2002, 2008) used Japanese war-time bombing as such an exogenous temporary shock, and found little evidence of path dependence for either the distribution of population as a whole or employment in individual industries.[43] Subsequent studies have provided evidence of path dependence using a variety of sources of quasi-experimental variation. Dell (2010) provides

[42] For evidence on the role of the expansion of the railroad network in promoting U.S. manufacturing activity through improved market access, see Hornbeck and Rotemberg (2021).

[43] Other research using war-time bombing as an exogenous shock includes Bosker, Brakman, Garretsen, and Schramm (2007), Brakman, Garretsen, and Schramm (2004), Miguel and Roland (2011), and Dell and Querubin (2018).

compelling evidence of persistent effects of the mining *mita*, which was an extensive forced mining labor system that was practiced in the areas of present-day Peru and Bolivia between 1573 and 1812. Redding, Sturm, and Wolf (2011) find path dependence in the location of Germany's air hub using the natural experiment of Germany's division and reunification.[44] Bleakley and Lin (2012) provides strong evidence that the temporary historic advantage of U.S. portage sites has had permanent effects on the location of economic activity. Hornbeck and Keniston (2017) find long-lived effects of the Boston fire on plot-level land values through the potential for large-scale rebuilding. Michaels and Rauch (2018) provide evidence of path dependence using data on the location of Roman cities.

Although these studies provide several convincing examples of path dependence, there remain many open questions of interpretation and areas for further research. In particular, empirical evidence of path dependence does not necessarily imply multiple equilibria. For example, historical advantages could lead to initial investments in durable infrastructure in a location. Once these initial investments have been incurred, it may be profitable to maintain them, and to continue to concentrate economic activity in that location, even after the original historical advantages have become obsolete. More generally, further theory and evidence is needed clarifying the conditions under which we either should or should not expect to observe path dependence, which relates to the debate in the economic growth literature regarding the role of history and expectations in selecting equilibria, as in Krugman (1991b) and Matsuyama (1991). In an important contribution in this area, Allen and Donaldson (2020) develop a dynamic economic geography model, in which there is either a unique steady-state equilibrium regardless of initial conditions, or there are multiple steady-state equilibria, with initial conditions determining which of these equilibria is selected.

4.8 Modeling the Geographic Incidence of Trade Shocks

In the baseline class of quantitative spatial models developed above, workers are perfectly mobile, and hence real income is equalized across all populated locations. Therefore, there is a single national labor market, and although an international trade shock can have heterogeneous effects on employment, nominal wages and price indexes across locations, it has exactly the same effect on welfare across all locations. Motivated by the reduced-form evidence on the distributional consequences of international trade across local labor markets, researchers have sought to develop quantitative spatial models that are consistent with these reduced-form moments and allow for the possibility that trade shocks can have uneven effects on welfare across local labor markets.

A particularly influential paper is Caliendo, Dvorkin, and Parro (2019), which incorporates a dynamic discrete choice model of household location decisions into a quantitative spatial model with multiple sectors and input-output linkages.[45] In this framework, the world consists of N locations and J sectors, where there is a competitive labor market in each region-sector combination. We use n or i to index locations and j or k to reference sectors. The timing of decisions is as follows. At the beginning of each period, each household starts out in a region-sector, where sector zero in each region corresponds to unemployment. Households observe the economic conditions in all labor markets and the realizations of idiosyncratic shocks to mobility costs. If they begin the period in a labor market, they work and earn the market wage. If they are unemployed in a region, they get home production. Then, both employed and unemployed households have the option to relocate.

[44] For structural estimations of models of the location of particular economic activities, see Holmes (2005) for headquarter location choices and Holmes (2011) for the expansion of Walmart's distribution and retail network.

[45] There is a rich tradition in international trade of using dynamic discrete choice models to analyze the distributional consequences of trade, including Artuç, Chaudhuri, and McLaren (2010), Dix-Carneiro (2014) and Traiberman (2019).

Under these assumptions, the lifetime utility of a household in region-sector nj at time t (v_t^{nj}) equals current instantaneous utility plus the expected continuation payoff from choosing a region-sector to maximize next period's lifetime utility:

$$v_t^{nj} = U\left(C_t^{nj}\right) + \max_{\{i,k\}_{i=1,k=0}^{N,J}} \left\{\beta E\left[v_{t+1}^{ik}\right] - \tau^{nj,ik} + \nu\epsilon_t^{ik}\right\} \tag{43}$$

$$U\left(C_t^{nj}\right) = \begin{cases} b^n & \text{if } j = 0 \\ w_t^{nj}/P_t^n & \text{otherwise} \end{cases}. \tag{44}$$

where b^n is home production; w_t^{nj} is the wage in sector j in location n at time t; P_t^n is the consumption goods price index in location n at time t; β is the discount factor; $\tau^{nj,ik}$ are common mobility costs of relocating from region-sector nj to region-sector ik; ϵ_t^{ik} is an idiosyncratic shock to mobility costs for region-sector ik; and ν is a parameter that scales the variance of this idiosyncratic shock.

Under the assumption that the idiosyncratic shock to mobility costs (ϵ_t^{ik}) is independently and identically distributed over time and distributed Type-I extreme value with zero mean, the expected lifetime utility of a representative household in region-sector nj at time t ($V_t^{nj} = E\left[v_t^{nj}\right]$) can be written as:

$$V_t^{nj} = U\left(C_t^{nj}\right) + \nu \log\left(\sum_{m=1}^{N}\sum_{h=0}^{J} e^{\left(\beta V_{t+1}^{mh} - \tau^{nj,mh}\right)^{1/\nu}}\right). \tag{45}$$

and the probability that a household chooses to relocate from region-sector nj to region-sector ik is:

$$\mu_t^{nj,ik} = \frac{e^{\left(\beta V_{t+1}^{ik} - \tau^{nj,ik}\right)^{1/\nu}}}{\sum_{m=1}^{N}\sum_{h=0}^{J} e^{\left(\beta V_{t+1}^{mh} - \tau^{nj,mh}\right)^{1/\nu}}}. \tag{46}$$

Using these relocation probabilities, the distribution of labor across region-sectors evolves over time as follows:

$$L_{t+1}^{nj} = \sum_{i=1}^{N}\sum_{k=0}^{J} \mu_t^{ik,nj} L_t^{ik}. \tag{47}$$

For a given allocation of labor across region-sectors at time t, the static (or temporary) equilibrium takes the same form as in the quantitative international trade literature. In particular, Caliendo, Dvorkin, and Parro (2019) considers a multi-sector version of the Eaton and Kortum (2002) model with input-output linkages between sectors following Caliendo and Parro (2015), which determines the real wage w_t^{nj}/P_t^n in instantaneous utility (44).

Two key differences between this dynamic model and the static model of the distribution of economic activity developed in Section 3 are as follows. First, this dynamic model implies a period of gradual adjustment in response to an exogenous shock (such as the China trade shock), because households only gradually relocate when they receive favorable shocks to idiosyncratic mobility costs. Second, in this dynamic model, there is an option value in each region-sector that depends on the expected value of relocating in the future, where the welfare gains from trade depend on both current instantaneous utility and this option value.

One of the key contributions of Caliendo, Dvorkin, and Parro (2019) is to show that this dynamic discrete choice formulation permits a dynamic version of the exact-hat algebra counterfactuals introduced in Section 3.9 above. In particular, counterfactuals can be undertaken using the observed values of the endogenous variables in an initial equilibrium and values of the model's structural parameters, without having to solve for unobserved fundamentals for each location. Using this approach, the paper analyzes the distributional consequences of the rise in China's trade using data on 22 sectors, 38 countries and 50 U.S. states. With each sector-state combination corresponding

to a separate labor market, this yields more than one thousand labor markets. Consistent with the reduced-form empirical findings discussed in Section 2 above, those region-sectors more exposed to the China trade shock experience larger reductions in manufacturing employment. In contrast to the earlier difference-in-difference specifications, the model can now be used to evaluate aggregate effects and compute model-based objects such as welfare. Overall, the China trade shock is predicted to reduce aggregate employment in U.S. manufacturing by around 0.55 million, which corresponds to about 16 percent of the observed decline from 2000 to 2007. Although the China trade shock increases aggregate U.S. welfare by around 0.2 percent, there is substantial dispersion in these welfare effects across region-sectors because of the mobility frictions, with the predicted welfare changes ranging from -0.8-1 percent. Taking these quantitative results together with the reduced-form evidence from Section 2 provides a good example of how these two methodologies can complement one another in shedding light on important economic issues.

There remain many exciting areas for further research on dynamic models of trade and geography. One substantive issue for which mobility frictions are salient is the response of the economy to environmental change. Balboni (2019) combines a dynamic discrete choice model of worker locations decisions with a static model of economic geography to examine how coastal flooding affects the returns to transport infrastructure investments. Although road investments concentrated in coastal regions between 2000 and 2010 had positive returns, they would have been outperformed by allocations concentrated further inland even in the absence of sea level rise. Future inundation considerably reduces the welfare effects of existing road investments. Under a central sea level rise scenario, a more foresighted road allocation that avoids the most vulnerable regions would have achieved 72 percent higher welfare.

Although migration decisions are one source of dynamics, there are several other sources of dynamics that are likely to be quantitatively important, including physical and human capital accumulation and innovation. Two path-breaking contributions to the geography of innovation are Desmet and Rossi-Hansberg (2014) and Desmet, Nagy, and Rossi-Hansberg (2018), which show how to tractably embed endogenous innovation decisions in high-dimensional spatial models. A key insight from these frameworks is that some of the largest effects of geography on economic development and welfare may occur through the rate of growth rather than the level of economic activity. Using data for the whole world economy at a 1 degree $\times$ 1 degree level of geographic resolution, Desmet, Nagy, and Rossi-Hansberg (2018) finds that fully liberalizing migration restrictions between countries would increase the present discounted value of utility about threefold, with much of this effect occurring through innovation and growth. Nagy (2020) shows that a model of trade and geography incorporating innovation is quantitatively successful in accounting for patterns of city formation in the United States prior to the Civil War. Desmet, Kopp, Kulp, Nagy, Oppenheimer, Rossi-Hansberg, and Strauss (2020) show that dynamics from both migration and innovation are quantitatively important in shaping the impact of coastal flooding on the global economy.

