

A Field Guide to Urban Economics

Matthew Turner

This version: October 23, 2025

Copyright Matthew Turner, 2025. All rights reserved. No part of this document may be reproduced, distributed, or transmitted in any form or by any means without the prior written permission of the author

Contents

1 The Monocentric City Model	1
1.1 The Realtor's mantra and spatial equilibrium	1
1.2 Land rent gradients in real life	6
1.3 The monocentric city model	10
1.3.1 The monocentric city model in two pictures	16
1.3.2 Three extensions	18
1.3.3 Comparative statics	24
1.3.4 Land rent and welfare	30
1.4 Application #1: Learning the value of school quality from real estate prices	33
1.5 Application #2: Detecting racist property tax assessments	36
1.5.1 Property taxes and rental prices	39
1.5.2 Land rent and capitalization	43
1.5.3 Fair assessment of property taxes	45
1.6 Conclusion	47
2 The Monocentric City Model vs. Data	55
2.1 Cities in real life	56
2.2 Rent gradients and Covid	61
2.3 Highways and decentralization	67
2.4 Highways and growth	72
2.5 Subways, decentralization and growth	74
2.6 Amenities and city size	77
2.7 Property taxes and land prices	79
2.8 Wages and rents	81
2.9 Conclusion	82
3 The Monocentric City Model with Housing	89
3.1 Population density gradients in real life	89
3.2 Households and housing	92

3.3	The construction sector	102
3.4	Conclusion	112
4	Urbanization in the Developed (mostly US) and Developing World	118
4.1	Urbanization in the US	119
4.2	Urban productivity premium	122
4.3	Excess urban mortality	127
4.4	The monocentric city model and urbanization in the developed world	131
4.5	Urbanization in the developing world	137
4.6	The Monocentric city model and rural amenities	156
4.7	Conclusion	166
5	White Flight, Gentrification, and Bid-Rent	169
5.1	Transportation costs and decentralization	172
5.2	The Great Migration and white flight	175
5.2.1	The American ghetto, 1890-1990	176
5.2.2	Causes of ghettos	182
5.3	Transportation costs, white Flight, and decentralization	191
5.4	Sorting and bid-rent	192
5.5	Conclusion	207
6	Quantitative Spatial Models and How Railroads Reorganized London	211
6.1	Distance in discrete space	213
6.2	The discrete choice problem	218
6.3	Extreme value distributions	226
6.4	A discrete linear city with heterogeneous households	231
6.5	Welfare	236
6.6	A discrete city with heterogeneous households and iceberg commute costs	240
6.7	The big prize: choosing discrete workplace <i>and</i> residence	242
6.8	Railroads, subways, and the economic geography of London, 1866-1921	245
6.9	Conclusion	257
6.10	Appendix: Proof of the discrete choice theorem	263
7	The Roback Model and the Value of Amenities	267
7.1	What is the value of amenities?	267
7.2	Amenities and productivity in the monocentric city model.	268
7.3	Roback Model	272
7.4	Roback Theorem	283

7.5	Comments	289
7.6	Application #1: Valuing climate	289
7.7	Application #2: Public finance in spatial equilibrium	292
7.8	Conclusion	296
8	Agglomeration Economies, or Why Are There Cities Anyway?	299
8.1	Why are there cities?	299
8.2	Returns to scale and a planner's problem	301
8.3	Returns to scale and equilibrium	304
8.4	How to measure agglomerations?	310
8.5	How to measure agglomeration economies?	317
8.6	Measuring agglomeration economies: Empirical results	320
8.7	Mechanisms	328
8.8	(Some of) what we know about mechanisms	330
8.9	Conclusion	339
9	Systems of Cities	344
9.1	Some basic facts about systems of cities #1	344
9.2	Zipf's law	351
9.3	Some basic facts about systems of cities #2	357
9.4	Path dependence and the locations of cities	361
9.5	Systems of cities	364
9.5.1	Spatial equilibrium	367
9.5.2	Planner's problem	374
9.5.3	Extension #1: Real estate developers	377
9.5.4	Extension #2: Allowing cities of size zero	378
9.6	Conclusion	379
10	Sorting, Voting with Your Feet, and a Simple Hedonic Model	385
10.1	Sorting versus causation: Why are people different in different places?	386
10.2	Spatial equilibrium and Tiebout sorting	405
10.3	The hedonic model	410
10.4	Conclusion	417

Chapter 1

The Monocentric City Model

1.1 The Realtor's mantra and spatial equilibrium

The late Lord Harold Samuel, a real estate tycoon in mid-20th century Britain, is reported to have said: “There are three things that matter in property: location, location, location.”¹ This sounds like a punchline, but it is important for two reasons. First, it is an expert’s observation about how the world works. Second, it highlights what is special about the economics of cities. You can’t study how cities are organized without thinking about where things are.

Consider two houses, house A and house B , alike in every detail except that the house B is downwind from a landfill and house A is not. Both are empty and available for rent, and there are many renters looking for houses in the neighborhood. The renters are all alike in the way they value both houses, and the landfill causes them all one dollar’s worth of unhappiness. What should happen?

¹For more detail on the origins of this saying, see William Safire’s June 26, 2009, column in the New York Times.

¹⁵ With many people looking for houses, both houses should end up rented. Moreover, the difference in their prices should be exactly equal to the one dollar of unhappiness that all renters suffer from being downwind from the landfill. This is exactly the Realtor's mantra. The difference in rent between the two houses is completely determined by their locations, downwind from a landfill and not.

²⁰ This argument just tells us that the rent for house *A* is a dollar more than for house *B*. It doesn't tell us the actual level of the rent for either house. How do we set the level of prices? It must be that the households in the two houses don't want to move to wherever the large pool of unsuccessful renters landed, some offstage "outside option", so the rent of house *A* should reflect the benefit from living in house ²⁵ *A* relative to this outside option, with the rent for house *B* a dollar less.

We have just worked out a simple model of spatial equilibrium for the housing market. In this equilibrium, identical people choose their favorite location from among the locations (and prices) available, and real estate prices adjust so that no one wants to move. An implication of this sort of equilibrium is that real estate prices are ³⁰ determined by the value of differences in the services that each location can provide to its occupant. That is, "location, location, location." As a starting point, this looks pretty good. The implications of our little economic model line up with the way the professionals think real estate markets should behave.

This model of how real estate markets work has an important implication for ³⁵ welfare. Which of the two successful renters is better off, the one downwind from the landfill and paying one dollar less rent, or the one without the nearby landfill? Suppose that the household near the landfill is worse off than the other. Recalling that the households are the same to start with, then this household has made a

mistake. They should have offered the landlord of the first house an extra penny
40 and moved into the house that is not near the landfill. At market rents, it must be that the two households are indifferent between the two houses. Lower rent must compensate the household downwind from the landfill for the unhappiness this causes or this household will move away. Even though the landfill affects just one of the two households, both are equally well off.

45 Suppose that, to promote environmental justice, you are asked to vote for a ballot initiative that will clean up the dump so that the downwind household can no longer notice it. If this ballot initiative is passed, who benefits from the cleanup? Absent the landfill, the resident of house B has an identical house to house A , but pays 1 dollar less in rent. This means that the resident of house A , or one of the many people who
50 did not find a place to live in the neighborhood, should bid up the price of house B to equal house A . This may take a while, the lease contract may run for a year, it may take people a while to figure out that the dump is cleaned up, but in a perfect world, this would happen pretty fast.

So who benefits from the dump clean up? Is it the household downwind from
55 the landfill? No. This household is indifferent between the two houses before the clean-up. After the clean-up, and rent adjustment, it is still indifferent between the two houses. The winner from the landfill clean-up is the landlord of the downwind house. Will this change your vote?

Now consider two more examples. Suppose that instead of being downwind from
60 a landfill, the second house is next to a gangster who collects one dollar in protection money every month, but is otherwise pleasant and unthreatening. This should operate much like being downwind from a landfill. The second house comes with a one dollar

monthly cost and the first does not, so the rent for the second house must be one dollar less. The real estate market should deal with both noxious neighbors in exactly
65 the same way.

Now for the interesting case. Suppose that the second house comes with a one dollar per month property tax bill, payable by the tenant, and the first does not. This has exactly the same implications for the resident of the second house as does the gangster neighbor; one dollar out of pocket each month. It should therefore have the
70 same implications for the real estate market. That is, the rent in the second house should be exactly one dollar lower than the first, and the two households should be indifferent between the two locations.

Why is this interesting? It means that the property tax (1) does not affect the welfare of the people who live in taxed houses, and (2) that the property tax does not
75 affect the total cost, rent paid to the landlord plus tax paid to the government, of a property. This is not how taxes on most other goods work, they usually raise prices at least a little.

This leads us to two bits of jargon. The first is easy. In all three examples, landfill,
gangster, and property tax, we say that real estate prices “capitalize” whatever is
80 special about the second house. That is, prices adjust to reflect differences in the value of the services provided by each location.

The second is “economic rent”, and it is a little slippery. Returning to the landfill example, we recall the household that occupies upwind house *A* is willing to pay an extra dollar of rent to avoid downwind house *B*. The need for jargon arises when we
85 replace the landfill with the gangster or tax payment. Here, the payment the tenant in house *B* makes to the landlord is one dollar less than that of the tenant in house

A, but the *total* payments for house *A* and *B* are the same. Without the landfill, the value that a tenant gets from each house is the same. The difference is that the landlord for house *A* receives the money equivalent of their tenant's value of living in the house, but the landlord for house *B* splits this value with the city government or the gangster. “Economic rent” is the value that a household gets from living in house *A* or house *B*. It is sometimes different than “contract rent”, what the tenant pays the landlord. In these examples, contract rent and economic rent coincide for house *A*, but economic rent is divided between a contract rent payment and a payment to the city government or the gangster in house *B*. From here on, “rent” always means “economic rent”, unless I explicitly note otherwise.

The object of this book is to understand the way people make the decisions that build and organize the cities where most of us live. The rest of this Chapter, and much of the rest of the book, revolves around applying the notion of spatial equilibrium that we have worked out here. That is, we ask what happens when people choose their favorite location from among the locations available and real estate prices adjust so that no one wants to move. However, instead of considering houses that are different from each other because of their proximity to gangsters and landfills, we consider houses that differ in their proximity to the center of a city where people work. This will give rise to one of the main theoretical tools that we have for thinking about the economic geography of cities, the “monocentric city model”.

1.2 Land rent gradients in real life

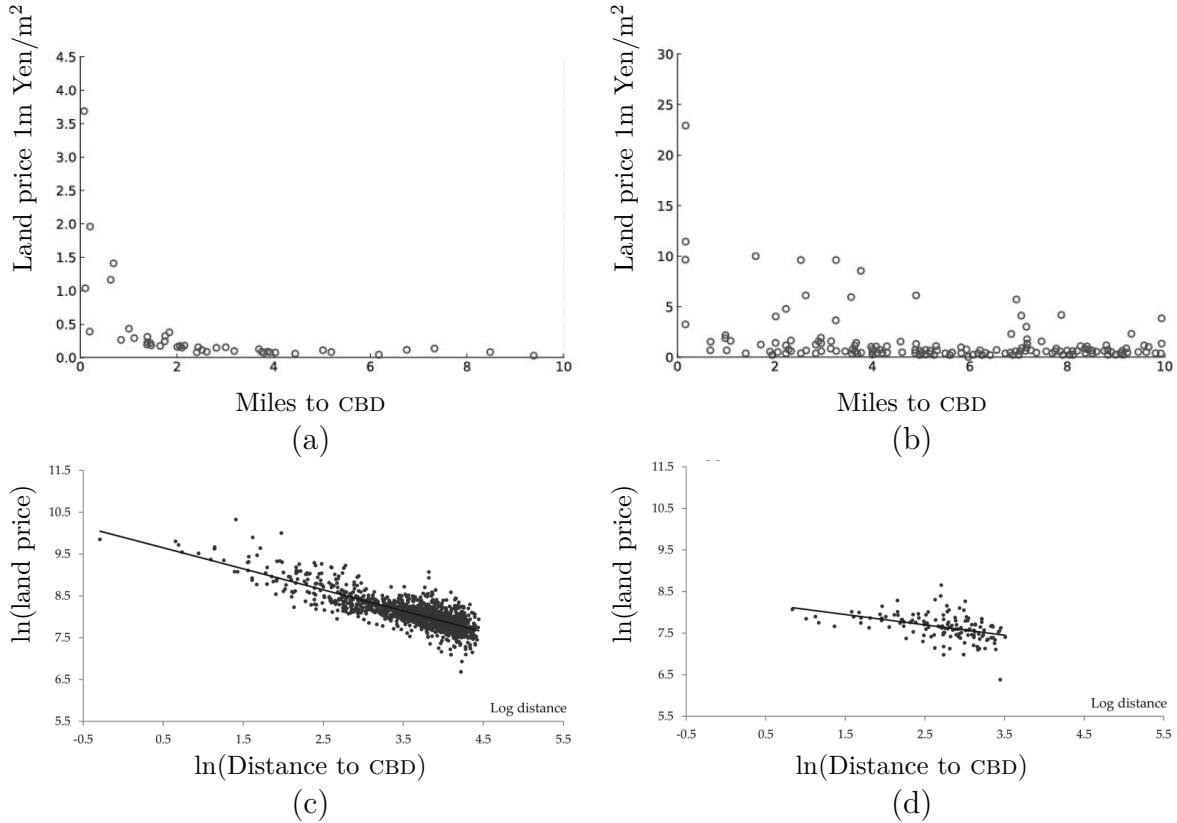
If we study the allocation of sugar donuts, we need just one price, the price of a sugar donut. But if we are studying the price of land, we need a price for each location.

¹¹⁰ If we think space is continuous, then we need a continuum of prices. A little more formally, in most of the rest of economics, prices are scalars. In urban economics, they are functions. We don't have a price of land, we have a land price function. This function assigns a price to each location. This land price function is often called a "land price gradient".

¹¹⁵ The goal for this book is to learn about the economics of cities, and land markets are central to this project. Figure 1.1 starts us off by describing land price gradients in two Japanese and two French cities. All four panels of figure 1.1 show how land prices change with distance to the city center. The two top panels show how land prices fall with distance from the city center for two cities in Japan in 1991, Hiratsuka and Yokohama. They fall fast. In panel (a) we see that a square meter of land near the center of Hiratsuka sells for 2 to 3 million Yen. A mile away, this price falls to half a million, and by 8 miles away, it has fallen to 100,000 or less. Panel (b) shows similar data for Yokohama, a much bigger city. Here, land near the city center sells for 10-20 million Yen per square meter, and shows the same rapid decline with distance to the center. Notice the mismatched units on both figures, metric on the y -axis and imperial on the x -axis.

¹²⁰ The bottom two panels of figure 1.1 differ from the top two in two ways. First, they are describing cities in France in 2012 instead of Japan in 1991. Second, both axes are in logarithms rather than levels. Panel (c) plots the logarithm of land prices in Paris against the logarithm of distance to the center. Panel (d) is the same, but

Figure 1.1: The relationship between land prices and distance to the center in four cities

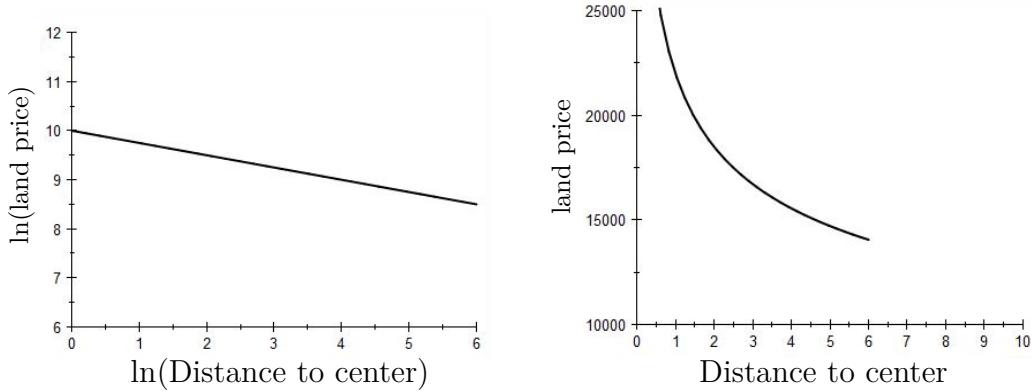


Note: (a) 1991 land prices in Hiratsuka, Japan; (b) 1991 land prices in Yokohama, Japan; (c) logarithm of 2012 land prices in Paris, France; (d) logarithm of 2012 land prices in Dijon, France. Figures (a,b) are based on figures from Lucas [2001] show how land rent declines (very fast) with radial distance from the center of two Japanese cities. Figures (c,d) from Combes et al. [2019] show the decline of the natural logarithm of rent with the logarithm of radial distance to the center. Panels (c) and (d) ©Oxford University Press.

for Dijon. These two figures also show a clear decline in land prices with distance to the center, but because the data is presented as logarithms, it is hard to tell if the decline is as fast as it is in the Japanese cities. For this, we need to do a little math.