Another exciting area for further research is applications of sufficient statistics methodologies to economic geography. Galle, Yi, and Rodriguez-Clare (2018) combines a standard quantitative model of international trade with a Roy model of the sorting of workers in each local labor market across sectors. The paper shows that the welfare gains from trade for workers from each local labor market can be expressed in terms of the domestic trade share for the aggregate economy in each sector and a measure of the degree of specialization of workers from each local labor market across sectors. Adão, Arkolakis, and Esposito (2019) uses a sufficient statistics approach to extend shift-share empirical specifications to incorporate indirect or general equilibrium effects that arise in spatial models from interactions between locations. These indirect effects are shown to play an important role in shaping the overall impact of

the China shock on each location. Kleinman, Liu, and Redding (2021) develops sufficient statistics for dynamic spatial models, in which the distribution of economic activity responds gradually to shocks, because of migration frictions for the mobile factor (labor) and the gradual accumulation of the immobile factor (capital structures). Closed-form solutions are derived for the response of both the steady-state equilibrium and the full transition path to shocks to productivity, amenities, trade costs and migration costs, in terms of the structural parameters of the model and four observable matrices for income shares, expenditure shares, inmigration shares and outmigration shares.

Finally, one striking feature of the reduced-form findings for the China shock in Section 2 is the importance of unemployment benefits, disability and income assistance as adjustment margins to the shock. Although one can interpret these adjustment margins as home production in neoclassical models, these findings raise the question of the potential relevance of other labor market frictions in explaining these responses, as explored in Kim and Vogel (2020) and Rodriguez-Clare, Ulate, and Vasquez (2018).

5 Modeling Economic Activity Within Cities

We now turn to quantitative spatial models of the organization of economic activity within cities, where commuting becomes relevant.[46] We consider a city that is embedded within a wider economy. The city consists of a discrete set of locations (N). The economy as a whole is populated by an endogenous measure of workers, who are geographically mobile, and choose whether to live in the city or the wider economy. Population mobility implies that the expected utility from living in the city equals the reservation level of utility in the wider economy $\bar{U}$.[47] If a worker chooses the city, she observes idiosyncratic preference draws for each possible pair of residence and workplace, and choose a residence n and a workplace i to maximize her utility. With a continuous measure of workers, the law of large numbers applies, and the expected values of variables for each residence and workplace pair are equal to their realized values.[48] Locations can differ from one another in terms of their attractiveness for both production and residence, as determined by productivity, amenities, the supply of floor space, and transport connections.

5.1 Workplace-Residence Choices

Worker preferences are defined over consumption goods and residential floor space. We assume that these preferences take the Cobb-Douglas form, such that the indirect utility for a worker ω residing in n and working in i is:[49]

$$U_{ni}(\omega) = \frac{B_n b_{ni}(\omega) w_i}{\kappa_{ni} P_n^{\alpha} Q_n^{1-\alpha}}, \qquad 0 < \alpha < 1, \tag{48}$$

where we suppress the time subscript, except where important; P_n is the price index for consumption goods, which may include both tradeable and non-tradeable consumption goods; Q_n is the price of floor space; w_i is the wage; κ_{ni} is an iceberg commuting cost; B_n captures residential amenities that are common across all workers and could be endogenous to the surrounding concentration of economic activity through agglomeration forces; and $b_{ni}(\omega)$ is an idiosyncratic amenity draw that captures all the idiosyncratic factors that can cause an individual to live and work in particular locations within the city.

[46] For a recent review of research on cities in the developing world, see Bryan, Glaeser, and Tsivanidis (2020).

[47] Although we take this reservation level of utility in the wider economy as given here, this model of a single city can be embedded in model system of cities that determines this common reservation level of utility across all cities, as in Monte, Redding, and Rossi-Hansberg (2018).

[48] For a relaxation of the assumption of a continuous measure of workers and an exploration of granularity, see Dingel and Tintelnot (2020)

[49] For empirical evidence using U.S. data in support of the constant housing expenditure share implied by the Cobb-Douglas functional form, see Davis and Ortalo-Magné (2011).

This specification of indirect utility (48) is consistent with an entire class of quantitative urban models, as shown in Heblich, Redding, and Sturm (2020). This class of models includes the canonical urban model with a single tradeable final good as in Lucas and Rossi-Hansberg (2002) and Ahlfeldt, Redding, Sturm, and Wolf (2015), an Armington model in which goods are differentiated by origin as in Armington (1969), Allen and Arkolakis (2014) and Allen, Arkolakis, and Li (2017), a Ricardian model in which regions specialize in different goods as a result of technology differences as in Eaton and Kortum (2002) and Redding (2016), and a new economic geography model in which an endogenous measure of firms in each location supplies horizontally differentiated varieties as in Helpman (1998), Redding and Sturm (2008) and Monte, Redding, and Rossi-Hansberg (2018). Although idiosyncratic heterogeneity is modeled in terms of worker preferences ($b_{ni}(\omega)$) in equation (48), there is a closely-related formulation in terms of heterogeneity in worker productivity (effective units of labor). Both formulations are isomorphic in terms of the choice probabilities, but differ in that the specification using effective units of labor interprets income in the data as wages times average effective units of labor. Similarly, although commuting costs (κ_{ni}) are modeled in terms of utility in equation (48), they enter the indirect utility multiplicatively with the wage, which implies that there is also a closely-related formulation in terms of the opportunity cost of time spent commuting.

We assume that idiosyncratic amenities ($b_{ni}(\omega)$) are drawn from an independent extreme value (Fréchet) distribution for each residence-workplace pair and worker:

$$G(b) = e^{-b^{-\epsilon}}, \qquad \epsilon > 1, \tag{49}$$

where we normalize the Fréchet scale parameter in equation (49) to one, because it enters the worker choice probabilities isomorphically to common amenities B_n in equation (48); the smaller the Fréchet shape parameter ϵ, the greater the heterogeneity in idiosyncratic amenities, and the less sensitive are worker location decisions to economic variables.[50]

Using standard results for extreme value distributions, the probability that a worker chooses to reside in n and work in i is given by:

$$\lambda_{ni} = \frac{L_{ni}}{L_N} = \frac{\left(B_n w_i\right)^{\epsilon} \left(\kappa_{ni} P_n^{\alpha} Q_n^{1-\alpha}\right)^{-\epsilon}}{\sum_{k \in N} \sum_{\ell \in N} \left(B_k w_\ell\right)^{\epsilon} \left(\kappa_{k\ell} P_k^{\alpha} Q_k^{1-\alpha}\right)^{-\epsilon}}, \tag{50}$$

where L_{ni} is the measure of commuters from n to i and L_N is the measure of workers that choose the city.

A first key implication of the extreme value specification for idiosyncratic amenities is that bilateral commuting flows in equation (50) satisfy a gravity equation. Therefore, the probability of commuting between residence n and workplace i depends on the characteristics of that residence n, the attributes of that workplace i and bilateral commuting costs and amenities ("bilateral resistance"). Furthermore, this probability also depends on the characteristics of all residences k, all workplaces ℓ and all bilateral commuting costs ("multilateral resistance"). A large reduced-form literature in urban economics provides empirical evidence that the gravity equation provides a good approximation to commuting flows, as reviewed in Fortheringham and O'Kelly (1989) and McDonald and McMillen (2010).

Summing across workplaces i, we obtain the probability that a worker lives in residence n ($\lambda_n^R = R_n / L_N$). Summing across residences n, we obtain the probability that a worker is employed in workplace i ($\lambda_i^L = L_i / L_N$):

$$\lambda_n^R = \frac{\sum_{i \in N} \left(B_n w_i\right)^{\epsilon} \left(\kappa_{ni} P_n^{\alpha} Q_n^{1-\alpha}\right)^{-\epsilon}}{\sum_{k \in N} \sum_{\ell \in N} \left(B_k w_\ell\right)^{\epsilon} \left(\kappa_{k\ell} P_k^{\alpha} Q_k^{1-\alpha}\right)^{-\epsilon}}, \qquad \lambda_i^L = \frac{\sum_{n \in N} \left(B_n w_i\right)^{\epsilon} \left(\kappa_{ni} P_n^{\alpha} Q_n^{1-\alpha}\right)^{-\epsilon}}{\sum_{k \in N} \sum_{\ell \in N} \left(B_k w_\ell\right)^{\epsilon} \left(\kappa_{k\ell} P_k^{\alpha} Q_k^{1-\alpha}\right)^{-\epsilon}}, \tag{51}$$

[50] Modeling idiosyncratic preferences using the extreme value distribution has a long tradition in transportation economics, dating back to McFadden (1974). A related literature models workers' migration decisions using extreme value distributed preferences, as in Grogger and Hanson (2011), Kennan and Walker (2011), Bryan and Morten (2019) and Morten and Oliveira (2018).

where R_n denotes employment by residence in location n and L_i denotes employment by workplace in location i.

A second key implication of the extreme value specification is that expected utility is equalized across all pairs of residences and workplaces within the city and is equal to the reservation level of utility in the wider economy:

$$\bar{U} = \delta \left[\sum_{k \in N} \sum_{\ell \in N} \left(B_k w_\ell \right)^\epsilon \left(\kappa_{k\ell} P_k^\alpha Q_k^{1-\alpha} \right)^{-\epsilon} \right]^{\frac{1}{\epsilon}}, \tag{52}$$

where the expectation is taken over the distribution for idiosyncratic amenities; $\delta \equiv \Gamma((\epsilon - 1)/\epsilon)$; and $\Gamma(\cdot)$ is the Gamma function.

The intuition for this second result is that bilateral commutes with attractive economic characteristics (high workplace wages and low residence cost of living) attract additional commuters with lower idiosyncratic amenities, until expected utility (taking into account idiosyncratic amenities) is the same across all bilateral commutes and equal to the reservation utility. A closely related implication is that workplaces and residences face upward-sloping supply functions in real wages for workers and residents respectively (as captured in the choice probabilities (50)). To obtain additional workers, a location must pay higher wages to attract workers with lower realizations for idiosyncratic amenities for that workplace. Similarly, to acquire additional residents, a location must offer a lower cost of living to entice residents with lower realizations for idiosyncratic amenities for that residence.