Let R indicate the price of land and x be the radial distance from the center of the

Figure 1.2: Comparing plots in logarithms and levels



Note: *The left panel plots equation (1.5). The right panel plots equation (1.6). Both panels describe the same relationship between R and x , but the one on the left is in logarithms and the one on the right is in levels.*

¹³⁵ city. The Japanese figures plot R against x . The French figures plot $\ln R$ against $\ln x$.

To compare these two types of plots, you need to remember the rules for logarithms.

$$\ln R = A + B \ln x \quad (1.1)$$

$$\implies \ln R = \ln e^A + \ln x^B \quad (1.2)$$

$$\implies \ln R = \ln e^A x^B \quad (1.3)$$

$$\implies R = e^A x^B \quad (1.4)$$

Equation (1.1) describes the line plotted for the two French cities, the logarithm of land price against the logarithm of distance to the center. Equation (1.2) uses the fact that logarithms and exponentiation are inverses, that is, $\ln e^x = x$. Equation

¹⁴⁰ (1.3) uses a rule of logarithms, $\ln(x) + \ln(y) = \ln(xy)$. The last equation uses the fact that logarithms and exponentiation are inverses again. This last equation is the one

that is plotted for the Japanese cities, so we see that the two different looking pairs of graphs are actually plotting the same information, but represented differently.

Eye-balling the figure for Paris, we see that the intercept is about 10 and the slope 145 about $-\frac{1}{4}$. Writing this out, we have

$$\ln R = 10 - \frac{1}{4} \ln x. \quad (1.5)$$

We can rearrange this to get

$$R = e^{10} x^{-\frac{1}{4}} \approx 22,000 x^{-\frac{1}{4}}. \quad (1.6)$$

The left panel of figure 1.2 plots the first of these equations. The right panel plots the second. Once we convert the French data from logarithms to levels, we see the same rapid, radial decline in land prices that we see for Japanese cities.

150 So, land rent behaves the same way in France as it does in Japan. This is pretty neat. It did not have to be true. In fact, cities almost everywhere show this sort of log-linear decline in rent with distance to the center.

Now, two asides. First, economists often find the world is well described by log-linear relationships like the ones illustrated for the French cities in figure 1.2, so it's 155 worth learning how to go back and forth between logarithms and levels (as we've just done). Second, log-linear relationships have another advantage. The coefficient B on $\ln x$ in equation (1.1) is an elasticity. It tells us the percentage change in rent that results from a one percent change in distance (see box 1.2.1). Elasticities are handy because you don't need to keep track of the units that you use to measure x and 160 R , or whatever variables you are interested in. You can use meters to describe your

y -axis and miles for the x -axis and, if you plot your data in logarithms, you won't get caught.

1.3 The monocentric city model

We have so far established two ideas. First, that the rental prices of properties ought to adjust to reflect differences in the value of living at the properties in such a way that no one wants to move. Second, the price of land falls rapidly as we move away from the center of cities.

The monocentric city model explains the second fact as a consequence of the first. That is, it assumes the price of real estate changes in order to keep people indifferent between all available locations, just as in our example with house A and house B , and uses this assumption to explain the decrease in real estate prices as we get further from the center. With our example of the landfill still in mind, you can guess how this is going to work. For prices to fall with distance from the center, something about more remote locations has to get worse. In the monocentric city model, the thing that gets worse with distance to the center is the cost of commuting to work. This is the central intuition of the model; land prices fall with remoteness from the center to exactly compensate for a more costly commute.

Before we develop the model, two comments. First, the data on land prices presented in section 1.2 describe the price of land, how much you have to pay to obtain the services of the land forever. In our examples, we've considered the rental price of land, what you have to pay to obtain the services of the land for some definite fixed time. For now, let's just name these two Asset Prices and Rental Prices, respectively,

and note that although they are not the same, they are close relatives. We'll work out the relationship between them later, but for now, you can think of them as synonyms.

¹⁸⁵ Second, in order to talk about prices (or rents) declining with distance from the center, we need to locate the center. It turns out that this is a surprisingly well defined concept. Ask yourself, and two or three other people, where the center of your hometown is. You will almost surely get the same answer from everyone. In Providence, where I am writing this Chapter, it is the plaza across from city hall.
¹⁹⁰ We will refer to this central location and the area around it as the “Central Business District”, or CBD.

Imagine a city located on a featureless plane or along a line. We begin with a linear city because it is a little simpler. Indicate locations on the line with x . There is a CBD located at $x = 0$ and $|x|$ is distance to the CBD, with $x < 0$ for a location ¹⁹⁵ to the left of the CBD and conversely. There is one unit of land at each x .² The city is populated by identical households (or workers), all of whom commute to the CBD where they earn wage w . Commuting costs t per unit distance, and so a household living at x pays $2t|x|$ in commuting costs. All households occupy a parcel of fixed size, \bar{l} , at whatever location x they choose. Households use their wage to pay the ²⁰⁰ land rent for their parcel, $R(x)\bar{l}$, to purchase a composite consumption good, c , and to pay the cost of commuting, $2t|x|$. Households derive utility from the consumption good according to the utility function $u(c)$, and we require that u is increasing, or equivalently, that $u' > 0$.

Land not used for urban residences remains in agricultural use and a farmer is

²This is a little bit fishy but let us avoid some arcane math. How can there be one unit of land at a point on a line? Strictly, and if you know some probability theory, there is a uniform “density” of land.

205 always willing to pay a reservation land rent of \bar{R} , the “agricultural land rent”. The total population of the city is N and the price of the consumption good, c , is set to $p_c = 1$ (so we don’t need to keep track of it). Finally, all land rent is collected by “absentee landlords”. This is an important bit of fiction. It means that all land rent leaves the model and we don’t need worry about the messy problem of keeping track 210 of who gets to spend it.

We make two assumptions about how households behave. First, that they choose their location x so as to make themselves as well off as possible. That is, households solve,

$$v(c, x) = \max_{c, x} u(c) \quad (1.10)$$

$$\text{s.t. } w = c + R(x)\bar{\ell} + 2t|x|.$$

215 This says that households choose their residential location x to maximize their consumption, subject to their budget. This is the standard assumption about rationality in economics: People make themselves as well off as they can given the choices available to them. As long as u is an increasing function this maximization problem is trivial; use everything left from the wage after paying for rent and commuting to buy the composite consumption good. No calculus required.

220 The second assumption we make is that no one wants to move. Formally, we require that households get the same utility at every x , and call this utility level \bar{u} . That is,

$$v(c, x) = \bar{u} \text{ at all occupied } x. \quad (1.11)$$

This is sometimes called a “free mobility condition”. If it is free to move, we expect people to move for any tiny improvement in their utility, and so utility must be the
225 same everywhere. We often call \bar{u} a “reservation utility level”.

This is the complete statement of the model.

Before we work out the implications of these assumptions, three comments are in order. First, as anyone who has ever moved knows, moving is not free, and so starting from an assumption of “free-mobility” seems inauspicious. But the assumption is
230 better than it seems. Consider someone who has already decided to move to a new city. All of their stuff is in the mail or on a truck. For this person, choosing a house on one block or another is essentially free. Moreover, this is one of the people who is actually participating in the market and helping to set prices. This is what the free mobility assumption really requires. Not that everyone can move costlessly, but that
235 the subset of households who are buying or renting properties can do so. This seems much easier to defend.

Second, the monocentric city model comes in two varieties, “open city” and “closed city”. In an open city model, city population adjusts until the free mobility condition is satisfied and everyone in the city is indifferent between all locations in the city and
240 the outside option. In a closed city model, we fix the population of the city and let the utility level, \bar{u} , adjust so that households are indifferent between all locations in the city. These are stylized cases, and reality probably lies somewhere between.

Finally, according to data assembled by the United Nations Population Division,
the share of the world’s population that lives in cities from 1960 until 2020 rose from
245 about 34% to about 56%, even as the world’s population is increasing. As a starting point, we probably want to think about a model where the population of a city can

adjust rather than one where it cannot. Open city models are also a little easier to work with, and so we'll start with this case.

We now turn to working out the implications of the monocentric city model. We
 250 would like to see what it implies about the extent of the city, its population, the land rent gradient, and the welfare of its residents.

To begin, invert the free mobility condition, equation (1.11), to find the level of consumption that households require to reach the reservation level of utility,

$$c^* = u^{-1}(\bar{u}). \quad (1.12)$$

For example, if $u(c) = c^{1/2}$ and $\bar{u} = 2$, then $(c^*)^{1/2} = 2$, so $c^* = 2^2 = 4$.

In a spatial equilibrium, everyone gets the same utility at every location, so consumption must be the same everywhere. Therefore,
 255

$$w - c^* = R(x)\bar{\ell} + 2t|x|, \quad (1.13)$$

for all locations x . With wages and consumption fixed for all households, commuting costs and land rent must vary in such a way that they always sum to a constant. Implicitly, we're also assuming $w > c^*$. Otherwise, no one would live in the city at
 260 all.

We can use equation (1.13) to find the extent of an equilibrium city. The limits of the city are defined by the most remote points where a city resident values the land more highly than a farmer. That is, where a city resident is just willing to pay the reservation land rent \bar{R} . Let \bar{x} denote the distance of these boundary points from

²⁶⁵ $x = 0$. At this location, we must have

$$w - c^* = \bar{R}\ell + 2t|\bar{x}|.$$

At the edge of the city, the cost to commute is such that a household can just pay the reservation land rent and commuting costs, and still buy the reservation consumption bundle. Reorganizing, we have that

$$\bar{x} = \frac{w - c^* - \bar{R}\ell}{2t},$$

is the most remote occupied point to the right of $x = 0$, and the city extends from

²⁷⁰ $-\bar{x}$ to $+\bar{x}$.

We see here how the open city assumption works. A household must obtain a utility level of at least \bar{u} to live in the city. This means consumption of at least c^* . The price of land at the unoccupied location nearest the CBD must be \bar{R} , so the price of the best unoccupied parcel is $\bar{R}\ell$. This means that the most remote occupied location is one where a household can just afford c^* after bidding land away from a farmer and paying for their commute. In this sense, the city is “open”, its size is determined by the number of people who choose to live there.

Because the city extends from $-\bar{x}$ to \bar{x} and each household consumes an exogenously fixed amount of land, the population of the city is,

$$N^* = \frac{2\bar{x}}{\ell}.$$

²⁸⁰ Note that we are here using the assumption that there is one unit of land at each x .

Using the equilibrium budget constraint, equation (1.13), and the equilibrium extent of the city, we can solve for the equilibrium rent gradient,

$$R^*(x) = \begin{cases} \frac{w - c^* - 2t|x|}{\bar{\ell}} & \text{if } |x| \leq \bar{x} \\ \bar{R} & \text{if } |x| > \bar{x}. \end{cases} \quad (1.14)$$

If we restrict attention to locations in urban use, then this is just

$$R^*(x) = \frac{w - c^* - 2t|x|}{\bar{\ell}}. \quad (1.15)$$

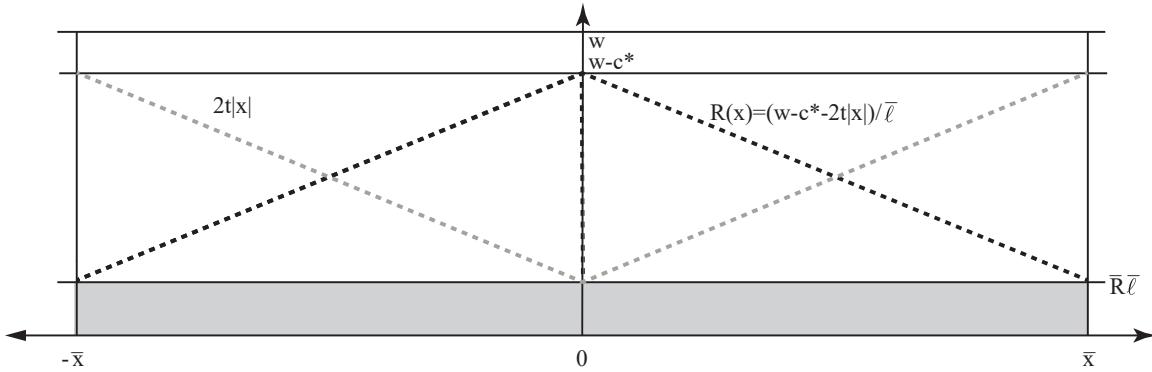
This is a bit simpler to write out, and I'll often cheat and write it this way.

285 This finishes the solution of the model. Box 1.3.1 works out an example. Assuming that households optimize and that no one wants to move, and given a reservation utility level, a price of commuting, and a wage for work in the central location, we've derived the size of the city and its configuration, along with the land rent gradient. Our next step is to present the same argument graphically. This is a little easier and 290 makes the intuition clearer. After that we want to think about some extensions of the model that make it a little more realistic, and to consider its other implications. So far, the model is able to predict the downward sloping rent gradient we observe. we'd like to have more predictions to check.

1.3.1 The monocentric city model in two pictures

295 The monocentric city model has a tidy graphical representation, given in figure 1.3. Let the x -axis indicate displacement away from the CBD at $x = 0$, and let the y -axis indicate units of consumption. Because w , c^* , \bar{R} and $\bar{\ell}$ are the same at all locations,

Figure 1.3: The monocentric city model in one easy picture



Note: *Illustration of the monocentric city model. x-axis is displacement from the CBD at $x = 0$ and y-axis is units of consumption. The wage, value of land in agriculture and household consumption level are all constant across all locations x . The gray dashed lines show how commute costs increase with distance from the CBD and the black dashed lines show how land rent decreases. The edges of the city occur where urban residents can no longer afford to commute and still outbid farmers for land.*

we can draw three horizontal lines, the first for the wage, at height w , the second for the wage net of consumption, at $w - c^*$, and the third at the value of land in agriculture, $\bar{R}\bar{\ell}$.

Looking again at the budget constraint, we have

$$w - c^* = R(x)\bar{\ell} + 2t|x|.$$

That is, every household gets the same wage and enjoys the same level of consumption. Once they have paid for this consumption, the rest of their earnings go to land rent for their parcel of size $\bar{\ell}$ and to the cost of commuting.

A household living right at the CBD, at $x = 0$, doesn't commute and so has zero

commute expenditure. In order for this household to satisfy their budget constraint, the rest of their earnings, $w - c^*$, goes to land rent. It follows that $R(0) = (w - c^*)/\bar{\ell}$ and that the $x = 0$ household's total expenditure on land is $R(0)\bar{\ell} = (w - c^*)$. As we move away from $x = 0$, commute costs increase linearly, at a rate of $2t$ per unit distance. This gives us the dashed gray commute cost gradient. Expenditure on land decreases by an exactly offsetting amount. This gives us the dashed black land rent gradient. The edge of the city occurs at a distance \bar{x} from the CBD where, once a household pays for their commute, they have just enough left over to bid a parcel away from the farmers.

Early in the industrial revolution many cities were “mill towns”. There was some big concentration of employment at the center, a mill, a collection of mills, a port or a railway depot. All the workers lived nearby and walked or took the train back and forth to the center. This is just the situation that the monocentric city model is meant to describe, and figure 1.4 shows that this is just how 19th century Providence was organized. Employment was highly centralized in the center, and there was no way to get back and forth to the CBD except on foot or by train.

1.3.2 Three extensions

We now consider three extensions to the basic model. First, we consider a closed city equilibrium. Although we will work primarily with the open city model, most current research is based on models of closed cities, so this is an important case to work out. Second, we consider a circular city. The linear city assumption is obviously silly. We want to work out the circular city model to demonstrate that the extra realism doesn't actually change anything beyond making our math a little messier.

Figure 1.4: The geography of Providence around 1896.



Note: *View of the city of Providence as seen from the dome of the new State House.*

Drawn by M. D. Mason, published in the Providence Sunday Journal, Nov. 15, 1896.

Figure courtesy of the Library of Congress, Geography and Map Division.

Finally, we want to introduce the idea of “amenities” to our model. Amenities, here

some feature of the city that affects utility directly, like sunshine or pollution, are important for determining city size in reality, and will play an important role in much of what we talk about later in the book.

Closed city equilibrium

The “closed city” version of the monocentric city model is exactly the same as the
 335 “open city” version we have worked out, except that instead of knowing the value for
 the reservation utility, \bar{u} , we know the number of people who live in the city, N .

Given population size N , because everyone consumes $\bar{\ell}$ of land, the length of the city must be $N\bar{\ell}$, and so the edges of the city must be at $|\bar{x}| = N\bar{\ell}/2$.

Once we know the most remote occupied locations, we can figure out consumption
 340 by requiring that rent at the edge of the city equal agricultural rent,

$$R(\bar{x}) = \frac{w - c^* - 2t|\bar{x}|}{\bar{\ell}} = \bar{R}.$$

Rearranging and substituting for \bar{x} , we have

$$c^* = w - \bar{R}\bar{\ell} - tN\bar{\ell}.$$

In equilibrium, consumption must be the same at all occupied locations. That is,

$$c^* = w - R(x)\bar{\ell} - 2t|x|.$$

Equating these two expression for c^* and solving for $R(x)$, we have

$$R(x) = \frac{(\bar{R} + tN)\bar{\ell} - 2t|x|}{\bar{\ell}}.$$

This is complicated looking, but the intuition is the same as what we have already
 345 done. Rent adjusts so that income net of commuting and rent is the same everywhere.