In this specification, workers are *ex ante* homogenous, and *ex post* heterogeneous in terms of their idiosyncratic preference draws. An interesting generalization is to introduce multiple groups of *ex ante* heterogeneous workers, as for example in Redding and Sturm (2016), Tsivanidis (2018), Almagro and Domínguez-Iino (2019), and Couture, Gaubert, and Handbury (2020). If these different types of workers have different dispersion parameters for idiosyncratic preferences, or if preferences are non-homothetic, there will be systematic spatial sorting of each group of workers across neighborhoods with different characteristics within the city.

5.2 First and Second-Nature Geography

We now combine this specification of workplace-residence choices with assumptions on production structure and the supply of floor space and show how this model of the internal structure of cities can be used to quantify the role of first and second-nature geography. We assume that the researcher observes employment by residence (R_n), employment by workplace (L_i), the price of floor space (Q_n), and geographical land area (K_n), and show how the model can be inverted to recover the unobserved locational fundamentals that exactly rationalize the observed data as an equilibrium outcome.

The remainder of our analysis proceeds in three steps. First, we use the observed data and the model's bilateral commuting predictions to recover unobserved wages (w_i). Second, we use assumptions on production structure and the observed data to recover unobserved productivity (A_i). Third, we use the observed data and the model's predictions for residential choice probabilities to recover unobserved amenities (B_i). Fourth, we use assumptions on the supply of floor space (H_n) to recover a measure of the density of development for each location (φ_i).

Starting with the first step for wages, using the commuting probability (λ_{ni}) in equation (50) and the residence probability (λ_n^R) in equation (51), the conditional probability that a worker commutes to workplace i conditional on living in residence n is:

$$\lambda_{ni|n}^R = \frac{\lambda_{ni}}{\lambda_n^R} = \frac{\left(w_i / \kappa_{ni} \right)^\epsilon}{\sum_{\ell \in N} \left(w_\ell / \kappa_{n\ell} \right)^\epsilon}, \tag{53}$$

Using this conditional commuting probability in the commuter market clearing condition that equates employment in each workplace with the number of residence commuting to that workplace, we have:

$$L_i = \sum_{n \in N} \frac{(w_i/\kappa_{ni})^{\epsilon}}{\sum_{\ell \in N} (w_\ell/\kappa_{n\ell})^{\epsilon}} R_n, \tag{54}$$

Given the observed data on employment (L_i) and residents (R_n) and a parameterization of commuting costs (κ_{ni}), equation (54) provides a system of N equations that can be solved for the unique wage in each location (w_n) that satisfies this commuter market clearing condition. A natural baseline parameterization of commuting costs is to assume that they are an exponential function of travel times (τ_{ni}), such that $\kappa_{ni} = e^{\kappa \tau_{ni}}$. Although this specification typically provides a good approximation to observed commuting data (see for example Ahlfeldt, Redding, Sturm, and Wolf 2015), it could be further enriched to introduce congestion, such that bilateral commuting costs for a residence-workplace pair depend on the flow of commuters for that pair.

Moving to the second step for productivity, we assume a single final good that is costlessly traded within the city and chosen as the numeraire ($P_n = 1$ for all n). Under these assumptions, unobserved productivity (A_n) can be recovered from the zero-profit condition that equates price and unit cost, together with the solutions for wages from above (w_n) and the observed price of floor space (Q_n):

$$A_i = w_i^{\alpha} Q_i^{1-\alpha}, \qquad 0 < \alpha < 1, \tag{55}$$

where we assume for simplicity no arbitrage between residential and commercial use of floor space, such that there is a single price for floor space in each location. This zero-profit condition has an intuitive interpretation: Higher wages (w_i) and higher prices of floor space (Q_i) in a location imply that productivity (A_i) must be higher, in order to satisfy the requirement that profits are zero if the final good is produced. In principle, productivity also could be endogenous to the surrounding concentration of economic activity through agglomeration forces.

Continuing to the third step for amenities, using expected utility ($\bar{U}$) from equation (52) and defining a measure of residents commuting market access (RMA_n), the residential choice probabilities (51) can be re-written as follows:

$$\lambda_n^R = \left(\frac{B_n}{\bar{U}/\delta}\right)^{\epsilon} Q_n^{-\epsilon(1-\alpha)} CMA_n^{\epsilon}, \qquad RMA_n = \left[\sum_{\ell \in N} (w_\ell/\kappa_{n\ell})^{\epsilon}\right]^{\frac{1}{\epsilon}}. \tag{56}$$

Given a choice of units in which to measure amenities (a normalization for $\bar{U}$), the observed data on residential choice probabilities (λ_n^R) and the price of floor space (Q_n), our parameterization of commuting costs (κ_{ni}), and our solutions for wages from above (w_n), we can determine unique values for residential amenities in each location (B_n) from these residential choice probabilities.[51]

Turning to the fourth and final step, we can recover the density of development (φ_n) from the market clearing condition that equates the demand and supply for floor space:

$$H_n^R + H_n^L = \varphi_n K_n, \tag{57}$$

where K_n is geographical land area; φ_n is the density of development (the ratio of floor space to geographical land area); and the demands for residential (H_n^R) and commercial (H_n^L) floor space use can be obtained from workers' and firms' first-order conditions using the observed data and the solutions from the previous steps above.

[51] The commuting gravity equation (50) can be re-written in terms of this measure of residents commuting market access (RMA_n) from equation (56) and an analogous measure of workers commuting market access, which play an analogous role to the consumer and firm market access measures in the model of economic activity between cities and regions in Section 4.3 above.

In the case in which productivity (A_n), amenities (B_n) and the density of development (φ_n) are exogenous, there exists a unique spatial equilibrium distribution of economic activity, as shown in Ahlfeldt, Redding, Sturm, and Wolf (2015). In contrast, if productivity and amenities are endogenous because of agglomeration forces, whether the model has a unique equilibrium or multiple equilibria depends on the strength of these agglomeration forces relative to the exogenous differences in characteristics across locations. Nevertheless, this quantitative urban model is invertible, in the sense that unique values of productivity (A_n), amenities (B_n) and the density of development (φ_n) can be recovered conditional on the observed equilibrium in the data, regardless of whether or not another equilibrium could have occurred. Intuitively, the equilibrium conditions in the model, such as utility maximization, population mobility, profit maximization and zero profits, contain enough information given the observed data to uniquely pin down these unobserved location characteristics.

A related property of this quantitative urban model is that it is recursive, in the sense that the overall values of productivity (A_n) and amenities (B_n) can be determined, without having to specify the extent to which these variables are exogenous or endogenous to agglomeration forces, and without having to specify the functional form of these agglomeration forces. Having recovered overall productivity (A_n) and amenities (B_n), we can undertake model-based decompositions to examine the relative importance of these variables and commuting market access in explaining the observed variation in internal city structure.

Exactly how the model is used to recover these unobserved location characteristics depends on both model assumptions and the data available to the researcher. From a data perspective, we assumed in this section that the researcher observed employment by workplace (L_i) and employment by residence (R_n), but not bilateral commuting flows. Therefore, we parameterized bilateral commuting costs (κ_{ni}) and solved for wages (w_n), which yields model predictions for bilateral commuting flows. In other applications, direct data on bilateral commuting flows may be available, as in Monte, Redding, and Rossi-Hansberg (2018), Heblich, Redding, and Sturm (2020) and Owens III, Rossi-Hansberg, and Sarte (2020). From a model perspective, we assumed in this section that there is a single tradeable final good and recovered productivity from the zero-profit condition. In contrast, in models in which goods are differentiated by origin or firm, productivity instead can be recovered from the market clearing condition that income in each location equals expenditure on the goods produced by that location.

5.3 Estimating Agglomeration Forces

One application of quantitative urban models is estimating the strength of agglomeration forces, which raises similar identification challenges as those for market access discussed above. Although high land prices and levels of economic activity in a group of neighboring locations are consistent with strong agglomeration forces, they are also consistent with shared amenities that make these locations attractive places to live (e.g., leafy streets and scenic views) or common natural advantages that make these locations attractive for production (e.g., access to natural water). To separate these two sets of determinants of levels of economic activity, one typically needs both some additional model structure and exogenous variation in the surrounding concentration of economic activity.

We begin by introducing this additional model structure, before discussing potential exogenous sources of variation in surrounding economic activity. We allow productivity (A_n) to depend on production fundamentals (a_n) and production externalities (Ω_n^L). Production fundamentals capture features of physical geography that make a location more or less productive independently of neighboring economic activity (e.g. access to natural water). Production

externalities capture productivity benefits from the surrounding employment density and are typically modeled in urban economics as depending on the travel time weighted sum of surrounding employment density:

$$A_i = a_i \left(\Omega_i^L\right)^{\eta^L}, \qquad \Omega_j^L \equiv \sum_{k=1}^{N} e^{-\rho^L \tau_{ik}} \left(\frac{L_k}{K_k}\right), \tag{58}$$

where η^L controls the relative importance of production externalities; ρ^L determines their rate of spatial decay; and recall that K_i is geographical land area.

Similarly, we allow residential amenities (B_n) to depend on residential fundamentals (b_n) and residential externalities (Ω_n). Residential fundamentals capture features of physical geography that make a location a more or less attractive place to live independently of neighboring economic activity (e.g. green areas). Residential externalities capture the effects of the surrounding density of residents and are modeled symmetrically to production externalities:

$$B_i = b_i \left(\Omega_i^R\right)^{\eta^R}, \qquad \Omega_i^R \equiv \sum_{k=1}^{N} e^{-\rho^R \tau_{ik}} \left(\frac{R_k}{K_k}\right), \tag{59}$$

where η^R controls the relative importance of residential externalities; and ρ^R determines their rate of spatial decay.

In separating agglomeration forces and locational fundamentals, we again use invertibility properties of this class of quantitative urban models. Given the values of productivity (A_n) and amenities (B_n) that were recovered from the equilibrium conditions of the model and the observed data in the previous section, and given values for the agglomeration parameters (η^L, ρ^L, η^R, ρ^R), we can solve for unique values of unobserved production fundamentals (a_n) and residential fundamentals (b_n). These unobserved fundamentals correspond to structural residuals that ensure that the model exactly rationalizes the observed data.