The difference is that in an open city equilibrium, the reservation utility is exogenous, and the population of the city adjusts until the marginal household pays just the agricultural rent. With a closed city equilibrium, population is fixed and utility adjusts.

³⁵⁰ **Circular city**

Suppose we relax the (silly) assumption that the city is on a line, and think about a circular city that is symmetric around a central point, still on a featureless plane, keeping everything else the same.

This barely changes the household's problem at all. We still have

$$\begin{aligned} & \max_{c,x} u(c) \\ \text{s.t. } & w = c + R(x)\bar{\ell} + 2tx \end{aligned}$$

³⁵⁵ and, in an open city, $u(c^*) = \bar{u}$.

Although this problem looks the same as the one for the linear city. It is slightly different. In this case, x is radial distance to the center, in whatever direction, and can only be positive. In contrast, for a linear city, x and $-x$ refer to particular coordinates on the line. Practically, this means that we don't need the absolute value on x to state the circular city problem, and we need to keep in mind that x refers to all of the locations at distance x from the CBD, rather than to a particular point.
³⁶⁰

Consumption must still be the same everywhere in a spatial equilibrium,

$$w - c^* = R(x)\bar{\ell} + 2tx.$$

Let \bar{x} denote the distance from the origin to the most remote occupied location. At this distance from the CBD, we must have

$$w - c^* = \overline{R\ell} + 2t\bar{x}.$$

³⁶⁵ Reorganizing, we have

$$\bar{x} = \frac{w - c^* - \overline{R\ell}}{2t}.$$

This is the same as for the linear city, but it is now on the edge of a circular city.

The area of our circular city is $\pi\bar{x}^2$, so population is

$$N^* = \frac{\pi\bar{x}^2}{\bar{\ell}}.$$

This is the big (and completely obvious) difference between the linear and circular city. With the linear city, the extent of the city increases at the same rate as the population. With a circular city, area increases much faster than radius, so for a given increase in the radius of the city, a circular city accommodates a lot more people than does a linear city.

Amenities

Suppose our city has an amenity A that affects the utility of residents. This could be ³⁷⁵ something like good or bad weather, crime, pollution, or parks. How does this affect equilibrium?

To illustrate ideas as simply as possible, suppose a household's utility is $u(Ac)$,

almost just as before. So, $A > 1$ is something good, $A < 1$ is something bad. How does this change the open city equilibrium? With an open city, we have

$$u(Ac^{**}) = \bar{u}.$$

³⁸⁰ Reorganizing, we have,

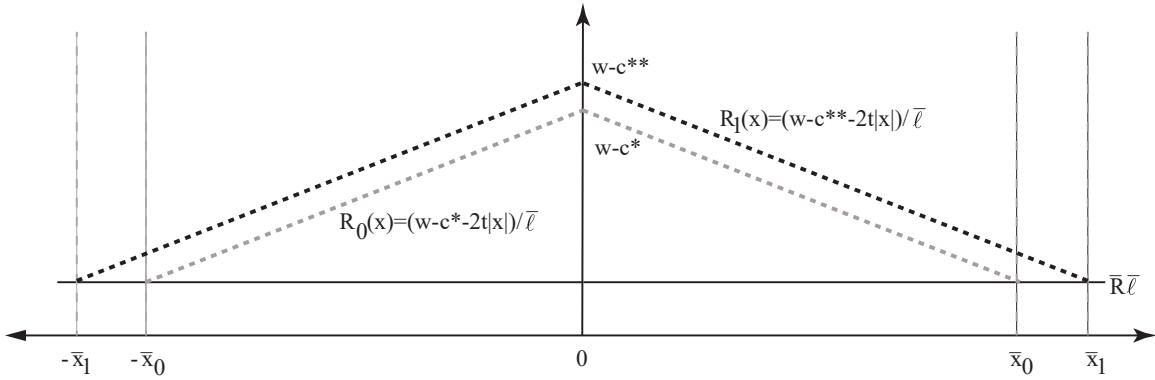
$$c^{**} = \frac{1}{A}u^{-1}(\bar{u})$$

If $A = 1$ we get back the basic case we've already covered and $c^{**} = c^*$. If $A > 1$, then $c^{**} < c^*$. If a city has an amenity that contributes to utility then households can attain their reservation utility level at lower levels of consumption. That is, people accept less consumption to get better weather. Nothing else about the model changes.

³⁸⁵ How does this affect the equilibrium city? As A increases and amenities get better, then; (1) equilibrium consumption falls, (2) the rent gradient intercept increases so rent goes up everywhere, and (3) the city grows in extent and population. This is illustrated in figure 1.3.

Sunny cities should be bigger and have higher rent than snowy ones, and the people ³⁹⁰ in sunny cities should also consume a little less than people in snowy cities. The people in sunny cities are achieving the reservation utility level partly with sunshine and partly with consumption. The people in snowy cities must rely more heavily on consumption.

Figure 1.5: The monocentric city with amenities



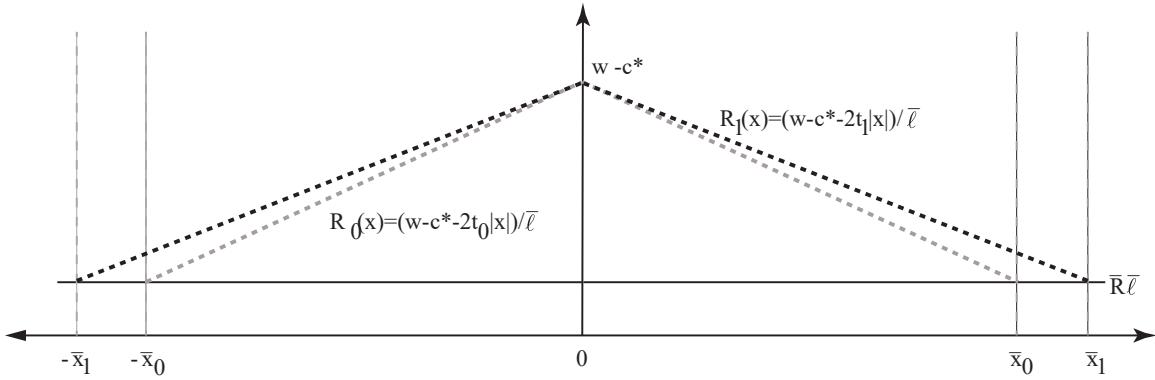
Note: The dashed gray line illustrates the land rent gradient for the baseline case when there are no amenities, really when $A = 1$. The dashed black line illustrates the land rent gradient when the city has some amenity that complements ordinary consumption, $A > 1$. In the city with the amenity, rent goes up everywhere and the city expands. With the amenity, it is possible to hit the reservation utility level with a little less consumption. This leaves more income to be divided between land rent and commuting.

1.3.3 Comparative statics

³⁹⁵ We have a model that assumes: transportation is costly, everyone wants to work in the center, people arrange themselves so that no one wants to move, i.e., spatial equilibrium. These assumptions imply the downward sloping rent gradient that characterizes land markets almost everywhere, and that figure 1.1 illustrates for Japan and France.

⁴⁰⁰ It would also be nice to work out whether the model makes other predictions we can check. If the model makes predictions that are obviously not in line with reality, we will know we have a problem. With that in mind, we now consider some “comparative statics”. That is, we ask how the monocentric city changes as we change, for example, wages or commuting costs while holding everything else fixed (we already

Figure 1.6: Monocentric city comparative statics as commuting costs change



Note: The dashed gray line describes an equilibrium land rent gradient in an open city when commuting costs are high, and the dashed black line describes an equilibrium land rent gradient in the same city when commuting costs fall. That is, $t_1 < t_0$. As commuting costs fall, land rent increases everywhere except at the CBD where the household's commute has length zero. As commuting becomes less expensive, at each location, the household has more income to divide between consumption and land rent. But the level of consumption is fixed by the reservation utility level. This means that the only thing that can adjust is the price of land. As land rent increases, it means that households can afford to bid a few marginal parcels away from farmers, and the edge of the city moves out from the CBD.

⁴⁰⁵ looked at what happens when amenities change). Once this is done, we will have a collection of predictions for the model that we can compare to what we observe in the real world.

Changes in an open city as commuting cost, t , changes

If commuting costs fall, then utility and consumption stay the same in an open city.

⁴¹⁰ The household at $x = 0$ has a free commute, so its commute cost is unaffected by the change in t , but commute costs fall for households a little further from the CBD. Because the sum of land rent and commute cost is the same at all locations, this

implies that the land rent gradient flattens, but its intercept stays the same. The
 415 only way consumption can remain constant at the reservation level, c^* , as commute cost falls is if land rent increases. At the edge of the city, land rent goes up and a few more households bid land away from farmers so the extent of the city increases. Because land rent increases everywhere, the total land rent paid to absentee landlords increases. Later on, we'll consider empirical evidence about this comparative static. For that purpose, it is helpful to note that as unit commute costs fall, a larger share
 420 of people live outside any fixed radius. This is all illustrated in figure 1.6

We can derive all of this analytically using partial derivatives (see box 1.3.2 if you need help with this). To proceed, take the partial derivative of the land rent gradient with respect to the commute cost,

$$\begin{aligned}\frac{\partial R(x)}{\partial t} &= \frac{\partial}{\partial t} \frac{w - c^* - 2t|x|}{\bar{\ell}} \\ &= \frac{-2|x|}{\bar{\ell}}\end{aligned}$$

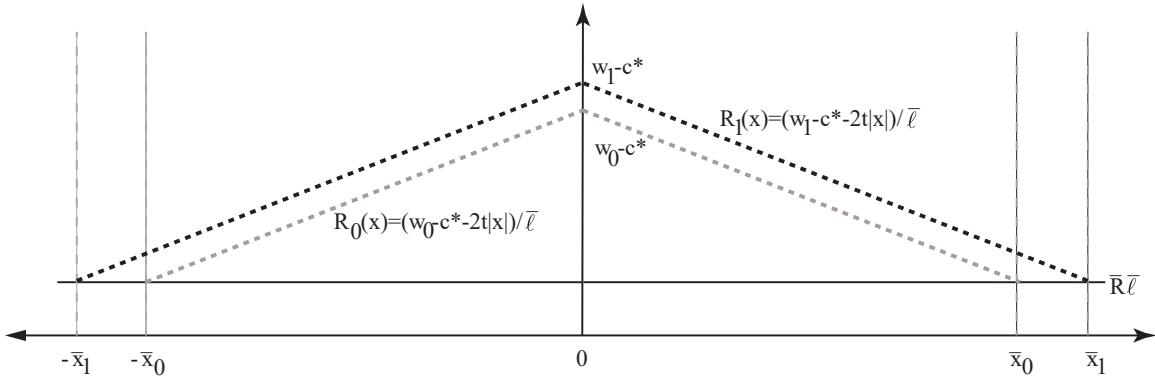
This derivative is negative, so as t increases, rent falls at each x , and conversely (we're
 425 ignoring the corner where $R = \bar{R}$.) We can do exactly the same thing to see what happens to the extent of the city as commute costs change,

$$\begin{aligned}\frac{\partial \bar{x}}{\partial t} &= \frac{\partial}{\partial t} \frac{w - c^* - \bar{R}\ell}{2t} \\ &= -\frac{w - c^* - \bar{R}\ell}{2t^2} < 0.\end{aligned}$$

This derivative is negative, too, so as t increases, the length of the city falls.

Finally, we can check what happens to city population as commute costs change.

Figure 1.7: Monocentric city comparative statics as the wage changes



Note: The dashed gray line shows an equilibrium land rent gradient for a low wage, and the dashed black line for a higher wage. As the wage increases in an open city, households everywhere see an equal increase in their income. Because consumption must stay constant, and because commute costs don't change, the only way to make sure the household budget balances if is land rent goes up by an amount that exactly offsets the wage increase. Because the value of land in residential use goes up everywhere, the extent of the city increases a little bit as a few more households are able to bid land away from farmers at the edge of the city.

Because $N = 2\bar{x}/\bar{\ell}$, we can use the chain rule to get,

$$\begin{aligned}\frac{\partial N}{\partial t} &= \frac{\partial}{\partial t} \frac{2\bar{x}}{\bar{\ell}} \\ &= \frac{2}{\bar{\ell}} \frac{\partial \bar{x}}{\partial t}\end{aligned}$$

⁴³⁰ and so the population of the city changes just like its extent. Thus, we obtain the same results analytically that we see in figure 1.6.

Changes in an open city as the wage changes

Figure 1.7 shows changes as wages increase in an open monocentric city. As wages rise, utility and consumption stay the same. This follows immediately from the
⁴³⁵ open city assumption and spatial equilibrium. The slope of the land rent gradient is determined by the unit commute cost, and this also remains fixed. The intercept of the land rent gradient increases by exactly the amount of the wage increase, $w_1 - w_0$. This is the only way that we can balance the budget for households at $x = 0$ and keep consumption constant. The same increase has to occur everywhere, and for the same
⁴⁴⁰ reason. Thus, an increase in the wage gives us a parallel shift up in the land rent gradient that offsets the wage increase. As a result, the extent of the city increases a little bit and population increases. Aggregate land rent increases by almost the exact amount as the total wage bill.

It bears repeating that almost all of the benefit of an increase in wages is collected by absentee landlords. Even though wages go up, residents' consumption is unchanged, so household budgets at any given location x can only balance if the wage increase is passed directly on to the landlords.
⁴⁴⁵

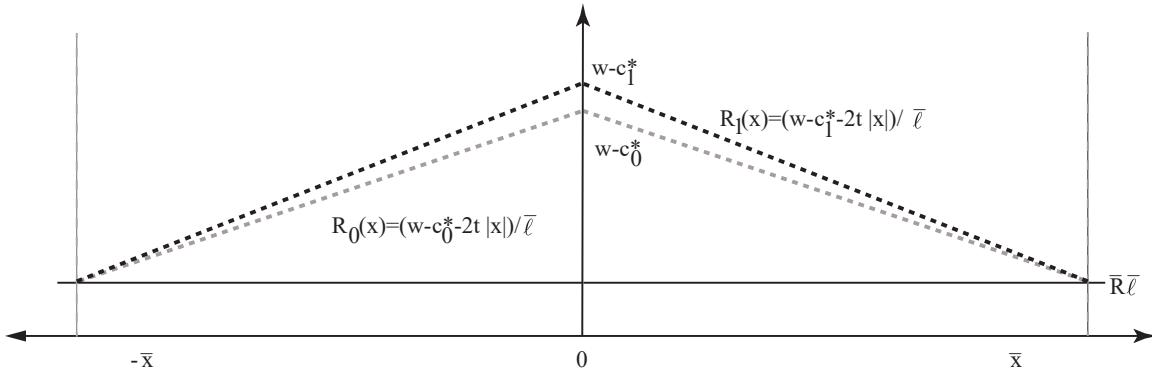
At the time of this writing, there is a lot of policy interest in the US about increases to the local minimum wage. What does this comparative static suggest about the
⁴⁵⁰ likely winner from these policies?

Changes in a closed city as commuting costs t change

Comparative statics in closed cities are quite different from those in open cities. Figure 1.8 illustrates what happens in a closed city when the cost of commuting increases.

Three things must stay fixed in a closed city as the cost of commuting increases.

Figure 1.8: Monocentric city comparative statics as commuting costs changes in a closed city.



Note: *The dashed gray line shows the equilibrium land rent gradient in a closed city with low commute costs, and the dashed black line shows the land rent gradient in the same city when commute costs increase. In a closed city, the population is fixed, this fixes the extent of the city. With the extent of the city fixed, the land rent gradient must adjust so that the most remote household can just bid land way from farmers. As commute costs rise, land rent must increase in order to keep the sum of land rent and commute costs constant. In the closed city, it is the rent at the most remote location that is fixed by our assumptions. In contrast, for an open city, the reservation utility level fixed the level of land rent at $x = 0$.*

- ⁴⁵⁵ First, the size of the city cannot change. It is fixed by the fact that the number of people and per capita land consumption are both fixed. Second, land rent at the edge of the city cannot change because the rent required to bid land away from farmers also doesn't change. Finally, in a spatial equilibrium, the sum of land rent and commuting must be the same at all occupied locations, otherwise consumption differs across places and we don't have an equilibrium.
- ⁴⁶⁰

How can we satisfy these three conditions as the unit cost of commuting rises? If commute costs increase, the rent gradient must get steeper to keep the sum of rent

and commuting constant. In addition, land rent must stay constant at \bar{x} . Together, this means that rent at $x < \bar{x}$ must increase, and in particular, that the rent at $x = 0$
⁴⁶⁵ goes up. Because income is fixed, the increase in rent and commute costs means that consumption and utility falls. This is quite different from the open city where the utility of households is fixed, and changes end up affecting land rent without affecting household utility.