The agglomeration parameters (η^L, ρ^L, η^R, ρ^R) can be estimated using orthogonality conditions on these structural residuals. In Ahlfeldt, Redding, Sturm, and Wolf (2015), these orthgonality conditions use the exogenous change in the surrounding concentration of economic activity from the division of Berlin in the aftermath of the Second World War and its reunification following the fall of the Iron Curtain. In particular, these orthogonality conditions assume that the log changes in production (a_n) and residential (b_n) fundamentals are uncorrelated with indicator variables for grid cells for the distance of West Berlin city blocks from the pre-war Central Business District (CBD) just East of the Berlin Wall. This identifying assumption requires that the change in the organization of economic activity within Berlin is explained by the model's agglomeration and dispersion forces, rather than by systematic changes in locational fundamentals. As this approach conditions on the observed equilibrium in the data, and the model is invertible in the sense that there is a one-to-one mapping from the observed variables to the structural residuals, no assumptions are required for parameter estimation about whether the model has a single equilibrium or multiple equilibria.[52] All that is required for the estimation to consistently estimate the model parameters is that the structural residuals for the observed equilibrium satisfy the assumed orthogonality conditions.

Using these orthogonality conditions, Ahlfeldt, Redding, Sturm, and Wolf (2015) find evidence of agglomeration economies in both production and residential decisions. For production, the estimated elasticity of productivity with respect to the density of employment is 0.07, which is towards the high end of the 3-8 percent range stated in the survey by Rosenthal and Strange (2004), but less than the elasticities from some quasi-experimental studies (e.g. Greenstone, Hornbeck, and Moretti 2010 and Kline and Moretti 2014a).[53] Consistent with other research using spatially-

[52] For an analysis of the existence and uniqueness of the general equilibrium in this class of models allowing agglomeration forces to spill over across locations, see Allen, Arkolakis, and Li (2020).

[53] For meta-analyses of empirical estimates of agglomeration forces, see Melo, Graham, and Noland (2009) and Ahlfeldt and Pietrostefani (2019).

disaggregated data within cities, such as Arzaghi and Henderson (2008), these productivity externalities are highly localized, such that they decline to close to zero by 10 minutes of travel time. For residential decisions, the estimated elasticity of amenities with respect to the density of residents is larger at around 0.15, which is in line with a growing body of research that emphasizes endogenous amenities, including Glaeser, Kolko, and Saiz (2001) and Diamond (2016). Again these residential externalities are highly localized, and decline to close to zero by 10 minutes of travel time. This finding of localized residential externalities is consistent with the results of Rossi-Hansberg, Sarte, and Owens (2010), which finds that housing externalities fall by approximately half every 1,000 feet.

An important area for further research is discriminating between alternative possible microfoundations for these specifications of agglomeration forces in production and residence. Many existing models of agglomeration were inspired by thinking about the concentration of manufacturing industries in cities in the late-19th century. Yet employment in cities today is overwhelmingly concentrated in service sectors. Furthermore, even within sectors, the types of economic activities performed in cities have changed dramatically over time, as shown in Michaels, Rauch, and Redding (2019) using the verbs from occupational descriptions to capture the tasks performed by workers in those occupations. Whereas the tasks most concentrated in cities in 1880 involved the manipulation of the physical world, such as Thread and Sew, those most concentrated in cities in 2000 involve human interaction, such as Advise and Confer.[54] Given these large-scale changes in the types of economic activities performed in urban versus rural areas over time, it is at least reasonable to ponder whether the nature and scope of agglomeration economies could have changed over time. Consistent with this idea, Autor (2019) finds substantial changes in the urban wage premium for workers with different levels of skills over time. At the beginning of the sample period in the 1970s, average wages were sharply increasing in population density for both low-skill workers (high-school or less) and high-skill workers (some college or greater). By the end of the sample period in 2015, this wage premium to population density had increased for high-skill workers but almost disappeared for low-skill workers.

Finally, most theories of agglomeration focus on production, yet the growing number of empirical findings of endogenous amenities call out for further research on the microfoundations for agglomeration forces in residential decisions. One important mechanism for endogenous amenities is agglomeration in the consumption of local services. Using barcode data for the consumer goods sector in the U.S., Handbury and Weinstein (2015) find substantial differences in the range of product varieties available across cities of different sizes. Using data on U.S. core-based statistical areas (CBSAs), Couture, Gaubert, Handbury, and Hurst (2018) provide evidence that non-homotheticities in consumption play an important role in gentrification and the spatial sorting of workers with different levels of income between the central city and the suburbs. Using restaurant location data, Couture (2016) and Davis, Dingel, Monras, and Morales (2019) provide evidence that both spatial and other frictions play an important role in determining demand for these non-traded services. Using Japanese smartphone global positioning system (GPS) data, Miyauchi, Nakajima, and Redding (2021) show that non-commuting trips within urban areas are both frequent and closely related to the availability of local services. Incorporating these consumption trips into a model of internal city structure, consumption access is shown to be quantitatively relevant for rationalizing both the observed spatial distribution of land prices and the counterfactual impact of transport infrastructure investments.

[54] For a model of a system of cities in which the costly exchange of ideas is the force for agglomeration, see Davis and Dingel (2019).

5.4 Quantifying the Impact of Transport Infrastructure Improvements

The constant elasticity structure of this class of quantitative urban models implies that exact-hat algebra techniques again can be used to quantify the impact of transport improvements on the location of economic activity. In particular, Heblich, Redding, and Sturm (2020), develop a methodology for evaluating the impact of transport infrastructure investments that holds in an entire class of quantitative urban models, because it only uses the assumptions of (i) a gravity equation in commuting, (ii) land market clearing, and (iii) payments for commercial and residential floor space are constant multiples of labor income by workplace and residence respectively. Combining this methodology with spatially-disaggregated data for London from 1801-1921, this class of quantitative urban models is shown to be remarkably successful in explaining the first large-scale separation of workplace and residence that occurred following the invention of the steam railway.

This methodology is well suited to the data-scarce environment of the 19th century, because it can be implemented using only bilateral commuting data in a baseline year and property values and employment by residence in other years. Within the structure of this class of quantitative urban models, these data provide sufficient statistics that can be used to isolate the effect of the change in commuting costs from the construction of the railway network on employment by workplace, while controlling for other unobserved changes over time in productivity, amenities, the costs of trading goods, the supply of floor space and the reservation level of utility in the wider economy.

We begin by deriving a combined land and commuter market clearing condition in this class of quantitative urban models. Using the assumption of Cobb-Douglas utility and production, expenditure on residential floor space is a constant multiple of income by residence, and expenditure on commercial floor space is a constant multiple of income by workplace. Combining both of these results, the land market clearing condition can be written as:

$$\mathbb{Q}_{nt} = (1-\alpha)\, v_{nt} R_{nt} + \frac{\beta^H}{\beta^L} w_{nt} L_{nt}, \tag{60}$$

where $(1-\alpha)$ is the share of residential floor space in household expenditure; β^H is the share of commercial floor space in firm costs; β^L is the share of labor in firm costs; and we have now made explicit the time subscript. In this land market clearing condition, the average per capita income of residents (v_{nt}) depends on the wage (w_{nt}) and the conditional commuting probabilities ($\lambda^R_{nit|n}$) as follows:

$$v_{nt} = \sum_{i \in N} \lambda^R_{nit|n} w_{it} \tag{61}$$

while employment (L_{nt}) and residents (R_{nt}) are linked through the commuter market clearing condition (54).

Substituting commuter market clearing (54) and average per capita income (61) into the land market clearing condition (60), we obtain a combined land and commuter market clearing condition. Rewriting this combined land and commuter market clearing condition in another year ($\tau \neq t$) in terms of relative changes ($\hat{x}_{nt} = x_{n\tau}/x_{nt}$) and the observed values of variables in the baseline year t, we obtain:

$$\begin{aligned} \hat{\mathbb{Q}}_{nt}\mathbb{Q}_{nt} &= (1-\alpha)\left[\sum_{i \in N} \frac{\lambda^R_{nit|n} \hat{w}^{\epsilon}_{it} \hat{\kappa}^{-\epsilon}_{nit}}{\sum_{\ell \in N} \lambda^R_{n\ell t|n} \hat{w}^{\epsilon}_{\ell t} \hat{\kappa}^{-\epsilon}_{n\ell t}} \hat{w}_{it} w_{it}\right] \hat{R}_{nt} R_{nt} \\ &\quad + \frac{\beta^H}{\beta^L} \hat{w}_{nt} w_{nt} \left[\sum_{i \in N} \frac{\lambda^R_{int|i} \hat{w}^{\epsilon}_{nt} \hat{\kappa}^{-\epsilon}_{int}}{\sum_{\ell \in N} \lambda^R_{i\ell t|i} \hat{w}^{\epsilon}_{\ell t} \hat{\kappa}^{-\epsilon}_{i\ell t}} \hat{R}_{it} R_{it}\right], \end{aligned} \tag{62}$$

where property values equal the price times the quantity of floor space: $\mathbb{Q}_{nt} = Q_{nt} H_{nt}$.

Given data on bilateral commuting ($\lambda^R_{nit|n}$), property values ($\mathbb{Q}_{nt}$), residents (R_{nt}) and wages (w_{nt}) in the baseline year of $t = 1921$, and data on relative changes in property values ($\hat{\mathbb{Q}}_{nt}$) and residents ($\hat{R}_{nt}$) for years going back to the

early 19th century, this combined land market clearing condition can be used to solve for the impact of changes in the transport network on historical employment by workplace and commuting patterns. Comparing these predictions to the available data, this class of quantitative spatial models is shown to closely replicate the sharp increase in workplace employment that occurs in the City of London in the second half of the 19th century, and successfully capture the fact that most workers lived within 5 kilometers of their workplace at the dawn of the railway age. These results are consistent with a large literature in economic history that discusses the transformation of central cities from residential to commercial activity following 19th-century transport improvements, including Jackson (1987).

5.5 Counterfactuals

An advantage of the methodology developed in the previous section is that it holds in an entire class of quantitative urban models and can be used to evaluate the impact of a transport improvements while controlling for changes in other determinants of economic activity (such as productivity and amenities). However, another question of interest is the counterfactual of how the spatial distribution of economic activity would have evolved if the only thing that changed were the railway network, holding constant other determinants of economic activity.

To address this question, Heblich, Redding, and Sturm (2020) undertake counterfactuals for removing the railway network under a range of alternative assumptions about the elasticity of supply of floor space and agglomeration forces. Three key findings from these counterfactuals are as follows. First, much of the increased separation of workplace and residence in Greater London during the 19th century can be explained by the pure change in commuting costs from the new transport technology alone. Under the assumption of an inelastic supply of floor space and no agglomeration forces, removing the entire railway network reduces net commuting into the City of London from around 370,000 in 1921 to 98,173 in 1831, which compares to 30,375 in the baseline quantitative analysis in the previous section that allows for changes in other determinants of economic activity. Second, this new transport technology has sizable effects on overall economic activity. Using the calibrated floor space supply elasticity and estimated values of both production and residence agglomeration forces, removing the entire railway network decreases the total population of Greater London by 51.45 percent to 3.59 million in 1831.

Third, the net present value of the counterfactual changes in property values exceeds historical estimates of the construction costs of the railway network for plausible values of the discount rate. Allowing for a positive floor space supply elasticity substantially enhances the impact of the railway network on property values, highlighting the role of complementary expansions in the supply of floor space in influencing the effects transport infrastructure improvements. Introducing agglomeration economies in production and residence further magnifies the impact of the railway network on property values, illustrating the relevance of endogenous changes in productivity and amenities for cost-benefit evaluations of transport improvements.

In the class of quantitative urban models considered here with a single type of worker, all workers are affected in the same way by the transport improvement. In settings with multiple groups of *ex ante* heterogeneous workers, improvements in transport infrastructure or other public policy interventions can have distributional effects on these different groups of workers, depending on patterns of spatial sorting. To analyze the impact of Bogotá's Bus Rapid Transit system, Tsivanidis (2018) develops a quantitative spatial model with low and high-skilled workers who have non-homothetic preferences over neighborhoods and transit modes. Although low-skilled workers use public transit the most, they have a larger elasticity of commuting decisions to commuting costs, and hence have larger labor

supply response to the transport improvement, thus inducing a subtle distributional effects between these two groups of workers. Similarly, using data from Dar es Salaam's Bus Rapid Transit system, Balboni, Bryan, Morten, and Siddiqi (2020) find distributional effects between low and high-income workers. Although these interventions are place-based in the sense that they are built in specific neighborhoods, they induce changes in patterns of spatial sorting throughout the city, and hence need not benefit the residents initially living in those neighborhoods.

6 Conclusions

The last decade has seen substantial research progress on geography and trade. A key contribution of recent empirical research has been to show that geography is an important dimension along which the distributional consequences of international trade occur. If industries are unevenly spatially distributed and factors of production are imperfectly mobile, different local labor markets within countries are differentially exposed to trade shocks. In influential work, Autor, Dorn, and Hanson (2013a) have shown that local labor markets more exposed to Chinese import growth experience larger reductions in manufacturing employment as a share of the working-age population, the employment to population ratio and mean log weekly earnings, as well as larger increases in per capita unemployment, disability, and income assistance transfer benefits. These empirical findings have stimulated a productive dialogue between theory and empirics, as researchers have developed models to rationalize these empirical findings.

A key theoretical advance has been the development of quantitative spatial models that are sufficiently rich as to connect directly with the observed data on many asymmetric locations. These frameworks incorporate differences across locations in both *first-nature* geography (locational endowments, such as access to the coast) and *second-nature* geography (the proximity of economic agents relative to one another in geographic space). Typically, these models have the property that they can be inverted to recover unobserved location characteristics (e.g. productivity, amenities and trade costs) that exactly rationalize the observed data on the endogenous variables of the model as an equilibrium outcome. Therefore, these frameworks can be used to quantify the roles of first and second-nature geography in explaining the observed distribution of economic activity. Nevertheless, these frameworks remain sufficiently tractable as to permit an analytical characterization of the existence and uniqueness of the general equilibrium and to be used for realistic counterfactuals. Importantly, these counterfactuals can be undertaken using the observed values of endogenous variables in an initial equilibrium to control for a wide range of unobserved locational fundamentals that affect the spatial distribution of economic activity.

We distinguish between models of regions or systems of cities (where goods trade and migration take central stage) and models of the internal structure of cities (where commuting becomes relevant). All three types of spatial interactions are well approximated by gravity equations in which spatial interactions decline sharply in the distance between economic agents. From models of regions or systems of cities, we now have a good understanding of the role of market access in shaping the spatial distribution of economic activity, the circumstances under which there is path dependence such that temporary shocks can have permanent effects on the spatial distribution of economic activity, and the role of imperfect factor mobility in rationalizing findings from the reduced-form literature on local labor markets. From models of the internal structure of cities, we now have a good understanding of overall magnitude of agglomeration forces and the implications of transport infrastructure improvements for the internal organization of economic activity within cities. In both these areas, reduced-form methods and quantitative models have comple-

mented one another: the empirical moments from reduced-form studies provide discipline for model parameters; the predictions of the quantitative models shed light on general equilibrium effects and model-based objects such as a welfare that are difficult to recover from reduced-form specifications alone.

Looking ahead, there is much that remains to be done. The explosion of new sources of geographical information systems (GIS) data provides an heretofore unprecedented level of detail on economic activity at a fine spatial scale. These new sources of data include ride-hailing data (e.g. Uber and Lyft), smartphone data with Global Positioning System (GPS) information, firm-to-firm transactions data from sales (VAT) tax records, credit card data with consumer and firm location, barcode scanner data with consumer and firm location, public transportation data on commuting, and satellite-imaging data. Whereas theoretical and empirical research on spatial linkages has traditionally focused on international trade between countries for reasons of data availability, we now increasingly have the ability to measure these linkages between locations at fine spatial scales. Promising areas for further research are combining quantitative spatial models with this wealth of empirical information to shed light on a whole host of issues, such as discriminating between alternative mechanisms for agglomeration, understanding the implications of new technologies for the organization or work, and assessing the causes, consequences and policy implications of spatial sorting.

References

Adão, R., C. Arkolakis, and F. Esposito (2019): "Spatial Linkages, Global Shocks, and Local Labor Markets: Theory and Evidence," *NBER Working Paper*, 25544.

Adão, R., M. Kolesár, and E. Morales (2019): "Shift-Share Designs: Theory and Inference," *Quarterly Journal of Economics*, 134(4), 1949–2010.

Ahlfeldt, G., and E. Pietrostefani (2019): "The Economic Effects of Density: A Synthesis," *Journal of Urban Economics*, 111, 93–107.

Ahlfeldt, G., S. Redding, D. Sturm, and N. Wolf (2015): "The Economics of Density: Evidence from the Berlin Wall," *Econometrica*, 83(6), 2127–2189.

Ahuja, R. K., T. L. Magnanti, and J. B. Orlin (1993): *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Upper Saddle River, NJ.

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

——— (2017): "The Welfare Effects of Transportation Infrastructure Improvements," Dartmouth College.

Allen, T., C. Arkolakis, and X. Li (2017): "Optimal City Structure," Yale University, mimeograph.

——— (2020): "On the Equilibrium Properties of Network Models with Heterogeneous Agents," *NBER Working Paper*, 27837.

Allen, T., C. Arkolakis, and Y. Takahashi (2019): "Universal Gravity," *Journal of Political Economy*, 128(2), 393–433.

Allen, T., and D. Donaldson (2020): "Persistence and Path Dependence in the Spatial Economy," *NBER Working Paper*, 28059, Dartmouth College, mimeograph.

Almagro, M., and T. Domínguez-Iino (2019): "Location Sorting and Endogenous Amenities: Evidence from Amsterdam," New York University, mimeograph.

Amior, M., and A. Manning (2018): "The Persistence of Local Joblessness," *American Economic Review*, 108(7), 1942–1970.

Amiti, M., M. Dai, R. C. Feenstra, and J. Romalis (2017): "How Did China's WTO Entry Affect U.S. Prices?," *NBER Working Paper*, 23487.

Anderson, J. E., and E. van Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 93(1), 170–192.

Angrist, J. D., and J.-S. Pischke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24(2), 3–30.

Arkolakis, C., A. Costinot, and A. Rodríguez-Clare (2012): "New Trade Models, Same Old Gains?," *American Economic Review*, 102(1), 94–130.

Armington, P. S. (1969): "A Theory of Demand for Products Distinguished by Place of Production," *IMF Staff Papers*, 16, 159–178.

Artuç, E., S. Chaudhuri, and J. McLaren (2010): "Trade Shocks and Labor Adjustment: A Structural Empirical Approach," *American Economic Review*, 100(3), 1008–45.

Arzaghi, M., and J. V. Henderson (2008): "Networking Off Madison Avenue," *Review of Economic Studies*, 75(4), 1011–1038.

Austin, B., E. Glaeser, and L. Summers (2018): "Jobs for the Heartland: Place-Based Policies in 21st-Century America," *Brookings Papers on Economic Activity*, Spring, 151–232.

Autor, D. (2019): "Work of the Past, Work of the Future," *American Economic Review*, 109(5), 1–32, Richard Ely Lecture.

Autor, D., D. Dorn, G. Hanson, and K. Majlesi (2020): "Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure," *American Economic Review*, forthcoming.

Autor, D., D. Dorn, and G. H. Hanson (2013a): "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, 103(6), 2121–68.

——— (2013b): "The Geography of Trade and Technology Shocks in the United States," *American Economic Review*, 103(3), 220–225, Papers and Proceedings.

——— (2020): "On the Persistence of the China Shock," MIT, mimeograph.

Autor, D., D. Dorn, G. H. Hanson, G. Pisano, and P. Shu (2020): "Foreign Competition and Domestic Innovation: Evidence from U.S. Patents," *American Economic Review: Insights*, forthcoming.

Autor, D., D. Dorn, G. H. Hanson, and J. Song (2014): "Trade Adjustment: Worker-Level Evidence," *Quarterly Journal of Economics*, 129(4), 1799–1860.

Autor, D. H., D. Dorn, and G. H. Hanson (2016): "The China Shock: Learning from Labor Market Adjustment to Large Changes in Trade," *Annual Review of Economics*, 8, 205–240.

——— (2019): "When Work Disappears: Manufacturing Decline and the Falling Marriage-Market Value of Young Men," *American Economic Review: Insights*, 1(2), 161–178.

Balboni, C. (2019): "In Harm's Way? Infrastructure Investments and the Persistence of Coastal Cities," MIT, mimeograph.

Balboni, C., G. Bryan, M. Morten, and B. Siddiqi (2020): "Transportation, Gentrification, and Urban Mobility: The Inequality Effects of Place-Based Policies," MIT, mimeograph.

Banerjee, A., E. Duflo, and N. Qian (2020): "On the Road: Access to Transportation Infrastructure and Economic Growth in China," *Journal of Development Economics*, 145, 1–36.