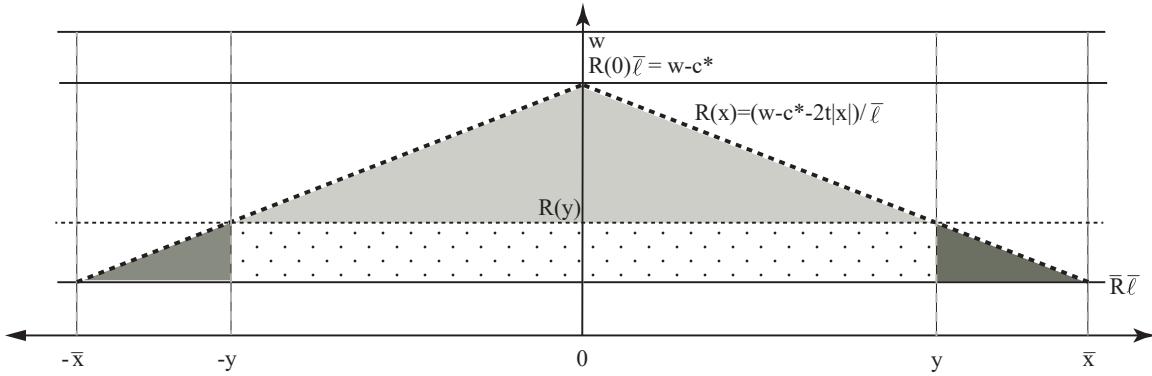
In an open city, the supply of people is perfectly elastic. All changes fall on
⁴⁷⁰ landlords, good or bad. With a closed city, the supply of people is perfectly inelastic, so some of the change in commuting cost falls on the households. This highlights the importance of knowing how responsive is migration to local economic conditions for understanding the distributive consequences of urban policy. If we want to change something in a closed city, it affects the welfare of residents. In an open city, the
⁴⁷⁵ welfare of residents is fixed, and payments to landlords change.

1.3.4 Land rent and welfare

In an open city equilibrium, each household gets $u(c^*) = \bar{u}$, and they get this payoff no matter how much rent they pay. In this sense, land rent is a measure of the benefit to a household from living in the city. They can get payoff \bar{u} in the reservation location.
⁴⁸⁰ In the city, they get this payoff and manage to pay land rent in addition. This suggests that aggregate land rent, the sum of land rent paid by all urban residents, is a measure of welfare. It is the collective willingness to pay to live in the city. It follows immediately, that changes in land rent indicate changes in welfare.

This is an important conclusion. Land rent is relatively easy to observe, much
⁴⁸⁵ easier than utility levels, and so the fact that land rent measures welfare gives us a

Figure 1.9: Aggregate land rent in the monocentric city



Note: The dashed black line describes a land rent gradient for an open city. \bar{x} is the edge of the city in equilibrium. A planner would like to choose an extent of the city to maximize aggregate land rent, taking as given that the city is open and households must satisfy the free mobility condition. If the planner chooses an extent of the city smaller than equilibrium, $y < \bar{x}$, then aggregate rent is less than for a city with edges at \bar{x} . In particular, the rent described by the two dark gray triangles is lost. If the planner chooses an extent greater than \bar{x} , then the planner must subsidize the marginal households to allow them to bid land away from farmers and still consume c^* .

way to use easily observable data to think about the welfare implications of changes in the urban environment. Drawing conclusions about welfare from things that are easy to observe is not something that economists get to do very often. Indeed, in the next section we will show how we can use this intuition to value school quality or
⁴⁹⁰ other place based attributes using data on real estate prices. There is an important caveat to this conclusion; it starts to break down once we start to think about models where not all households are the same. We'll return to this problem in Chapter 6.

Now that we have a way of measuring welfare in a city, it is natural to ask whether the equilibrium city maximizes welfare. A little more precisely, we have described
⁴⁹⁵ how the monocentric city arises as an equilibrium outcome when everyone pursues

their own narrow self interest. What would happen if a rent maximizing planner organized the city, subject to free mobility for the households? Would the resulting rent maximizing city be different from an equilibrium city?³

In our optimal city, we still allow free mobility, so, as in the equilibrium city, we

500 must have

$$w - c^* = R^*(x)\bar{\ell} + 2t|x|.$$

Rearranging, we have

$$R^*(x) = \frac{w - c^* - 2t|x|}{\bar{\ell}}$$

at all occupied locations. Given this, our planner wants to choose the extent of the city, y to maximize total land rent, taking as given that rent is given by this expression at any location in urban use.

505 To avoid an involved calculus problem, figure 1.9 makes the argument graphically. This figure is like those we have used throughout the Chapter to illustrate the monocentric city model. In particular, \bar{x} is the equilibrium extent of the city. We would like to consider whether a planner choosing the extent of the city to maximize aggregate rent would choose something different.

510 When the planner chooses \bar{x} as the extent of the city, then aggregate rent is just what we would have for the equilibrium city. It is the sum of the light gray, dark gray and dotted regions under the land rent gradient. Suppose our planner chooses a

³You might recognize the parallel between this question and the one answered by the first fundamental theorem of welfare economics. This theorem states that, under weak conditions, if a market equilibrium exists then it is Pareto optimal. We will find something similar here.

slightly smaller extent for the city, $y < \bar{x}$? Then aggregate land rent is just the area under the land rent gradient between $-y$ and y . This is the sum of the light gray and 515 dotted regions. This is clearly less than the aggregate land rent that results if the planner chooses \bar{x} as the extent of the city. Now what if the planner chooses $y > \bar{x}$. Then the marginal increase in urban land rent does not offset foregone agricultural land rent. In fact, the planner has to subsidize the marginal urban resident in order to allow them to bid land away from a farmer and still afford enough consumption 520 that they don't want to move away. It follows that the monocentric city that emerges in equilibrium is "optimal" in the sense that it maximizes land rent.⁴

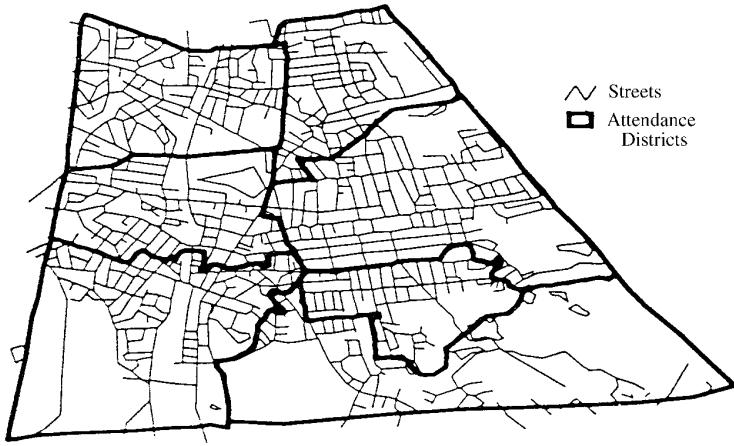
1.4 Application #1: Learning the value of school quality from real estate prices

An interesting implication of the monocentric city model is that land rent can never be 525 discontinuous. To see this, imagine the rent gradient drops discontinuously as we move away from the CBD. In this case, the household at the high side of the discontinuity can move to the low side, experience almost zero change in commute costs, and a discrete drop in rent. This contradicts the idea that this was an equilibrium to start with. A household can move and make themselves better off. A similar argument 530 works if there is a discontinuous increase in land rent.

This means that we can have a discontinuous rent gradient only if amenities vary discontinuously. In this case, spatial equilibrium requires that rent vary discontinu-

⁴This is slightly weaker than the first welfare theorem because rent maximization is implied by Pareto optimality, but not conversely.

Figure 1.10: School district boundaries in Melrose Massachusetts around 1990



Note: *Heavy black lines show school attendance zone boundaries in Melrose Massachusetts around 1990. Lighter lines are streets.* Reproduced from Black [1999], ©Oxford University Press.

ously in order to equalize utility across locations. This intuition motivates the “border discontinuity design” for learning about the value of amenities that vary discretely as we move across the landscape.

Black [1999] uses this idea to examine the value of school quality by looking at how housing prices vary when we cross a school district boundary where school quality varies. She considers the relationship between school quality and real estate prices for three counties in Massachusetts between 1993 and 1995. Figure 1.10 illustrates this geography for a single city.

Black matches data describing elementary school average test scores (a proxy for school quality) and real estate transaction data to this map. School quality varies discretely at an attendance zone boundary. How much is this worth? The logic of spatial equilibrium tells us that as long as nothing else changes at an attendance

⁵⁴⁵ zone boundary, the land price gradient should be continuous. If we see a jump, it must mean that the value of the properties is changing to reflect the different value of attending schools in the different attendance zone.

To measure this gap in real estate prices (if it is present), Black restricts attention to transactions within a few hundred yards of a school attendance zone boundary, ⁵⁵⁰ and estimates the following regression,

$$\ln(\text{House price}_i) = A_0 + A_1 \text{test score}_i + A_2 \text{border indicators} + \text{controls}_i + \varepsilon_i$$

The parameter A_1 tells us the size (in log points) of the change in house prices at the border. If real estate markets are in “spatial equilibrium” this should tell us the value of improving test scores. Black finds that A_1 ranges between 0.013 and 0.031, so a 1 point increase in test scores increases the logarithm of housing prices by between 0.013 ⁵⁵⁵ and 0.031, which works out to between a 1-3% increase in housing prices. In Black’s sample, about 90% of all houses lie in attendance zones with test scores between 25.2 and 29.8, so moving from the 10th to the 90th percentile of school district quality results in an increase in a house price increase of between about 4 and 12%.

It’s worth taking a minute to think about how neat this is. Suppose you did not ⁵⁶⁰ know about this trick, how would you go about figuring out the value of an improvement in test scores? It would likely involve tracking what happened to students who were otherwise similar, but went to better and worse schools, trying to figure out how their lives turned out, and then trying to attach a dollar value to this difference. This border discontinuity design using real estate prices is much simpler. It lets us work ⁵⁶⁵ out the value of better schools in one step.

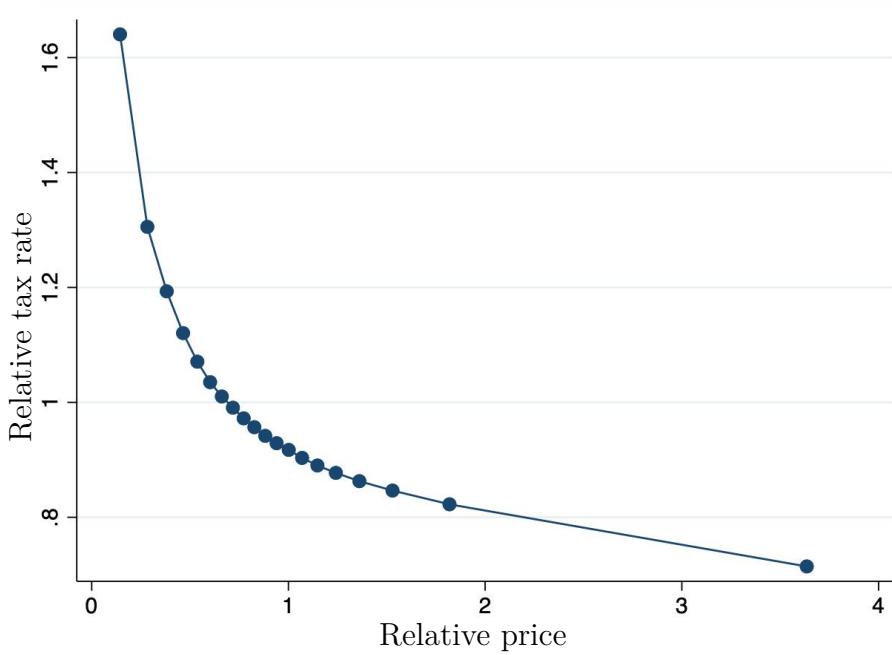
It's also worth noting the problems with this method. First, we're getting the value of school quality to the parents not to the students. If you think parents don't value their childrens' education the way they should, then this could be a problem.
570 Second, it is possible that school attendance zones follow features of the landscape that divide the nice places from the unpleasant. For example, they might follow rivers where one bank is swampy and the other is not. This is a well known problem with these sorts of border discontinuity research designs, and Black follows good practice and carefully excludes attendance zone boundaries where this sort of problem might obviously arise. Third, it may be that what this exercise is picking up is not the value
575 of better schools at all, but the value of living near people who value better schools.

In a similar exercise done a few years after Black's study, Bayer et al. [2007] found that people living on the high score side of a school district boundary had higher incomes and were more likely to be college educated and white. Like Black, Bayer et al. find that real estate prices are higher on the high score side of a boundary,
580 but the fact that people are sorting themselves into better and worse school districts on the basis of other characteristics means that we can't rule out the possibility that part of what people value is proximity to affluent, college educated white people, not access to better schools. We consider such sorting in more detail in Chapter 10.

1.5 Application #2: Detecting racist property tax assessments

Consider the problem of unfair property tax assessments in Chicago. The June 20, 2024 edition of the *Chicago Tribune* reports that

Figure 1.11: Plot of relative tax rate versus relative house price in for the US 2000-2016



Note: *Relative price is property sale price divided by jurisdiction average price in year of sale. Relative tax rate is property's tax rate divided by jurisdiction average tax rate in the year of sale. The tax rate is the tax due in the year of sale divided by the sale price.*

Binned scatter plot shows average relative tax rate and average relative price by 20 quantiles of relative sale price. Based on 26 million residential sales. Figure and figure note reproduced from Berry [2021].

590

An unprecedented analysis by the Tribune reveals that for years the county's property tax system created an unequal burden on residents, handing huge financial breaks to homeowners who are well-off while punishing those who have the least, particularly people living in minority communities.

The problem lies with the fundamentally flawed way the county assessor's office values property.

595 The valuations are a crucial factor when it comes to calculating
property tax bills, a burden that for many determines whether
they can afford to stay in their homes. Done well, these estimates
should be fair, transparent and stand up to scrutiny.

600 But that's not how it works in Cook County, where Assessor Joseph
Berrios has resisted reforms and ignored industry standards while
his office churned out inaccurate values. The result is a staggering
pattern of inequality.⁵

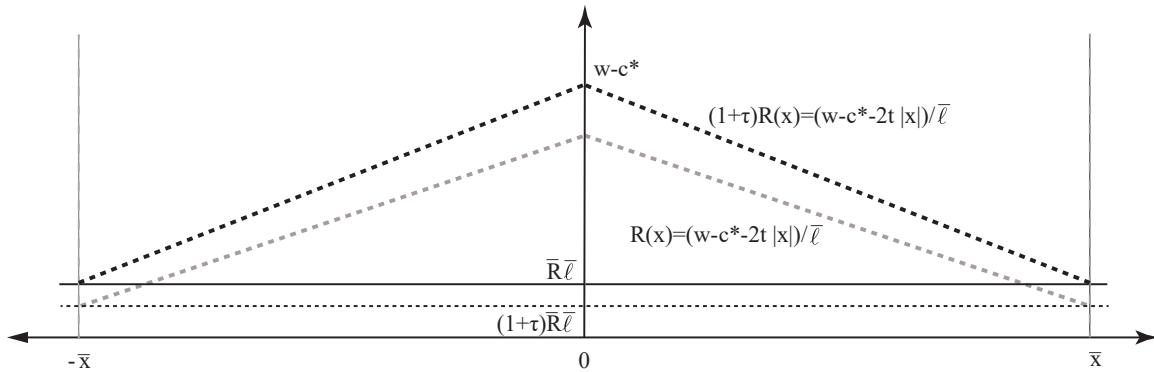
The figure 1.11 illustrates the extent of the problem using a national sample of
real estate transactions and property tax bills. This figure is based on data describing
605 the sales price on the x -axis, and the ratio of the property tax bill to the sale price on
the y -axis. Both quantities are adjusted statistically for differences in averages across
counties and school districts. The steep downward slope means that less expensive
houses are paying a greater share of house value as taxes. The people who live in less
expensive houses, people who are disproportionately black and Hispanic, have a bigger
610 property tax bill relative to the price of their houses than the disproportionately white
people who live in more expensive houses.

This looks bad for tax assessors nationwide. However, while it is plausible (and
even likely) that there has been misbehavior by tax assessors, at least in Chicago,
this figure is not the smoking gun it first appears.

615 In particular, if property taxes are based on market prices and market prices
capitalize tax assessments, the relationship we see in figure 1.11 is just what we
would expect when assessors are behaving fairly. The argument has three steps.

⁵ <https://apps.chicagotribune.com/news/watchdog/cook-county-property-tax-divide/assessments.html>, September 20, 2024.

Figure 1.12: Contract rent and economic rent in an open city with a property tax



Note: The dashed black line gives the land rent in a monocentric city without a property tax. This is the “economic rent gradient” and it is exactly the same as we have seen in other open city examples. The dashed gray line is the “contract rent gradient”, what the tenant pays the landlord, before paying a property tax. The difference between the dashed black and gray lines is the tenant’s tax payment.

First is looking at how property rental prices and property taxes are related. Second is looking at how rental and asset prices are related, and third is putting the first two together.