Bartelme, D. (2020): "Trade Costs and Economic Geography: Evidence from the US," University of Michigan, mimeograph.

Bartik, T. J. (1991): *Who Benefits from State and Local Economic Development Policies?* W. E. Upjohn Institute for Employment Research, Kalamazoo MI.

Baum-Snow, N. (2007): "Did Highways Cause Suburbanization?," *Quarterly Journal of Economics*, 122(2), 775–780.

Baum-Snow, N. (2019): "Urban Transport Expansions and Changes in the Spatial Structure of US Cities: Implications for Productivity and Welfare," *Review of Economics and Statistics*, forthcoming.

Baum-Snow, N., L. Brandt, J. Henderson, M. A. Turner, and Q. Zhang (2017): "Roads, Railroads and Decentralization of Chinese Cities," *Review of Economics and Statistics*, 99(2), 435–448.

Bernard, A. B., J. B. Jensen, and P. K. Schott (2006): "Survival of the Best Fit: Exposure to Low-Wage Countries and the (Uneven) Growth of U.S. Manufacturing Plants," *Journal of International Economics*, 68, 219–37.

Bernard, A. B., S. J. Redding, and P. K. Schott (2013): "Testing for Factor Price Equality with Unobserved Differences in Factor Quality or Productivity," *American Economic Journal: Microeconomics*, 5(2), 135–163.

Bernhofen, D., Z. El-Sahli, and R. Kneller (2016): "Estimating the Effects of the Container Revolution on World Trade," *Journal of International Economics*, pp. 36–50.

Bleakley, H., and J. Lin (2012): "Portage: Path Dependence and Increasing Returns in U.S. History," *Quarterly Journal of Economics*, 127(2), 587–644.

Bloom, N., M. Draca, and J. Van Reenen (2016): "Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT and Productivity," *Review of Economic Studies*, 83, 87–117.

Bloom, N., K. Handley, A. Kurman, and P. Luck (2020): "The impact of Chinese trade on US Employment: The Good, The Bad and the Debatable," Stanford University, mimeograph.

Borusyak, K., P. Hull, and X. Jaravel (2019): "Quasi-Experimental Shift-share Research Designs," University College London, mimeograph.

Borusyak, K., and X. Jaravel (2018): "The Distributional Effects of Trade: Theory and Evidence from the U.S.," London School of Economics, mimeograph.

Bosker, M., S. Brakman, H. Garretsen, and M. Schramm (2007): "Looking for Multiple Equilibria When Geography Matters: German City Growth and the WWII Shock," *Journal of Urban Economics*, 61, 152–167.

Brakman, S., H. Garretsen, and M. Schramm (2004): "The Strategic Bombing of German Cities during WWII and its Impact on City Growth," *Journal of Economic Geography*, 4(2), 201–218.

Brancaccio, G., M. Kaloupsidi, and T. Papageorgiou (2020): "Geography, Transportation and Endogenous Trade Costs," *Econometrica*, 88(2), 657–691.

Breinlich, H. (2006): "The Spatial Income Structure in the European Union - What Role for Economic Geography?," *Journal of Economic Geography*, 6(5), 593–617.

Brooks, L., N. Gendron-Carrier, and G. Rua (2019): "The Local Impact of Containterization," George Washington University, Washington DC.

Brooks, L., and Z. D. Liscow (2019): "Infrastructure Costs," *SSRN Working Paper*, pp. 385–399, http://dx.doi.org/10.2139/ssrn.3428675.

Brülhart, M., C. Carrère, and F. Trionfetti (2012): "How Wages and Employment Adjust to Trade Liberalization: Quasi-experimental Evidence from Austria," *Journal of International Economics*, 86, 68–81.

Bryan, G., E. Glaeser, and N. Tsivanidis (2020): "Cities in the Developing World," *Annual Review of Economics*, 12, 273–297.

Bryan, G., and M. Morten (2019): "The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia," *Journal of Political Economy*, 127(5), 2229–2268.

Caliendo, L., M. Dvorkin, and F. Parro (2019): "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock," *Econometrica*, 87(3), 741–835.

Caliendo, L., and F. Parro (2015): "Estimates of the Trade and Welfare Effects of NAFTA," *Review of Economic Studies*, 82(1), 1–44.

Caliendo, L., F. Parro, E. Rossi-Hansberg, and P.-D. Sarte (2018): "The Impact of Regional and Sectoral Productivity Changes on the U.S. Economy," *Review of Economic Studies*, 85(4), 2042–2096.

Chandra, A., and E. Thompson (2000): "Does Public Infrastructure affect Economic Activity? Evidence from the Rural Interstate Highway System," *Regional Science and Urban Economics*, 30(4), 457–490.

Che, Y., Y. Lu, J. R. Pierce, P. K. Schott, and Z. Tao (2020): "Did Trade Liberalization with China Influence U.S. Elections?," Yale of School of Management, mimeograph.

Costinot, A., D. Donaldson, M. Kyle, and H. Williams (2019): "The More We Die, The More We Sell? A Simple Test of the Home-Market Effect," *Quarterly Journal of Economics,*, 134(2), 843–894.

Couture, V. (2016): "Valuing the Consumption Benefits of Urban Density," Vancouver School of Economics, mimeograph.

Couture, V., C. Gaubert, and J. Handbury (2020): "Income Growth and the Distributional Effects of Urban Spatial Sorting," University of Chicago, mimeograph.

Couture, V., C. Gaubert, J. Handbury, and E. Hurst (2018): "Income Growth and the Distributional Effects of Urban Spatial Sorting," University of California Berkeley, mimeograph.

Davis, D. R., and J. Dingel (2019): "A Spatial Knowledge Economy," *American Economic Review*, 109(1), 153–70.

——— (2020): "The Comparative Advantage of Cities," *Journal of International Economics*, 123, 1–27.

Davis, D. R., J. Dingel, J. Monras, and E. Morales (2019): "How Segregated is Urban Consumption?," *Journal of Political Economy*, 127(4), 1684–1738.

Davis, D. R., and D. E. Weinstein (1999): "Economic Geography and Regional Production Structure: An Empirical Investigation," *European Economic Review*, 43(2), 379–407.

——— (2002): "Bones, Bombs, and Break Points: The Geography of Economic Activity," *American Economic Review*, 92(5), 1269–1289.

——— (2003): "Market Access, Economic Geography and Comparative Advantage: an Empirical Test," *Journal of International Economics*, 59(1), 1–23.

——— (2008): "A Search for Multiple Equilibria in Urban Industrial Structure," *Journal of Regional Science*, 48(1), 29–65.

Davis, M. A., and F. Ortalo-Magné (2011): "Housing Expenditures, Wages, Rents," *Review of Economic Dynamics*, 14(2), 248–261.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2014): "The Role of Entrepreneurship in US Job Creation and Economic Dynamism," *Journal of Economic Perspectives*, 28, 3–24.

Dekle, R., and J. Eaton (1999): "Agglomeration and Land Rents: Evidence from the Prefectures," *Journal of Urban Economics*, 46(2), 200–214.

Dekle, R., J. Eaton, and S. Kortum (2007): "Unbalanced Trade," *American Economic Review*, 97(2), 351–355.

Dell, M. (2010): "The Persistent Effects of Peru's Mining Mita," *Econometrica*, 78(6), 1863–1903.

Dell, M., and P. Querubin (2018): "Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies," *Quarterly Journal of Economics*, 133(2), 701–764.

Desmet, K., and J. V. Henderson (2015): "The Geography of Development Within Countries," in *Handbook of Regional and Urban Economics*, vol. 5, chap. 22, pp. 1457–1517. North-Holland, Amsterdam.

Desmet, K., R. Kopp, S. A. Kulp, D. K. Nagy, M. Oppenheimer, E. Rossi-Hansberg, and B. H. Strauss (2020): "Evaluating the Economic Cost of Coastal Flooding," *American Economic Journal: Macroeconomics*, forthcoming.

Desmet, K., D. K. Nagy, and E. Rossi-Hansberg (2018): "The Geography of Development," *Journal of Political Economy*, 126(3), 903–983.

Desmet, K., and E. Rossi-Hansberg (2014): "Spatial Development," *American Economic Review*, 104(4), 1211–1243.

Diamond, R. (2016): "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000," *American Economic Review*, 106(3), 479–524.

Ding, X., T. Fort, S. J. Redding, and P. K. Schott (2019): "Structural Change Within Versus Across Firms: Evidence from the United States," Princeton University, mimeograph.

Dingel, J. I. (2017): "The Determinants of Quality Specialization," *Review of Economic Studies*, 84, 1551–1582.

Dingel, J. I., and F. Tintelnot (2020): "Spatial Economics for Granular Settings," *NBER Working Paper*, 27287.

Dix-Carneiro, R. (2014): "Trade Liberalization and Labor Market Dynamics," *Econometrica*, 82(3), 825–885.

Dix-Carneiro, R., and B. K. Kovak (2017): "Trade Liberalization and Regional Dynamics," *American Economic Review*, 107(10), 1908–2946.

Dix-Carneiro, R., J. Pessoa, R. Reyes-Heroles, and S. Traiberman (2020): "Globalization, Trade Imbalances, and Labor Market Adjustment," New York University, mimeograph.

Donaldson, D. (2015): "The Gains from Market Integration," *Annual Review of Economics*, 7, 619–647.

——— (2018): "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure," *American Economic Review*, 108(4-5), 899–934.

Donaldson, D., and R. Hornbeck (2016): "Railroads and American Economic Growth: A Market Access Approach," *Quarterly Journal of Economics*, 131(2), 799–858.

Ducruet, C., R. Juhász, D. K. Nagy, and C. Steinwender (2020): "All Aboard: The Aggregate Effects of Port Development," CREI, Pompeu Fabra, mimeograph.

Duranton, G., P. Morrow, and M. A. Turner (2014): "Roads and Trade: Evidence from the US," *Review of Economic Studies*, 81(2), 681–724.

Duranton, G., G. Nagpal, and M. A. Turner (2020): "Transportation Infrastructure in the US," *NBER Working Paper*, 27254.

Duranton, G., and M. A. Turner (2011): "The Fundamental Law of Road Congestion: Evidence from US Cities," *American Economic Review*, 101(6), 2616–52.

——— (2012): "Urban Growth and Transportation," *Review of Economic Studies*, 79(4), 1407–1440.