1.5.1 Property taxes and rental prices

Consider a monocentric city and suppose that land is subject to a property tax rate τ . How does this change the equilibrium? For the purpose of this problem, it’s going to be important to discriminate between economic rent and contract rent. Recall economic rent is the whole value of the property to the tenant, and contract rent is what the tenant pays to the landlord. Let R_C be the contract rent in the taxed city, and suppose that the household at x pays property tax $\tau R_C(x)$, so the household’s

total payment for the property is $(1 + \tau)R_C(x)$ every month. Then the household's problem is

$$\begin{aligned} & \max_{c, x} u(c) \\ \text{s.t. } & w = c + (1 + \tau)R_C\bar{\ell} + 2t|x|. \end{aligned} \tag{1.16}$$

- 630 This is still an open city, so we should have $u(c^*) = \bar{u}$ at all occupied locations. This requires constant consumption of $c^* = u^{-1}(\bar{u})$, just as in a city without property taxes.

Substituting c^* into the budget constraint and rearranging, we get

$$(1 + \tau)R_C = (w - c^* - 2t|x|)/\bar{\ell}.$$

- 635 If we compare this expression to the expression for the land rent gradient in an untaxed city in equation (1.15), we see that the sum of the contract rent and taxes is exactly equal to the economic rent in a city without a property tax. Contract rent plus taxes sums to economic rent. In math, we have

$$R^*(x) = (1 + \tau)R_C(x). \tag{1.17}$$

- This means that adding taxes to the household's problem does not change anything about the city, except that some of the money that would have been collected by 640 absentee landlords is collected by the government. This is exactly the same conclusion we reached in section 1.1.

Figure 1.12 illustrates. The dashed black line in this figure gives the land rent

gradient in the untaxed city. This is the “economic rent gradient”. After households pay this amount of rent and pay for their commute, they are just able to purchase
645 the reservation consumption bundle, c^* . The dashed gray line describes the “contract rent gradient”. This is what the household pays the landlord. This is just enough below the dashed black line, that the tax payment makes up the difference. Because the household doesn’t care whether it pays the city or the landlord, think back to our example of the friendly gangster in section 1.1, the household makes decisions on
650 the basis of contract rent plus taxes. That is, on the basis of economic rent. But this means that nothing about the household’s decision changes.

This is a pretty surprising result. It says that property taxes don’t change behavior at all. That bears repeating. Property taxes don’t change behavior at all. This is a special feature of property taxes.

655 To understand, why this is important, consider the problem of a legislature that needs to raise 100\$ per person in tax revenue. It has the choice of a tax which simply collects 100\$ from everyone, or a tax which collects 100\$ from everyone who stands on one foot for a minute on April 15, and 200\$ from everyone else. We expect both systems of taxation to raise the same revenue, but one makes the average taxpayer
660 worse off by whatever discomfort they endure standing on one foot for a minute. That is, the second tax system creates an incentive for tax avoidance behavior, and tax avoidance behavior is usually wasteful. People engage in it not because they like it, but because it reduces their tax burden.

Taxes that raise revenue without creating an incentive for avoidance behavior are
665 rare, and they are special because they allow the government to raise one dollar of revenue at the cost of only one dollar of harm to the taxpayer. With avoidance

behavior, the cost of a dollar of revenue is always more. It is one dollar plus the cost of the avoidance behavior. A city with a property tax is full of people who act in exactly the same way as they would in a city without a property tax. In this sense,
670 a property tax is at least as good as any other way of raising government revenue.

This result is widely known as the “Henry George Theorem” (although it is not easy to find it stated explicitly anywhere, in particular in Henry George’s writings.) Two caveats apply. First, if people are not all identical, this result starts to break down. Second, it’s important to also tax agricultural land. Otherwise there is an
675 effect on the extensive margin. Some people move away from the edges of the city because untaxed farmers outbid them for land.

In theory, the way property taxes are assessed is as follows. First, an assessor assigns your house a value. Usually, this value is deliberately close to the house’s market value.⁶ Second, the municipal government chooses a “mill rate”, typically around 1%.
680 Each homeowner’s tax bill is the product of the mill rate and their assessed value. In practice, there are lots of opportunities for unfairness and malfeasance, but for the present purpose, we’ll suppose that this has a small impact on tax rates.

We’ve finished working out how property taxes and property rental prices are related. However, property prices depend on *asset prices* not *rental prices*, so knowing
685 the relationship between property taxes and rental prices is not enough to understand the whole process. Our next step is to figure out how rental prices are related to asset prices.

⁶If you ever own a house, you will invariably conclude that the assessor thinks your house is much more valuable than anyone else in the world. On the other hand, California’s infamous Proposition 13 prevented changes in assessed value except when a property changes hands, this means that the assessed value of properties that have not sold for a long time are often much lower than similar houses that have changed hands more recently.

There is a caveat to this argument. If we are being precise, we have so far considered the sale of land, rather than the sale of houses. Property taxes are almost always collected as taxes on the joint value of land and any structure on the land.
 690 This means that “property taxes” are also a tax on houses. This matters because it creates an opportunity for avoidance behavior. “Over-taxed” houses transact for less money, and their owners write larger checks to the city each year. The problem is that you pay property taxes on improvements to your house, too. If you add a room, you
 695 pay property tax on the value of this addition forever. If you are subject to a higher property tax, home improvements cost more. This means that a high property tax disincentivizes home improvement and maintenance, or said another way, incentivizes blight.

1.5.2 Land rent and capitalization

700 How are rent and asset prices related? To answer this question, we need to work out the mathematics of “discounted present values”. Let ρ be the real interest rate. One dollar today turns into $1 + \rho$ in a year. P is the purchase price of a property and R the rental price for one year. If $\rho P < R$ then renters should buy their properties and pocket the difference. If $\rho P > R$ then owners should sell and become renters. Only
 705 when $\rho P = R$ is there no opportunity for intertemporal arbitrage. So, we should have $\rho P = R$. That is, rent equals one year of interest on the asset price of the property.

There is a second way to work out the relationship between rental and asset prices. It’s a little more complicated, but it gives a better intuition about how capitalization works. Suppose the rent on a property is R every year, forever. The sales price is the
 710 value today of this stream of payments. R in one year is worth $R_1 = \frac{1}{1+\rho}R$ today. R

in two years is worth $R_2 = \frac{1}{(1+\rho)^2}R$ today, and so on. R every year forever, starting in one year is worth

$$\begin{aligned} V &= \frac{1}{(1+\rho)}R + \frac{1}{(1+\rho)^2}R + \frac{1}{(1+\rho)^3}R + \dots \\ &= \sum_{t=1}^{\infty} \frac{1}{(1+\rho)^t}R. \end{aligned} \tag{1.18}$$

These sorts of sums of streams of payments are called the “discounted present value” or sometimes just “present value”. While present value calculations look really
⁷¹⁵ complicated, they turn out to be pretty easy to work with. To see this, start by defining $\delta = \frac{1}{(1+\rho)}$ (δ is called the “discount factor”). We can now rewrite equation (1.18) more compactly as,

$$V = \sum_{t=1}^{\infty} \delta^t R. \tag{1.19}$$

This is still a complicated expression, but it is not too hard to transform it into something much simpler.

⁷²⁰ Multiplying both sides of equation (1.19) by δ we get,

$$\delta V = \delta \sum_{t=1}^{\infty} \delta^t R. \tag{1.20}$$

Subtracting equation (1.20) from (1.19),

$$\begin{aligned} V - \delta V &= \sum_{t=1}^{\infty} \delta^t R - \delta \sum_{t=1}^{\infty} \delta^t R \\ \implies (1 - \delta)V &= \delta R + \delta^2 R + \delta^3 R + \dots \\ &\quad - \delta^2 R - \delta^3 R - \delta^4 R - \dots \\ &= \delta R \end{aligned}$$

Substituting in the definition of δ and rearranging, we get $\rho V = R$. That is, the rental price of land is equal to the interest payment on the asset price. Thus, the two ways of figuring out how rent and asset price are related are equivalent (this is pretty neat).⁷²⁵

1.5.3 Fair assessment of property taxes

What does all of this mean for the relationship between property taxes and the sale price of houses?

Restating equation (1.17), we have,

$$R^*(x) = (1 + \tau)R_C(x). \quad (1.21)$$

⁷³⁰ Now we need some notation. Let $V(x)$ be the “economic asset price”. That is, the discounted present value of economic rent $R^*(x)$, and let $V_C(X)$ be the “contract asset price”, that is, the discounted present value of contract rent.

Starting from equation (1.21), we have

$$\sum_{t=1}^{\infty} \delta^t R^*(x) = \sum_{t=1}^{\infty} (1 + \tau) \delta^t R_C(x)$$

or

$$V(x) = (1 + \tau)V_C(x).$$

⁷³⁵ If we take logs of both sides, and recall that $\ln(1 + x) \approx x$ for x small, we get

$$\begin{aligned} \ln V(x) &= \ln(1 + \tau) + \ln V_C(x). \\ &\approx \tau + \ln V_C(x). \end{aligned}$$

Rearranging, we get

$$\tau \approx \ln V(x) - \ln V_C(x). \quad (1.22)$$

Notice that in a city with a property tax, we will never observe $V(x)$. These are the transaction prices that would occur in the (counterfactual) absence of a property tax, but we will observe the tax rate and $V_C(x)$.

⁷⁴⁰ Now suppose we conduct a regression of the tax rate on observed asset prices. What would this look like? Letting i index transactions, it will be something like this,

$$\tau_i = A_0 + A_1 \ln V_C(x)_i + \varepsilon_i. \quad (1.23)$$

From equation (1.22) we expect that A_1 would be about -1 . If we were going to plot this, it would show a rapid decrease in the tax rate with value of the property,
 745 exactly what we see in figure 1.11.

The current system of property tax assessment in Chicago, or in the US as a whole may well be terribly corrupt and unfair, but figure 1.11 does not make this case. That the tax rate declines with property price is an implication of the way that property taxes are capitalized into property prices.

750 This is a dramatic and, to me, surprising result. Why does it work? We know that property prices affect property taxes. This is given in the rules for how property taxes are calculated. But property taxes also affect property prices. This is the logic of capitalization. This means that the relationship we see in figure 1.11 has to reflect both of these relationships. The math we've just worked out shows how these two
 755 relationships work together to create a downward sloping relationship between the tax rate and transaction price, with no racism required.

1.6 Conclusion

We've now developed the basic version of the monocentric city model pretty thoroughly. This model assumes: spatial equilibrium, costly commuting, and central
 760 employment.

The open city model makes the following predictions. First, $R^*(x)$ decreases in x . We've seen that this is correct, and there is more evidence on this point to come.

Second, as commuting costs, t , decrease, utility, \bar{u} , spatial equilibrium immediately implies that equilibrium consumption, c^* , also stays constant. This, in turn

765 requires that; the rent gradient gets flatter and its intercept stays the same. This implies, in turn, that the extent and population of the city increases. We will see some evidence about this later. Third, as wages, w , increase, utility and consumption, \bar{u} and c^* , stay constant. The slope of the rent gradient is unchanged, but its intercept increases by the same amount as the wage increase. The extent of the city and its
770 population both increase, and aggregate rent increases by about the same amount as the aggregate wage bill. We haven't worked out what happens as agricultural rent changes. This is straightforward, but there is not much empirical evidence about this comparative static, so I am going to leave it aside. Fourth, as amenities, A , increase, utility stays constant, but consumption c^{**} falls. The slope of rent gradient is un-
775 changed, and the intercept increases. The city gets longer, population and aggregate land rent both increase. Fifth, changes in property taxes do not change anything except how much rent is collected by absentee landlords. This is called the Henry George Theorem. Finally, spatial equilibrium requires that rent gradients be continuous, unless something that people value about the location changes discontinuously.
780 This intuition gives rise to the widely used border discontinuity research design for evaluating location specific attributes.

This is a good start, but leaves open a few questions. First, the shape of the rent gradient that the model predicts is wrong. The model is predicting a linear rent gradient when in reality it decreases much faster than this. In its current form, the
785 monocentric city model is a model of land allocation. Adding a description of housing (as opposed to just land) in Chapter 3 will help with this. Second, why are people in the center? This is a central assumption. Implicitly, there is a mill or big factory in the CBD where people are more productive than if they work elsewhere. So far, we've

just assumed that people want to be in the center. It would be nice to understand a
790 little bit more about why. We'll come back to this when we talk about agglomeration economies in Chapter 7.

Finally, the assumption that cities are all monocentric is contradicted every time we set foot in a suburban big-box store. Indeed, the assumption that people work only in the CBD is so obviously at variance with observation as to bring the usefulness
795 of the monocentric city model into question. There are two responses to this. First, is to point to the body of empirical evidence confirming the predictions of the monocentric city model. The next chapter describes this evidence. Second is to work with models that allow both firms and households to choose their locations. This line of investigation is very technically demanding and was pioneered in a pair of papers,
800 Fujita and Ogawa [1982] and Ogawa and Fujita [1980], and later refined and extended in Lucas and Rossi-Hansberg [2002]. More recently, it has been the subject of a literature on Quantitative Spatial Models, which are the subject of Chapter 6.

Box 1.2.1: Regression coefficients in log-linear specifications are elasticities

Suppose that we can write n as a log-linear function of r . That is,

$$\ln n = A + B \ln r. \quad (1.7)$$

To see that B is an elasticity, suppose we increase r by 1%, from r^0 to $r^1 = 1.01r^0$, all else equal. Then, we have

$$\ln n^0 = A + B \ln r^0 \quad (1.8)$$

and

$$\begin{aligned} \ln n^1 &= A + B \ln r^1 \\ &= A + B \ln(1.01)r^0 \\ &= A + B(\ln(1.01) + \ln r^0). \end{aligned} \quad (1.9)$$

Subtracting equation (1.8) from (1.9), we have

$$\ln n^1 - \ln n^0 = B \ln(1.01).$$

or

$$\ln \frac{n^1}{n^0} = B \ln(1.01).$$

Next, define ρ as the proportional change in n , so that $\frac{n^1}{n^0} = 1 + \rho$, and recall that $\ln(1 + x) \approx x$ for x small, and we have

$$\rho = B \times 0.01.$$

Multiplying by 100, we have that $100\rho = B$. That is, B is the percentage change in r that results from a 1% change in n . In the jargon, B is the elasticity of n with respect to r .

Box 1.3.1: Example: Monocentric city

Suppose that, $u(c) = \ln(c)$, $\bar{R} = 0$, $\bar{u} = 0$ and $\bar{l} = 1$. The household's problem is to choose location and consumption to solve,

$$\begin{aligned} & \max_{c, x} \ln(c) \\ \text{s.t. } & w = c + R(x) + 2t|x|. \end{aligned}$$

Suppose the city is open, so that people migrate in or out until the utility level at all locations in the city is equal to the reservation utility level. Then, spatial equilibrium requires $\ln(c) = \bar{u} = 0$ so that $c^* = 1$ everywhere.

Using $c^* = 1$ in the budget constraint, we have

$$w - 1 = R(x) + 2t|x|$$

which means that,

$$R(x) = \begin{cases} w - 1 - 2t|x| & \text{if } |x| \leq (w - 1)/2t \\ 0 & \text{if } |x| > (w - 1)/2t. \end{cases}$$

The edges of the city are at $\bar{x} = \pm(w - 1)/2t$ and because $\bar{l} = 1$ this means that $N^* = (w - 1)/t$.

Box 1.3.2: Partial differentiation

Given a univariate function $f : R \rightarrow R$, or $f(x) \in R$, we have

$$\frac{df}{dx} = \lim_{\epsilon \rightarrow 0} \frac{f(x + \epsilon) - f(x)}{\epsilon}$$

This is the “instantaneous slope” of f at x .

Partial differentiation is the generalization of this idea to surfaces. Consider a function $F : R^2 \rightarrow R$, or $F(x_1, x_2) \in R$. This function describes a surface, a height for each point in the plane. How do we think about the slope of such a surface? What we want is a tangent plane rather than a tangent line.

With partial differentiation, we think about the slope of such a plane along one axis. Thus, given $F(x_1, x_2)$, we define

$$\frac{\partial F}{\partial x_1} = \lim_{\epsilon \rightarrow 0} \frac{F(x_1 + \epsilon, x_2) - F(x_1, x_2)}{\epsilon}$$

This is exactly analogous to the univariate derivative, if we imagine that we are finding the slope of a “slice” of the surface parallel to the x_1 axis.

Mechanically, treat the “other variables” as constant and use all the rules you know from univariate differentiation. For example, if $F(x, y) = 2x + 3y^2 + 2xy$ then $\frac{\partial F}{\partial x} = 2 + 2y$ and $\frac{\partial F}{\partial y} = 6y + 2x$. This should be in your calculus book.

Problems

1. In this problem, we will work through an example of the monocentric city model.

805 Assume we have a linear, open city. Let $w=3$, $\bar{l} = 1$, $p_c = 1$, $\bar{R} = 0.5$, $\bar{u} = 0$, and $A = 1$. Let $u(c) = \ln(c - 1)$.

- (a) Set up the household's problem. Assume we are in a spatial equilibrium, so everyone is optimizing and no one wants to move. Call consumption in this equilibrium c^* . What is $u(Ac^*)$ equal to?
- (b) Find c^* .
- (c) Using the constraint from the household's problem, find an expression for \bar{x} in terms of w, c^*, \bar{R}, \bar{l} and t .
- (d) Use the assumption that there is one unit of land at each x to derive an expression for N^* in terms of \bar{x} and \bar{l} .
- (e) Use the household's equilibrium budget constraint and the equilibrium extent of the city to solve for the equilibrium rent gradient, $R^*(x)$.
- (f) Take derivatives of your expressions for \bar{x}, N^* , and $R^*(x)$ with respect to t . How do the city extent, population, and equilibrium rent gradient change as transportation costs increase? Provide some intuition.
- (g) Assume that transportation costs increase from $t_0 = 1$ to $t_1 = 2$. What is the boundary of the city now? What is $R^*(0)$? Use these three points to draw a picture of how the rent gradient changes when t increases. Please label $R^*(0), \bar{R}$ and \bar{x} .

825 (h) How would total land rent within the boundaries of the city change if we go from $t_0 = 1$ to $t_1 = 2$?

2. In this problem, we will analyze property taxes in the monocentric city model.

(a) Assume we have an open, linear city with property tax rate τ_0 . $R_0(x)$ is the land rent in this city. Set up the household's problem (you don't need to solve it).

830 (b) Assume the tax rate increases from τ_0 to τ_1 , where $1 + \tau_1 = (1.10)(1 + \tau_0)$.

Set up the household's problem with this new tax rate.

(c) Using what you know about c^* in an open city equilibrium, solve for $R_1(x)$ in terms of $R_0(x)$. How does the sum of rent and property taxes change?

835 (d) Suppose landlords are responsible for paying the property tax. What does this suggest about the relationship between what tenants pay and property taxes?

840 3. In this problem, we will examine rental gradients in practice. Using Zillow or some similar real estate website, pick a radial road out from the center of a city you know well (for example, along Angel Street from Kennedy Plaza in Providence) and plot the prices of at least 15 similar properties as distance to the center increases. What do you find? You can do this for another city if you would like.

Chapter 2

The Monocentric City Model vs.

Data

845 The language around hypothesis testing in econometrics and statistics is both awkward and particular. We check if “the data fails to reject” a particular hypothesis. We’re not trying to conclude that we’re right, only that we’re not clearly wrong. Gravity is just a theory. It explains the way apples fall from trees pretty well, but we’re still keeping an open mind. The humility implicit in this awkward turn of phrase, “fail to reject”, has always struck me as central to the scientific method. Our models, like the monocentric city model are always simplified, stylized descriptions of the world, and while they may make a lot of good predictions, eventually, we will find some sort of economic dark matter that forces us to revise or abandon any model.

850 With that said, the monocentric city model holds up better than just about any economic model I know. Even though it is highly stylized, cities don’t really exist on a featureless line or plane and not everyone works in the center, it is hard to reject

any of the comparative statics we worked through in Chapter 1 on the basis of the available data.

860 This Chapter surveys the evidence for this claim. In particular, the monocentric city model makes predictions about what should happen to land rent, and the extent and size of the city as transportation costs, amenities, and property taxes change. We here compare these predictions with the available empirical evidence.

2.1 Cities in real life

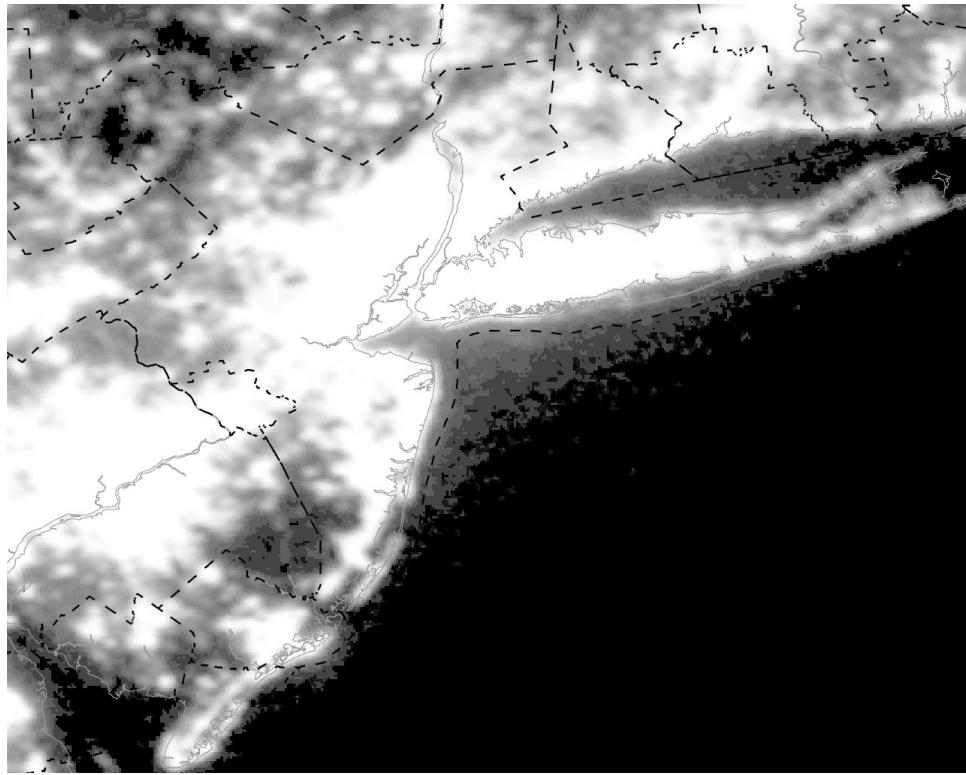
865 To check the predictions of the monocentric city, we need some real group of people to try to match up with the theoretical city. If you think carefully about this, it's pretty hard. The monocentric city model is an abstraction from reality. Finding some real object that matches it closely is going to be hard and there is not going to be a single "right" answer.¹

870 A central feature of the monocentric city model is that it is a labor market. People work in the center if and only if they live in the city. This suggests that we want to be looking at some real world area that seems like a "labor market" in this same sense; an area in which most residents also work and where not many non-residents work.

875 In the US, the unit that satisfies these criteria most closely is probably the "metropolitan statistical area", or MSA. MSAs are metropolitan areas of at least 50k people, built from counties. They are purely reporting units and are defined by the US Census Bureau. There are a few different flavors, "micropolitan statistical areas", "core based

¹Indeed, the whole of the September 2025 issue of the Journal of Urban Economics is about just this issue.

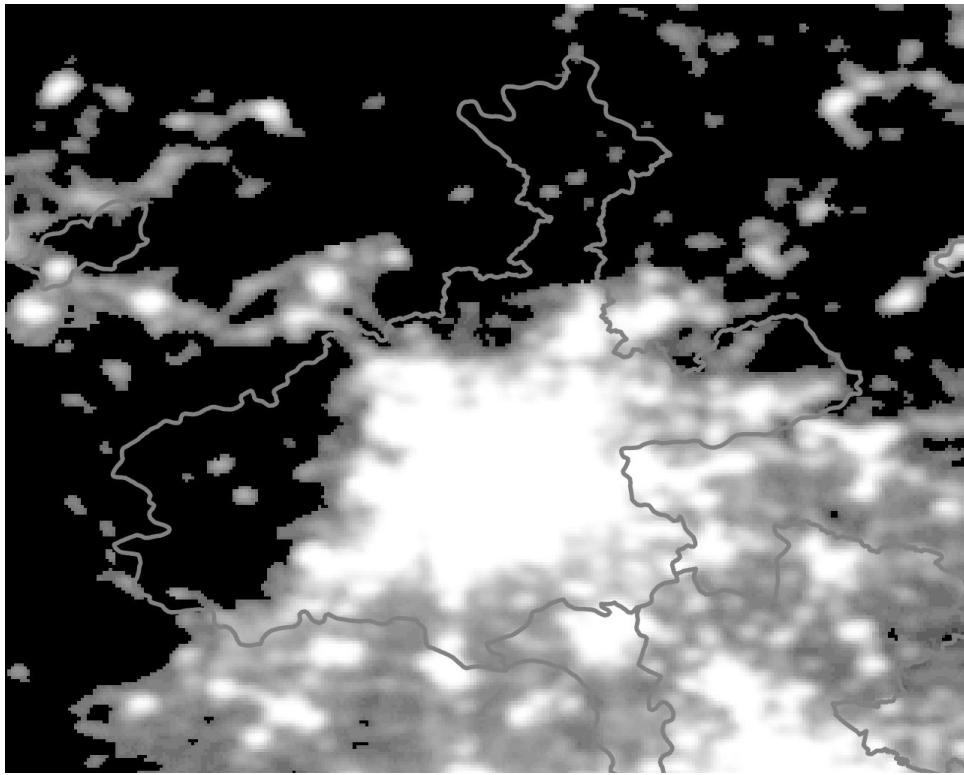
Figure 2.1: MSA boundaries and lights at night for New York City



Note: *New York region and lights at night in 2013. MSA boundaries are given by dashed lines. The New York MSA is in the center of the picture.*

statistical areas” (CBSAs), “consolidated metropolitan statistical area” (CMSAs). Definitions are easy to find on the census website. In every case, the idea is to build a metropolitan labor market from counties. Many of the US based empirical papers we discuss will use MSAs, or one of their variants, as the definition of “city”. Figure 2.1 illustrates MSA boundaries in 2019 for the Northeastern US. The background shows lights at night (from 2013) to show the extent of development. Notice that MSAs often contain a lot of undeveloped land. The dividing line between one MSA and the next is sometimes not matched with a dividing line in night lights, so while MSAs are

Figure 2.2: Prefectural boundaries and lights at night for Beijing



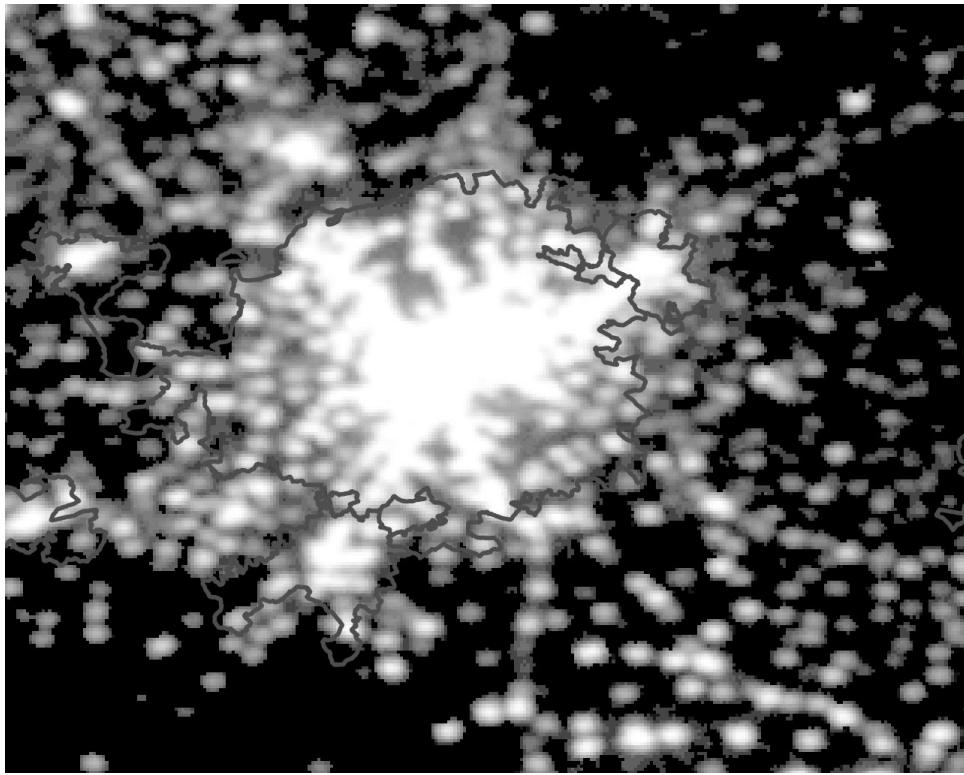
Note: *Madrid region and lights at night in 2013. Beijing is central. Prefectural cities are the nearest analog to US MSAs. Boundaries based on Baum-Snow et al. [2017].*

probably the best empirical analog to the monocentric city that we have available, it is clear that reality is more complicated than our simple model allows.

Besides MSAs, there are many other candidate geographies that describe “cities”.

Municipal boundaries come immediately to mind. That is, the administrative boundaries of a city. This would be a natural unit to consider if we were thinking about issues related to municipal finance or services, but for thinking about the implications of the monocentric city model, they seem like a bad fit. Many municipalities do not contain the employment center where most of their residents work. Conversely, many

Figure 2.3: Functional Urban Area Boundaries and lights at night for Madrid



Note: *Madrid region and lights at night in 2013. The Functional Urban Area containing Madrid is central. Functional Urban Area are a possible analog to US MSAs defined for Europe.*

895 of the workers in center city municipalities do not live in the municipality where they work. Many people who live in suburban Rye, NY, just north of Manhattan, work in Manhattan, not Rye. Conversely, many of the people who work in Manhattan do not live in New York City. The municipality is not the geography that the monocentric city model describes.

900 The census also describes “urban areas”. While the census definition of urban area stretches over several paragraphs, the goal is to delimit the places where people

reside, and the boundaries of urban areas often line up closely with remote sensing data showing built up areas. The boundaries of urban areas don't tell us anything about commuting patterns within the urban area. It follows that, they are also not obvious real world analogs to the geography of the monocentric city model.

Other countries often keep track of pretty similar units, either based on administrative or reporting boundaries. For example, in China, the unit that corresponds most closely to an MSA is the "Prefectural City". China covers about the same land area as the US and, like the US, has about 3000 counties and about 30 provinces, so US and Chinese counties are about the same size and Chinese provinces are a little larger than US states. However, while there is no administrative unit between county and State in the US, in China, the "prefecture" is an administrative unit in between county and province, with each prefecture being made up of a collection of counties in the same province. Gathering up the urban counties in a prefecture gives us the closest Chinese geographical unit to the US MSA. Figure 2.2 illustrates 2005 prefecture city boundaries for Beijing. Notice that, like the US, there is a lot of sparsely populated area in many of the prefectural cities.

Prefectural cities are unlike US MSAs in two regards, First, every Chinese county is part of a prefecture, while many rural counties in the US are not part of an MSA. Second, the prefecture is an administrative unit in China, but the US MSA does not have any administrative authority. It is a geography created solely for reporting purposes. Recalling that we want our empirical unit to describe a labor market, no one commutes outside the area boundary to work, it is not clear whether Chinese prefectural cities are better or worse analogs of the monocentric city model than US MSAs.

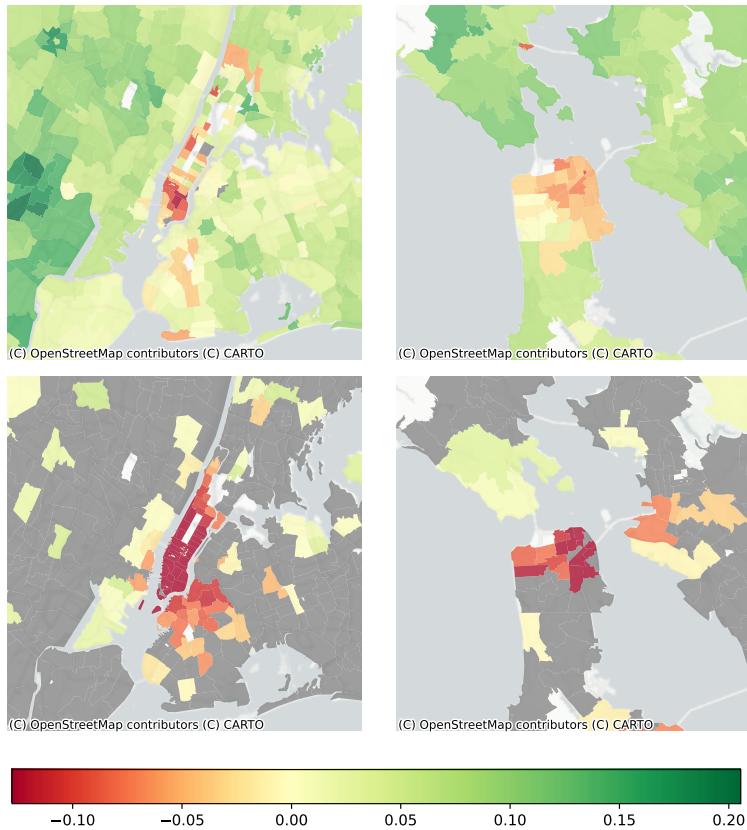
There is less consensus on what is the best European unit to study and the empirical papers the we will discuss rely on different units. One possibility is the Functional Urban Unit illustrated for the case of Madrid in figure 2.3.

2.2 Rent gradients and Covid

930 Among its many effects on our lives, the Covid pandemic that began in 2020 reduced commuting costs for a large part of the workforce. In response to the pandemic, many people were able to work from home at least part of the week, which meant fewer commute trips. In the language of the model, this means that t decreased more-or-less in proportion to the decrease in commute trips. We saw in section 1.3.3 that a
935 decrease in t implies that the rent gradient flattens, its intercept stays the same, and the city gets longer and more populated.