Eaton, J., and S. Kortum (2002): "Technology, Geography, and Trade," *Econometrica*, 70(5), 1741–1779.

Eaton, J., S. Kortum, and B. Neiman (2016): "Obstfeld and Rogoff's International Macro Puzzles: A Quantitative Assessment," *Journal of Economic Dynamics and Control*, 72, 5–23.

Ehrlich, M., and T. Seidel (2018): "The Persistent Effects of Place-Based Policy: Evidence from the West-German Zonenrandgebiet," *American Economic Journal: Economic Policy*, 10(4), 344–374.

Eriksson, K., K. N. Russ, J. C. Shambaugh, and M. Xu (2019): "Trade Shocks and the Shifting Landscape of US Manufacturing," *NBER Working Paper*, 25646.

Faber, B. (2014): "Integration and the Periphery: The Unintended Effects of New Highways in a Developing Country," *Review of Economic Studies*, 81(3), 1046–1070.

Fajgelbaum, P., and C. Gaubert (2020): "Optimal Spatial Policies, Geography, and Sorting," *Quarterly Journal of Economics*, 135(2), 959–1036.

Fajgelbaum, P., and A. Khandelwal (2016): "Measuring the Unequal Gains from Trade," *Quarterly Journal of Economics*, 131, 1113–1180.

Fajgelbaum, P., E. Morales, J. C. Suárez Serrato, and O. Zidar (2019): "State Taxes and Misallocation," *Review of Economic Studies*, 86, 333–376.

Fajgelbaum, P., and S. Redding (2018): "Trade, Structural Transformation and Development: Evidence from Argentina 1869-1914," *NBER Working Paper*, 20217.

Fajgelbaum, P. D., and E. Schaal (2020): "Optimal Transport Networks in Spatial Equilibrium," *Econometrica*, 88(4), 1411–1452.

Feenstra, R. C., H. Ma, and Y. Xu (2019): "US Exports and Employment," *Journal of International Economics*, 120, 46–58.

Feyrer, J. (2009): "Distance, Trade, and Income - The 1967 to 1975 Closing of the Suez Canal as a Natural Experiment," *NBER Working Paper*, 15557.

Fogel, R. W. (1964): *Railroads and American Economic Growth: Essays in Econometric History*. Johns Hopkins University Press, Baltimore, MD.

Fortheringham, S., and M. O'Kelly (1989): *Spatial Interaction Models: Formulations and Applications*. Kluwer, Dordrecht.

Fujimoto, T., and U. Krause (1985): "Strong Ergodicity for Strictly Increasing Nonlinear Operators," *Linear Algebra Applications*, 71(1), 101–12.

Fujita, M., P. Krugman, and A. J. Venables (1999): *The Spatial Economy: Cities, Regions, and International Trade*. MIT Press, Cambridge MA.

Galle, S., M. Yi, and A. Rodriguez-Clare (2018): "Slicing the Pie: Quantifying the Aggregate and Distributional Consequences of Trade," University of Calirfornia, Berkeley, mimeograph.

Garcia-López, M.-À., A. Holl, and E. Viladecans-Marsal (2015): "Suburbanization and Highways: When the Romans, the Bourbons and the First Cars Still Shape Spanish Cities," *Journal of Urban Economics*, 85, 52–67.

Gaubert, C., P. Kline, and D. Yagan (2020): "Place-based Redistribution," University of California, Berkeley.

Glaeser, E. L., and J. Gyourko (2005): "Urban Decline and Durable Housing," *Journal of Political Economy*, 113(2), 345–375.

Glaeser, E. L., and J. E. Kohlhase (2004): "Cities, Regions and the Decline of Transport Costs," *Papers in Regional Science*, 83, 197–228.

Glaeser, E. L., J. Kolko, and A. Saiz (2001): "Consumer City," *Journal of Economic Geography*, 1(1), 27–50.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020): "Bartik Instruments: What, When, Why, and How," *American Economic Review*, 110(8), 2586–2624.

Greenland, A., J. Lopresti, and P. McHenry (2019): "Import Competition and Internal Migration," *Review of Economics and Statistics*, 101(1), 44–59.

Greenstone, M., R. Hornbeck, and E. Moretti (2010): "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings," *Journal of Political Economy*, 118(3), 536–598.

Grogger, J., and G. Hanson (2011): "Income Maximization and the Selection and Sorting of International Migrants," *Journal of Development Economics*, 95(1), 42–57.

Handbury, J., and D. E. Weinstein (2015): "Goods Prices and Availability in Cities," *Review of Economic Studies*, 82, 258–296.

Handley, K., and N. Limão (2017): "Policy Uncertainty, Trade and Welfare: Theory and Evidence for China and the United States," *American Economic Review*, 107(9), 2731–83.

Hanson, G. H. (1996): "Localization Economies, Vertical Organization, and Trade," *American Economic Review*, 86(5), 1266–1278.

——— (1997): "Increasing Returns, Trade, and the Regional Structure of Wages," *Economic Journal*, 107(1), 113–133.

——— (2005): "Market Potential, Increasing Returns, and Geographic Concentration," *Journal of International Economics*, 67(1), 1–24.

Harris, C. D. (1954): "The Market as a Factor in the Localization of Industry in the United States," *Annals of the Association of American Geographers*, 44(4), 315–348.

Head, K., and T. Mayer (2006): "Regional Wage and Employment Responses to Market Potential in the EU'," *Regional Science and Urban Economics*, 36(5), 573–594.

Head, K., and J. Ries (2001): "Increasing Returns versus National Product Differentiation as an Explanation for the Pattern of U.S.-Canada Trade," *American Economic Review*, 91(4), 858–876.

Heblich, S., S. Redding, and D. Sturm (2020): "The Making of the Modern Metropolis: Evidence from London," *Quarterly Journal of Economics*, 135(4), 2059–2133.

Heblich, S., S. Redding, and Y. Zylberberg (2020): "The Distributional Consequences of Trade: Evidence from the Repeal of the Corn Laws," Princeton University, mimeograph.

Helpman, E. (1998): "The Size of Regions," in *Topics in Public Economics: Theoretical and Applied Analysis*, ed. by D. Pines, E. Sadka, and I. Zilcha, pp. 33–54. Cambridge University Press, Cambridge.

——— (2018): *Globalization and Inequality*. Harvard University Press, Cambridge MA.

Helpman, E., O. Itskhoki, M.-A. Muendler, and S. J. Redding (2017): "Trade and Inequality: From Theory to Estimation," *Review of Economic Studies*, 84(1), 357–405.

Helpman, E., O. Itskhoki, and S. J. Redding (2010): "Inequality and Unemployment in a Global Economy," *Econometrica*, 78(4), 1239–1283.

Hirschman, A. (1958): *The Strategy of Economic Development*. Yale University Press, New Haven.

Holmes, T. J. (2005): "The Location of Sales Offices and the Attraction of Cities," *Journal of Political Economy*, 113(3), 551–81.

——— (2011): "The Diffusion of Walmart and the Economies of Density," *Econometrica*, 79(1), 253–302.

Hornbeck, R., and D. Keniston (2017): "Creative Destruction: Barriers to Urban Growth and the Great Boston Fire of 1872," *American Economic Review*, 107(6), 1365–98.

Hornbeck, R., and M. Rotemberg (2021): "Railroads, Market Access, and Aggregate Productivity Growth," Chicago Booth School of Business, mimeograph.

Hsieh, C.-T., and E. Moretti (2019): "Housing Constraints and Spatial Misallocation," *American Economic Journal: Macroeconomics*, 11(2), 1–39.

Hsu, W.-T., and H. Zhang (2014): "The fundamental law of highway congestion revisited: Evidence from national expressways in Japan," *Journal of Urban Economics*, 81, 65–76.

Hulten, C. R. (1978): "Growth Accounting with Intermediate Inputs," *Review of Economic Studies*, pp. 511–518.

Jackson, K. T. (1987): *Crabgrass Frontier: The Suburbanization of the United States*. Oxford University Press, Oxford.

Jacobson, L. S., R. J. LaLonde, and D. G. Sullivan (1993): "Earnings Losses of Displaced Workers," *American Economic Review*, 83(4), 685–709.

Jedwab, R., and A. Storeygard (2020): "The Average and Heterogeneous Effects of Transportation Investments: Evidence from Sub-Saharan Africa 1960-2010," *NBER Working Paper*, 27670.

Kaplan, G., and S. Schulhofer-Wohl (2017): "Understanding the Long-Run Decline in Interstate Migration," *International Economic Review*, 58(1), 57–94.

Kennan, J., and J. R. Walker (2011): "The Effect of Expected Income on Individual Migration Decisions," *Econometrica*, 79(1), 211–251.

Kim, R., and J. Vogel (2020): "Trade Shocks and Labor Market Adjustment," *American Economic Review: Insights*, forthcoming.

Kleinman, B., E. Liu, and S. Redding (2020): "International Friends and Enemies," *NBER Working Paper*, 27587.

——— (2021): "Sufficient Statistics for Dynamic Spatial Economics," Princeton University, mimeograph.

Kline, P., and E. Moretti (2014a): "Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority," *Quarterly Journal of Economics*, 129(1), 275–331.

——— (2014b): "People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Policies," *Annual Review of Economics*, 6, 629–662.

Kovak, B. K. (2013): "Regional Effects of Trade Reform: What is the Correct Measure of Liberalization?," *American Economic Review*, 103(5), 1960–1976.

Krugman, P. (1991a): *Geography and Trade*. MIT Press, Cambridge MA.

——— (1991b): "History Versus Expectations," *Quarterly Journal of Economics*, 106(2), 651–667.

——— (1991c): "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99(3), 483–499.

Krugman, P., and A. Venables (1995): "Globalization and the Inequality of Nations," *Quarterly Journal of Economics*, 110(4), 857–880.

Lucas, R. E., and E. Rossi-Hansberg (2002): "On the Internal Structure of Cities," *Econometrica*, 70(4), 1445–1476.

Magyari, I. (2017): "Firm Reorganization, Chinese Imports, and US Manufacturing Employment," Columbia University, mimeograph.

Manski, C. F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60(3), 531–542.

Martincus, C. V., J. Carballo, and A. Cusolito (2017): "Roads, Exports and Employment: Evidence from a Developing Country," *Journal of Development Economics*, 125, 21–39.

Matsuyama, K. (1991): "Increasing Returns, Industrialization, and Indeterminacy of Equilibrium," *Quarterly Journal of Economics*, CVI, 617–50.

Mayer, T., and S. Zignago (2011): "Notes on CEPII's Distances Measures: The GeoDist Database," *CEPII Working Paper*, 25.

McCaig, B. (2011): "Exporting out of Poverty: Provincial Poverty in Vietnam and U.S. Market Access," *Journal of International Economics*, 85(1), 102–113.

McCaig, B., and N. Pavcnik (2018): "Export Markets and Labor Allocation in a Low-Income Country," *American Economic Review*, 108(7), 1899–1941.

McDonald, J., and D. McMillen (2010): *Urban Economics and Real Estate: Theory and Policy*. John Wiley & Sons, Hoboken, NJ.

McFadden, D. (1974): "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 3(4), 303–328.

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71(6), 1695–725.

Melo, P. C., D. J. Graham, and R. B. Noland (2009): "A Meta-analysis of Estimates of Urban Agglomeration Economies," *Regional Science and Urban Economics*, 39(3), 332–342.

Michaels, G. (2008): "The Effect of Trade on the Demand for Skill: Evidence from the Interstate Highway System," *Review of Economics and Statistics*, 90(4), 683–701.

Michaels, G., and F. Rauch (2018): "Resetting the Urban Network: 117-2012," *Economic Journal*, 128(608), 378–412.

Michaels, G., F. Rauch, and S. Redding (2019): "Task Specialization in U.S. Cities from 1880-2000," *Journal of the European Economic Association*, 17(3), 754–798.

Michaels, G., F. Rauch, and S. J. Redding (2012): "Urbanization and Structural Transformation," *Quarterly Journal of Economics*, 127, 535–586.

Miguel, E., and G. Roland (2011): "The Long-run Impact of Bombing Vietnam," *Journal of Development Economics*, 96, 1–15.

Mitchell, J. S., and D. M. Keirsey (1984): "Planning Strategic Paths through Variable Terrain Data," *Technical Symposium East (International Society for Optics and Photonics, 1984)*, pp. 172–179.

Miyauchi, Y., K. Nakajima, and S. Redding (2021): "Consumption Access and Agglomeration: Evidence from Smartphone Data," *CEPR Discussion Paper*, 6741-1613580728.

Monte, F., S. Redding, and E. Rossi-Hansberg (2018): "Commuting, Migration and Local Employment Elasticities," *American Economic Review*, 108(12), 3855–3890.

Moretti, E. (2013): *The New Geography of Jobs*. Mariner Books, New York.

Morten, M., and J. Oliveira (2018): "The Effects of Roads on Trade and Migration: Evidence from a Planned Capital City," Stanford University, mimeograph.

Nagy, D. K. (2020): "Hinterlands, City formation and Growth: Evidence from the U.S. Westward Expansion," CREI, Pompeu Fabra, mimeograph.

Nakajima, K. (2008): "Economic Division and Spatial Relocation: The Case of Postwar Japan," *Journal of the Japanese and International Economics*, 22, 383–400.

Nakamura, E., and J. Steinsson (2018): "Identification in Macroeconomics," *Journal of Economic Perspectives*, 32(3), 59–86.

Neal, D. (1995): "Industry-specific Human Capital: Evidence from Displaced Workers," *Journal of Labor Economics*, 13, 653–677.

Overman, H. G., S. J. Redding, and A. J. Venables (2003): "The Economic Geography of Trade, Production and Income: A Survey of Empirics," in *Handbook of International Trade*, ed. by E. Kwan-Choi, and J. Harrigan, pp. 353–87. Basil Blackwell, Oxford.

Overman, H. G., and L. A. Winters (2006): "Trade Shocks and Industrial Location: the Impact of EEC Accession on the UK," *CEP Discussion Paper*, 588.

Owens III, R., E. Rossi-Hansberg, and P.-D. Sarte (2020): "Rethinking Detroit," *American Economic Journal*, 12(2), 258–305.

Pascali, L. (2017): "The Wind of Change: Maritime Technology, Trade, and Economic Development," *American Economic Review*, 107(9), 2821–54.

Pavcnik, N. (2017): "The Impact of Trade on Inequality in Developing Countries," *The Jackson Hole Economic Policy Symposium Proceedings, Fostering a Dynamic Global Economy*, pp. 61–114.

Pierce, J. R., and P. K. Schott (2016): "The Surprisingly Swift Decline of US Manufacturing Employment," *American Economic Review*, 106(7), 1632–62.

——— (2020): "Trade Liberalization and Mortality: Evidence from U.S. Counties," *American Economic Review: Insights*, 2(1), 47–64.

Platt Boustan, L. (2020): *Competition in the Promised Land: Black Migrants in Northern Cities and Labor Markets*. Princeton University Press, Princeton.

Redding, S. J. (2010): "The Empirics of New Economic Geography," *Journal of Regional Science*, 50(1), 297–311.

——— (2011): "Economic Geography: a Review of the Theoretical and Empirical Literature," in *Palgrave Handbook of International Trade*, ed. by D. Bernhofen, R. Falvey, D. Greenaway, and U. Kreickemeier. Palgrave Macmillan, London.

——— (2016): "Goods Trade, Factor Mobility and Welfare," *Journal of International Economics*, 101, 148–167.

Redding, S. J., and E. Rossi-Hansberg (2017): "Quantitative Spatial Models," *Annual Review of Economics*, 9, 21–58.

Redding, S. J., and P. K. Schott (2003): "Distance, Skill Deepening and Development: Will Peripheral Countries Ever Get Rich?," *Journal of Development Economics*, 72(2), 515–41.

Redding, S. J., and D. M. Sturm (2008): "The Costs of Remoteness: Evidence from German Division and Reunification," *American Economic Review*, 98(5), 1766–1797.

——— (2016): "Neighborhood Effects: Evidence from the Streets of London," Princeton University, mimeograph.

Redding, S. J., D. M. Sturm, and N. Wolf (2011): "History and Industry Location: Evidence from German Airports," *Review of Economics and Statistics*, 93(3), 814–831.

Redding, S. J., and M. A. Turner (2015): "Transportation Costs and the Spatial Organization of Economic Activity," in *Handbook of Regional and Urban Economics*, ed. by G. Duration, J. V. Henderson, and W. Strange, chap. 20, pp. 1339–1398. Elsevier, Amsterdam.

Redding, S. J., and A. J. Venables (2004): "Economic Geography and International Inequality," *Journal of International Economics*, 62(1), 53–82.

Reyes-Heroles, R. (2016): "The Role of Trade Costs in the Surge of Trade Imbalances," Federal Reserve Board, Washington DC.

Rodriguez-Clare, A., M. Ulate, and J. P. Vasquez (2018): "New Keynesian Trade: Understanding the Employment and Welfare Effects of Sectoral Shocks," University of California, Bekeley.

Rosenthal, S. S., and W. C. Strange (2004): "Evidence on the Nature and Sources of Agglomeration Economics," in *Handbook of Regional and Urban Economics*, ed. by J. V. Henderson, and J. Thisse, vol. 4. Elsevier, Amsterdam.

Rossi-Hansberg, E., P.-D. Sarte, and R. I. Owens (2010): "Housing Externalities," *Journal of Political Economy*, 118(3), 485–535.

Rossi-Hansberg, E., P.-D. Sarte, and F. Schwartzman (2020): "Cognitive Hubs and Spatial Redistribution," Princeton University, mimeograph.

Ruggles, S., S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, and M. Sobek (2018): "IPUMS USA," Minneapolis, MN.

Saiz, A. (2010): "The Geographic Determinants of Housing Supply," *Quarterly Journal of Economics*, 125(3), 1253–1296.

Serrato, J. C. S., and O. Zidar (2016): "Who Benefits from State Corporate Tax Cuts? A Local Labor Market Approach with Heterogeneous Firms," *American Economic Review*, 106(9), 2582–2624.

Sethian, J. (1996): "A Fast Marching Level Set Method for Monotonically Advancing Fronts," *Proceedings of the National Academy of Sciences*, 93, 1591–1595.

Shoven, J. B., and J. Whalley (1992): *Applying General Equilibrium*. Cambridge University Press, Cambridge.

Storeygard, A. (2016): "Farther on Down the Road: Transport Costs, Trade and Urban Growth," *Review of Economic Studies*, 83(3), 1263–1295.

Tirado, D. A., E. Paluzie, and J. Pons (2002): "Economic Integration and Industrial Location: the Case of Spain Before WWI," *Journal of Economic Geography*, 2(3), 343–63.

Topalova, P. (2010): "Factor Immobility and Regional Impacts of Trade Liberalization: Evidence on Poverty from India," *American Economic Journal: Applied Economics*, 2(4), 1–41.

Traiberman, S. (2019): "Occupations and Import Competition: Evidence from Denmark," *Econometrica*, 109(12), 4260–4301.

Tsitsiklis, J. (1995): "Efficient Algorithms for Globally Optimal Trajectories," *Transactions on Automatic Control*, 40, 1528–1538.

Tsivanidis, N. (2018): "The Aggregate And Distributional Effects Of Urban Transit Infrastructure: Evidence From Bogotá's TransMilenio," University of California, Berkeley, mimeograph.

United Nations (2018): *World Urbanization Prospects: The 2018 Revision*. United Nations, Department of Economic and Social Affairs, New York.

Wang, Z., S.-J. Wei, X. Yu, and K. Zhu (2018): "Re-Examining the Effects of Trading with China on Local Labor Markets: A Supply Chain Perspective," *NBER Working Paper*, 24886.

Waugh, M. (2010): "International Trade and Income Differences," *American Economic Review*, 100(5), 2093–2124.

Wilkerson, I. (2011): *The Warmth of Other Suns: The Epic Story of America's Great Migration*. Vintage, New York.

Wolf, N. (2007): "Endowments versus Market Potential: What Explains the Relocation of Industry after the Polish Reunification in 1918?," *Explorations in Economic History*, 44(1), 22–42.

Wuthnow, R. (2019): *The Left Behind: Decline and Rage in Small-Town America*. Princeton University Press, Princeton.

Zabreyko, P., A. Koshelev, M. Krasnoselskii, S. Mikhlin, L. Rakovshchik, and V. Stetísenko (1975): *Integral Equations: A Reference Text*. Noordhoff Int. Publ., Leiden, Netherlands.