The Covid pandemic also changed the amenity value of living close to the center of the city. Prior to the pandemic, a central residential location meant access to restaurants, bookstores and violin lessons. During the pandemic, the center of the
940 city became a minefield of opportunities to contract an infection. Because the risk of infection was increasing with the number of people around, the risk of infection was lower in suburbs and rural areas, and so the second effect of the pandemic was to increase the amenity value of the suburbs relative to the center city. This is not quite the case we looked at (we had changes in amenities the same everywhere), but it is
945 close. We expect this sort of change in amenities to have different on land rent in the center than the suburbs. It is going to decrease rent in the center. In the suburbs it will decrease rent by less than in the center, or even increase it. Adding the two

Figure 2.4: Pandemic asset price and rent growth in New York City and San Francisco from December 2019 to December 2020



Note: Year-over-year changes in prices (top two panels) and rents (bottom two panels) at the ZIP code level for the New York and San Francisco MSAs from Dec 2019 to Dec 2020. The bottom two rows zoom in on the city center. Darker green colors indicate larger increases, while darker red colors indicate larger decreases. Figure reproduced from Gupta et al. [2022], ©Journal of Financial Economics.

effects of Covid, one operating through commute costs, and one through amenities, we should see urban land rent gradients flatten and decrease at the center. The total effects on city extent and population are ambiguous.

Gupta et al. [2022] look at how the housing market changed during the first year

of Covid. To do this, they assemble data describing real estate transactions and their distance from center of the city. For rental and sales prices they rely on Zillow price indexes. These indexes are available at the zipcode-month level from the real estate website of the same name, and are intended to describe the price or rent for a “standard” house.

Gupta et al. [2022] use these data to estimate house price and rental price gradients for US cities, just as we already saw done for two Japanese and French cities in figure 1.1. The difference is that they use US data, and they estimate gradients for rental prices and for sale prices.

Letting V be sales prices, R be rental price, x be distance to the CBD, and letting i index zipcodes, Gupta et al. [2022] estimate

$$\ln V_i = A_1 + B_1 \ln x_i + \epsilon_i,$$

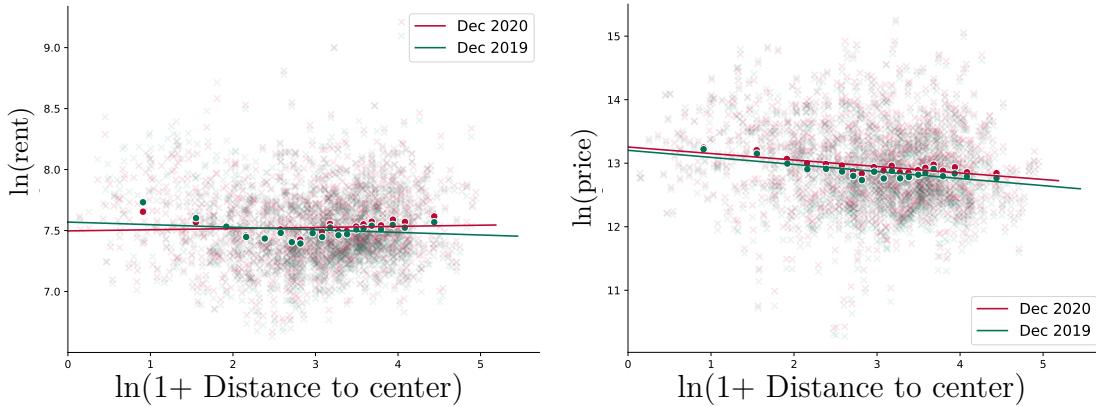
and

$$\ln R_i = A_2 + B_2 \ln x_i + \mu_i,$$

pooling data for the 30 largest MSAs in the US. This gives them a sort of average rent gradient for the largest cities in the US.

They conduct both regressions twice, once before the pandemic, and once after. Figure 2.5 reports their results. The left panel reports rent gradients with the logarithm of distance to the CBD on the x -axis and the logarithm of rent on the y -axis. The green line reports the pre-Covid rent gradient and the red line the post-Covid rent gradient.

Figure 2.5: Changes in real estate prices and rents from the Covid pandemic



Note: The left panel shows the relationship between log distance from the city center and log rent before (green) and after (red) the pandemic. The right panel is identical but reports sale price gradients. Lighter points indicate ZIP codes, while colored points indicate averages by 5% distance bins. Figure reproduced from Gupta et al. [2022], ©Journal of Financial Economics.

As we expect, (1) the pre-Covid rent gradient is downward sloping, (2) the post-Covid, the rent gradient is flatter, and (3) post Covid, the intercept of the rent gradient is lower. Surprisingly, the rent gradient is actually slightly upward sloping post-Covid. This could reflect one of four things. First, that the Covid (dis)amenity so strongly favored suburban locations that it reversed slope of the rent gradient. Second, these rent gradients are actually averages over a sample of 30 MSAs, and that this averaging is partly responsible for the changes we see.² Third, these figures are based on a “Zillow rent index”, which is supposed to be the rent for a standard house. This could be a problem if the “standard” rental unit in a suburban location changed from a small apartment before the pandemic to a large house during, but

²For example, if suburban rents increased more in larger cities, which are more heavily represented in the sample of “big x” rents.

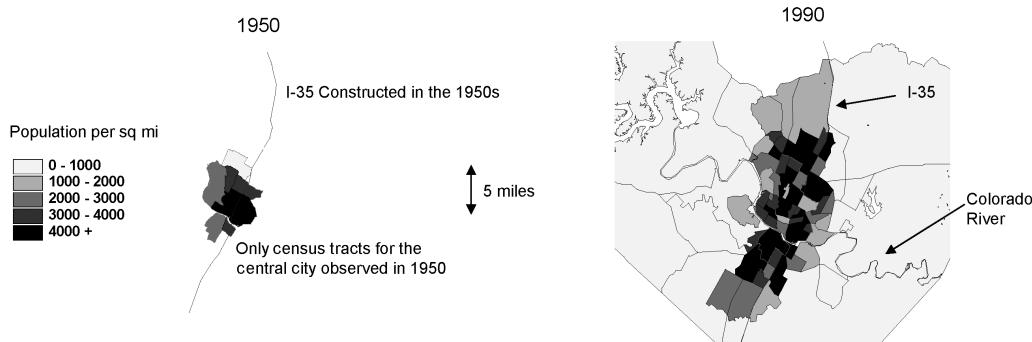
no corresponding change happened in central locations. Fourth, there is some serious problem with the monocentric city model, and this upward sloping rent gradient is alerting us to this problem.

The right panel of figure 2.5 is identical to the left, except that it reports gradients
985 for sale prices, rather than rent. Here we see that both gradients are downward sloping and that the post-Covid gradient is flatter than the pre-Covid gradient. We do not see a decrease in the intercept of the post-Covid price gradient however. This probably reflects an increase in the overall demand for housing during the pandemic. This suggests an important omission from the monocentric city model, and one that we
990 will address in Chapter 3. The monocentric city model does not allow people to choose their housing consumption. Everyone must choose housing, really land, $\bar{\ell}$.

Notice that average rents move more than average sales prices, and in particular, central rents decrease dramatically while central sale prices increase. Much was written during the pandemic about the possibility that Covid would lead to the end of cities as we know them. Some feared that as people retreated to remote work suburbs
995 the central cities would remain only as blighted husks of their former selves.

Recall our discussion of how rental prices and asset prices are related in section 1.5.2. Asset prices are the discounted present value of rental prices. With this in mind, what does figure 2.5 suggest about the permanence of Covid related changes to the
1000 rent gradient? If people anticipated that the Covid related changes were permanent, then the rent and price gradients would change in exactly the same way. That the rent gradient declines in the center, even as asset prices increase slightly, must mean that the buyers of those central houses expect central cities to recover at least some of their pre-Covid attractiveness. Thus, pandemic fears about the death of cities were

Figure 2.6: Development patterns around Austin, Texas.



Note: *Population per square mile in census tracts near central Austin in 1950 and 1990. Between 1950 and 1990, Interstate 35 was constructed more-or-less north to south through central Austin. Over time, the city decentralized along the highway. Figure reproduced from Baum-Snow [2007], ©Oxford University Press.*

¹⁰⁰⁵ not widely shared by people who were actually deciding where to live.

Summing up, the monocentric city model seems to do pretty well at predicting the way that rent and price gradients respond to the pandemic decrease in transportation costs and change in central versus suburban amenities. Both rental and price gradients flattened out as we expect, and rental prices fell in the center. Two features of price ¹⁰¹⁰ and rent changes are noteworthy. First, we observe an upward sloping rent gradient post-Covid. This is not strictly inconsistent with the model, if the changes in amenities are large enough it could reverse the slope of the rent gradient, but it is suspicious. Second, asset prices don't fall in the CBD but rental prices do. This probably the expectation that rents will rise in the future after the pandemic has passed.

¹⁰¹⁵ 2.3 Highways and decentralization

The US Interstate Highway System is a system of limited access highways built where no roads previously existed, or as upgrades of smaller highways. Construction of the Interstate began around 1955 with most of the network was complete by 1970. Most construction since 1970 has been of expansion lanes on existing routes. Not many other changes to US infrastructure more clearly reduced the cost of travel in general, and commuting in particular. We here consider how these new highways changed US cities, and whether these changes are consistent with what the monocentric city model predicts should happen when commuting costs fall.

In a classic paper, Baum-Snow [2007] investigates how US cities changed between 1950 and 1990 in response to the new highway network. He starts by constructing constant boundary MSAs for 139 MSAs. For each MSA he also constructs constant boundary central cities for these MSAs. These are the “old downtowns” as of 1950. He then asks what happened to population in these old downtowns during the period from 1950-1990 when the Interstate was constructed.

He finds that between 1950 and 1990, the total population of the 139 MSAs in his sample increased by 72%. Over the same period, the population of the old center cities *decreased* by 17%. That is, as the population of US MSAs nearly doubled, the population of their centers fell by almost one fifth. This sort of spreading out and growth is exactly what we see in the monocentric city model when commuting cost t falls.

Figure 2.6 illustrates the changes that took place in Austin, Texas around the time that Interstate 35 was constructed running North to South through the center of the city. Between 1950 and 1990, the city expanded linearly along the highway.

Although Austin is not really a linear city, the changes illustrated in this figure look
 1040 like the comparative static that we found in section 1.3.3, people spread out along
 the road as the cost of traveling on the road falls.

As a way to check whether this phenomena is general, Baum-Snow [2007] estimates
 a population density gradient, much like the land rent gradients we've already seen.
 To understand this regression, introduce the following notation. Let P_{ij} be population
 1045 in census tract j of MSA i . Let $\text{dis}_{ij}^{\text{cbd}}$ be the distance from this tract to the CBD, and
 let $\text{dis}_{ij}^{\text{hwy}}$ be the distance to nearest interstate. Baum-Snow estimates,

$$\ln P_{ij} = \alpha_i + \beta \ln \text{dis}_{ij}^{\text{cbd}} + \gamma \ln \text{dis}_{ij}^{\text{hwy}} + \epsilon_{ij} \quad (2.1)$$

using data for 1950, and again for 1990.

This is a population density gradient, but elaborated to allow for the fact that
 transportation costs are not the same in every direction traveling outward from the
 1050 CBD. Transportation costs ought to be lower along a highway than not.

Equation (2.1) is a population density gradient. As described in Chapter 1, the
 monocentric city model does not allow for variation in population density. Every
 household lives on a parcel of size \bar{l} . Thus, mechanically, equation (2.1) seems unlikely
 to help us evaluate the monocentric city model. This is a fair point. However, equation
 1055 (2.1) will let us look at how people spread out from the center when transportation
 costs fall, a prediction of the model.

Baum-Snow finds that β increases from $-.132$ to $-.114$ between 1950 and 1990.
 Because the regression is conducted in logarithms, we can interpret the coefficients as
 elasticities. Thus a 1% increase in distance to the CBD leads to a 0.13% decrease in

1060 density in 1950, but only a 0.11% decrease in 1990. That is, the population density gradient got flatter. This is just what the monocentric city model predicts will happen when transportation costs fall. Baum-Snow also finds γ decreases from about -0.014 to -0.019 , so population density falls faster as we travel away from a highway in 1990 than in 1950. This is not something the monocentric city model can address. The
1065 model assumes that the city is radially symmetric, and this regression is looking for changes that are not radially symmetric.

There is a problem with these estimates of equation (2.1), however. We cannot be sure whether changes in population density occur because people change their choice of residential location in response to the change in transportation costs, or whether
1070 the Federal Highway Administration cleverly anticipated where people were moving and built roads to meet these changes. In all likelihood, some of both is going on, and this means that interpreting equation (2.1) is difficult.

To resolve this problem, we would like to do something in the spirit of a clinical trial. That is, we imagine assigning each MSAs to a “treated” or “control” sample
1075 on the basis of a coin toss (heads and Brownsville is “Treated”, tails and Spokane is “Control”) and then assigning highways to the treated, but not the control sample in 1950. In 1990, we could learn the effects of highways on the treated cities by comparing them to the control cities. If we were able to do this, then we could be pretty sure that the treated cities were different from the control cities, on average,
1080 only because they got highways.

Simply describing this process, makes it clear that it’s impractical. However, one of the important innovations in economics over the past 20-30 years has been the development of econometric techniques that allow us to simulate, sometimes more

convincingly than others, exactly this sort of experiment. This is just what Baum-Snow does to resolve the problem of reverse causation, and much of his paper is about these econometric details. This allows Baum-Snow to estimate how much cities change in response to the construction of the interstate highway network.

Baum-Snow's econometric method requires that he restrict himself to a particular question. Define a “radial interstate ray” to be just what it sounds like, an interstate highway that travels from the constant boundary center city out of the MSA. In figure 2.6, I35 counts as two rays, one going north and one south. Let rays_i denote the count of radial interstate rays in city i , e.g., two for Austin, and let N_i^c be the center city population. The operator Δ indicates changes from 1950-90, and “controls” is a list of other variables whose purpose is to solve econometric problems beyond the scope of this book. Baum-Snow's main estimating equation is

$$\Delta \ln N_i^c = \delta_1 + \delta_2 \text{rays}_i + \text{controls}_i + \epsilon_i. \quad (2.2)$$

Because there were no interstate rays in 1950, the count of rays in 1990 is also the change in the number of rays, so this is really a regression of the change in the logarithm of central city population on the change in interstate rays.

To understand what this equation is doing, you need to recall some of the rules

₁₁₀₀ for manipulating logarithms,

$$\begin{aligned}\Delta \ln N_i^c &= \ln N_{1990i}^c - \ln N_{1950i}^c \\ &= \ln \frac{N_{1990i}^c}{N_{1950i}^c} \\ &= \ln(1 + r_i) \\ &\approx r_i\end{aligned}$$

So, as long as the rate of change is small enough that $\ln(1 + x) \approx x$ is a good approximation, the coefficient of rays in equation (2.2), δ_2 , tells us the change in central city population caused by each ray as a share of the initial value.

Baum-Snow's big result is that δ_2 is about -0.11 . This means that each radial ₁₁₀₅ interstate ray reduces central city population by about 11%. Because an average MSA received about 1.5 radial interstate rays, for an average MSA, the interstate caused about a 16% decline in the population of the constant boundary central city between 1950 and 1990. Recalling that the population of constant boundary central cities declined by about 17% during this time, this means that the interstate caused almost ₁₁₁₀ the entire decline in central city population between 1950 and 1990.

If we think that the main effect of highways on cities is to reduce transportation costs, then this looks pretty good for the monocentric city model. The monocentric city model predicts a decreasing share of population near the center as t falls, just what Baum-Snow finds.

₁₁₁₅ **2.4 Highways and growth**

A second prediction of the (open) monocentric city model is that cities will grow when transportation cost falls. Duranton and Turner [2012] examine this hypothesis by looking at how MSA population (really employment) changes with lane kilometers of interstate highway between 1983 and 2003.

₁₁₂₀ In their sample of 227 MSAs, average employment grew from about 250 thousand to about 410 thousand between 1983 and 2003, an increase of about 65%. During this same time, kilometers of interstate highway in an average MSA increased from 234 to 255, an increase of about 9%. This works out to annual growth rates of about 2.8% for population, and 0.5% for highway kilometers.³

₁₁₂₅ We hope that the Federal Highway Administration builds highways in cities where people want to move, so we should be concerned that correlations between changes in MSA highways and changes in employment could reflect reverse causation. Duranton and Turner rely on the same basic econometric technique as Baum-Snow [2007] to try to overcome this problem, and like Baum-Snow [2007], much of the Duranton and ₁₁₃₀ Turner paper is devoted to describing this technique. As I did in the discussion of Baum-Snow, I'm going to skip over these details.

Duranton and Turner want to check whether cities with more highways grow faster. For this purpose, let n_{it} be employment in MSA i at year t , let r_{it} be lane kilometers of interstate in MSA i in year t , and let x_{it} be control variables that address ₁₁₃₅ econometric problems beyond the scope of this book. The main estimating equation

³If you are paying close attention, you will notice that the number of MSAs varies across the different studies I've described. This reflects three things: (1) sometimes, not all data is available for all MSAs, (2) the number of MSAs increases over time as more metropolitan areas cross the 50,000 population threshold, and (3) the studies may rely on slightly different definitions of MSA.

from Duranton and Turner [2012] is,

$$\Delta \ln n_{it+1} = A_0 + A_1 \ln r_{it} + A_2 \ln n_{it} + A_3 x_{it} + \varepsilon_{it} \quad (2.3)$$

This looks a lot like the Baum-Snow regression, equation (2.2), but there is an important difference. Because Baum-Snow started his study when there were zero interstates, his control for highway rays was really “change in rays”. That’s not what’s happening here. Here employment growth is a function of the initial level of highway lane kilometers, so it’s actually not easy to compare the two regressions, even though they look a lot alike.
1140

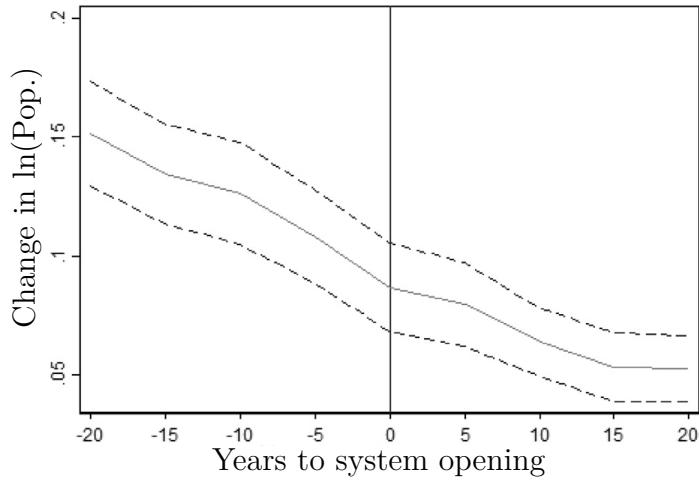
As before,

$$\begin{aligned} \Delta \ln n_{it+1} &= \ln n_{it+1} - \ln n_{it} \\ &= \ln(n_{it+1}/n_{it}) \\ &= \ln(1 + \rho_n) \\ &\approx \rho_n \end{aligned}$$

So that A_1 tells us the effect on the growth rate of the MSA employment, ρ_n from a change in initial lane kilometers of interstate.
1145

The main empirical result in Duranton and Turner is that A_1 is about 0.15. This means that a 1% increase in lane kilometers increases the annual employment growth rate for an average MSA by about $0.15 \times 1\% = 0.15\%$. Recalling that the stock of interstate kilometers in an average MSA grows by about 0.5% per year during their sample, highway construction contributes $0.15 \times 0.5\% = 0.075\%$ per year to the
1150

Figure 2.7: Average city population growth rate as time to subway system opening varies



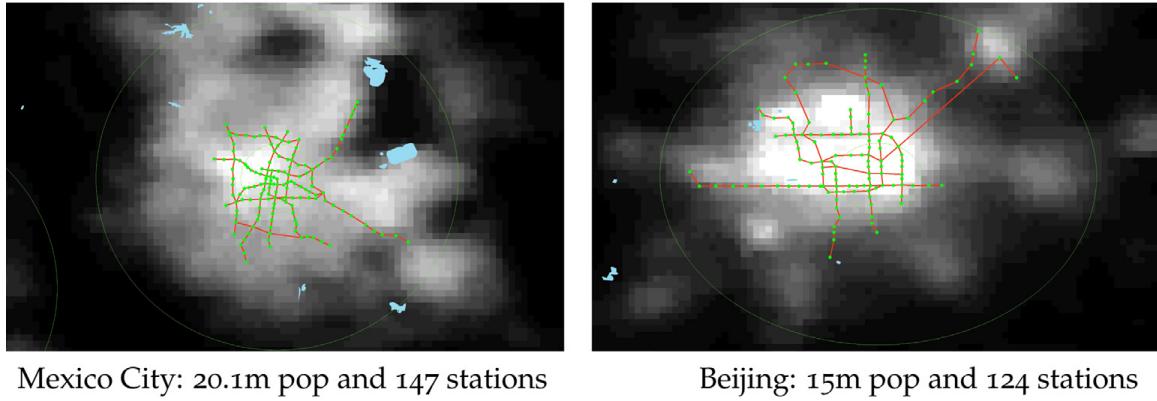
Note: *Subway system opening and population growth (constant sample of 61 cities).* The graph depicts mean change in city log population according to time to system opening. $t = 0$ indicates the year in which a city's subway system was inaugurated. We impose a constant sample of cities on either side of $t = 0$. Graph based on constant sample of 61 cities. It does not look like subways cause population growth. Figure reproduced from Gonzalez-Navarro and Turner [2018], ©Journal of Urban Economics.

baseline 2.8% annual growth rate of MSA employment. The monocentric city model predicts that a city will get bigger when transportation costs fall. This is just what Duranton and Turner find. As we build more roads in a city, more people come, but the magnitude is tiny.

1155 2.5 Subways, decentralization and growth

If changes to transportation costs affect the way cities are organized, it shouldn't matter if changes result from better roads, more telecommuting, or better subways.

Figure 2.8: Lights at night and subways in Mexico City and Beijing in 2010



Note: 2010 lights at night, subway route maps, and all subway stations constructed prior to 2010 in Mexico City (left) and Beijing (right). The gray/green ellipses in each figure are projected 5 km and 25 km radius circles to show scale and light blue is water. Figure reproduced from Gonzalez-Navarro and Turner [2018], ©Journal of Urban Economics.

We've just checked telecommuting and roads.

Gonzalez-Navarro and Turner [2018] check subways. Their paper is based on three main sources of data. The first is a census of all subway systems in the world. The second is a panel of data describing the population of all 632 cities in the world that had a population above 750,000 sometime between 1950 and 2010. The third is “lights at night” data for these cities. This is the same data we saw in figures 2.1 and 2.2.

As of 2010, among these 632 large cities there are 138 with subway systems. Four subways were in operation around 1860, Liverpool, Boston, London and New York. Cities in Asia began constructing subways in the 1970 and account for most of the new systems since that time. Overall, cities in Europe are much better provided with subways than anywhere else. As of 2010, an average system consists of 77 route kilometers and 57 stations. On average, the 138 subway cities had a population of

1170 about 4.7m people in 2010 and their populations grew at an average rate of just above 2% per year between 1950 and 2010.

1175 Among the 138 subway cities that Gonzalez-Navarro and Turner [2018] study, only 61 have data that is complete enough to allow the calculation of population growth rates 20 years before and 20 years after the system opening. The solid line in figure 2.7 reports the average population growth rate in these cities as a function of the time to the subway system opening. The dashed lines report confidence intervals around the mean. That this line slopes downward tells us that, on average, the growth rate is falling in these cities. This is a common empirical finding. Researchers often find that the growth rate of cities falls as they get larger, and this sample consists of large 1180 cities, getting larger. More interestingly, we see no change in the trend around the time when each city's subway system opened (zero on the x -axis). This is one of the main findings from Gonzalez-Navarro and Turner [2018]; the opening of a subway system does not seem to affect the level or growth of population in the cities where they open.

1185 Gonzalez-Navarro and Turner also investigate whether subways decentralize cities. To do this, they use lights at night data to estimate a light gradient and then ask whether this gradient flattens after the subway system opens.

1190 This means estimating more gradients, this time to see how the intensity of lights at night varies with distance to the center. For each of 138 subway cities, for each year when they observe night lights (1995, 2000, 2005, 2010), they calculate mean light intensity in a series of donuts, 0-1.5km, 1.5-5km, 5-10km, 10-25km and 25-50km, centered on the CBD. The 5km and 25km donuts are faintly visible in figure 2.8. Next, let y_{itd} be the mean light intensity in donut d for city i in year t , and let x_{itd}

be distance of the midpoint of the donut from center, e.g., 7.5 km for 5-10km donut.

¹¹⁹⁵ They can now estimate city-year specific light density gradients,

$$\ln y_{itd} = A_{it} + B_{it} \ln x_{itd} + \epsilon_{itd}.$$

That is, Gonzalez-Navarro and Turner estimate the slope of the light gradient 138 times in each of the four years they observe lights at night. This gives a separate B_{it} for each city-year, each describing the change in mean light intensity as we move from the smallest to the largest of the five donuts. These are exactly the same log-linear gradients we've already seen, but explaining the intensity of lights at night rather than property prices or population density.
¹²⁰⁰

With the slope B_{it} for city-years it in hand, Gonzalez-Navarro and Turner can check if subways cause cities to spread out with the following regression,

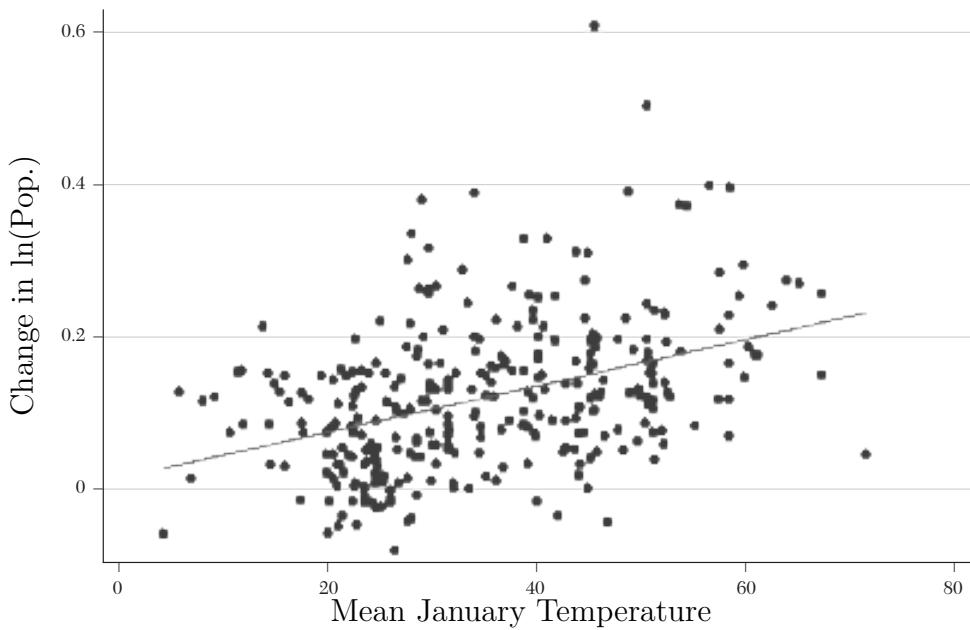
$$\Delta B_{it} = A_0 + A_1 \Delta \ln(\text{Subway Stations}_{it}) + A_3 \text{Controls}_{it} + \epsilon_{it}$$

because light gradients are downward sloping, if subways cause cities to decentralize, ¹²⁰⁵ they will increase the B_{it} , and this is just what they find. That is, as the number of subway stations increases, B increases and the light gradient gets flatter. This is just what the monocentric city model predicts will happen when transportation costs fall.

2.6 Amenities and city size

Another prediction of the monocentric city model is that cities will be bigger as their ¹²¹⁰ amenities are better. To check this, Glaeser and Gottlieb [2009] collect data on 316

Figure 2.9: Population growth and mean January temperature for US MSAs between 1990 and 2000



Note: *y-axis is the change in the logarithm of MSA population between 1990 and 2000. x-axis is mean January temperature. US MSAs with milder winters grow faster. Figure reproduced from Glaeser and Gottlieb [2009] ©American Economic Association.*

US MSAs in 1990 and 2000, and ask how population growth is related to weather.

They find a strong relationship between good weather and growth in population.

Figure 2.9 illustrates this result. On average, a one degree Fahrenheit increase in mean January temperature is associated with a 2% increase in population over the course of a decade. This is also consistent with the predictions of the monocentric city model.

2.7 Property taxes and land prices

One of the more interesting consequences of spatial equilibrium is that property taxes are capitalized into land prices in a really mechanical way. One dollar of property taxes equals one dollar of rent, and one dollar of property taxes per year equals the discounted present value of one dollar per year in asset prices.

In reality, things ought to be more complicated for two reasons. First, property taxes are assessed on the value of land *and house*. If you put an addition on your house, you need to pay property taxes on the value of the addition forever. Similarly for a new paint job, etc. Thus, if we allow a little more realistic description of the world, we might expect that 1\$ of property taxes will decrease the value of house and land by more than 1\$ because it will lead to sub-optimal maintenance.

Second, up to now, we have implicitly assumed that property taxes leave the model. They go to the city government and are entirely wasted. In fact, property taxes are used, at least in part, to provide important public services like schools, trash collection, fire and police protection, parks and roads. These things will operate like amenities, and hopefully, have value of at least 1\$ per dollar of taxes collected. Strictly, the monocentric city model predicts that property taxes decrease the value of a property *all else equal*, but it is not easy to find real world examples where taxes changes and services do not.

One case that seems pretty close occurred in 2008 in Toronto. In 2008, Toronto imposed a “land transfer tax”. This is a property tax that you pay when you sell your property, rather than every year, as property taxes are usually collected. This tax was imposed in Toronto, but not in neighboring municipalities. Because the land transfer tax is attached to a municipality, we should expect their effect on real estate

markets to vary discretely at municipal borders. If taxes are capitalized into real estate prices as the monocentric city model suggest, then we should see prices fall in Toronto by about the magnitude of the tax, net of whatever value of public services the tax will purchase. Dachis et al. [2012] argue that the circumstances surrounding 1245 the imposition of the 2008 Toronto land transfer tax make it unlikely that the new tax affects the provision of municipal services. They then do exactly the experiment described above. They find that real estate prices fall by about one dollar for every dollar of tax assessed (although their estimates are not very precise).

Palmon and Smith [1998] finds another way to look for changes in property tax 1250 rates, all else equal. They study data describing house prices in 50 subdivisions in the Houston suburbs. All 50 subdivisions are similar. They are served by three school districts of similar quality, and were constructed within a few years of each other.

Water and sewer service is the same for every subdivision and was provided by the private developers who built out each subdivision. Developers financed the construction of water and sewer infrastructure by issuing bonds, with the payment on 1255 these bonds financed by property taxes collected from homeowners in the subdivision. The interest rate on the bonds, and hence the subdivision property tax rate, varies with the interest rate that prevailed when construction occurred. This means that the different subdivisions are paying different prices for the same water and sewer 1260 service.

Palmon and Smith find that about 65 cents of every dollar of property tax is capitalized into property prices. This suggests that the basic logic of capitalization that emerges from the monocentric city model is economically important. But it is also a bit less than the predicted 100% capitalization, so the model, somehow, not

1265 exactly right. It's worth noting that this analysis is based on only about 500 real estate transactions, so we should worry about how precise their estimates are and whether they apply outside of the Houston suburbs.

2.8 Wages and rents

1270 Another prediction of the monocentric city model, and the last one we'll check, is that a 1\$ increase in wages leads to a 1\$ increase in land rent.

We can find some evidence about this in Davis and Ortalo-Magné [2011]. This paper looks at the relationship between income and expenditure on housing using a large census data set. Table 2.1 presents their findings. For each MSA and census year, they calculate the ratio of rent to income for each household. The table presents 1275 the value of this ratio for the median household in each MSA. The last row of the table reports the average of these values over all MSAs. That is, the average median rent to income ratio. The median share of income devoted to rent ranges between about 0.21 and 0.29, with most MSAs even closer to 0.25. That is, people spend about 25% of their income on housing, no matter where they live (in the US) or how much 1280 money they make.

Glaeser and Gottlieb [2009] also make this point. In particular, for MSAs in the US in 2006, they regress the logarithm of median MSA income on the logarithm of median home value. That is,

$$\log(\text{MSA median income})_i = A_0 + A_1 \log(\text{MSA median home value})_i + \epsilon_i.$$

They find that $A_1 = 0.34$. This means that a 1% increase in MSA mean income is

1285 associated with a 0.34% increase in MSA mean home value. This is a little larger than the 0.25 that Davis and Ortalo-Magné [2011] finds, but the data are a little different too.

1290 Clearly wages and housing expenditure move together as the monocentric city model requires. But, equally clearly, the effect is much less than the one-for-one relationship that the model predicts.

2.9 Conclusion

The first prediction of the monocentric city model is that the rent gradient, $R(x)$, should decrease with distance to the center. This is broadly consistent with observation. We see it for cities in France and Japan in figure 1.1, for housing price gradients in the US both before and post-Covid, and for apartment rental prices before Covid in figure 2.5. The post-Covid rental price gradient for apartments increases slightly as we move away from the center. This contradicts the prediction of the simplest version of the monocentric city model, though if we allow residential locations to contribute directly to utility, i.e., “amenities”, then this slight upward slope could reflect capitalization of Covid risk in a way that is consistent with the model.

1295 The next prediction of the monocentric city model is that rent gradients should flatten as commuting costs fall. This is exactly what we saw during Covid, when commuting costs fell as people began to work remotely. In figure 2.5 both sale and rental price gradients flattened in response, although part of this response was surely due to the greater risk of Covid in denser more central locations.

1305 The monocentric city model predicts that cities will spread out and grow in size

as transportation costs fall. Between 1950 and 1990, the population of constant boundary center cities in the US fell by 17%, even as the population of MSAs that contained them increased by 72%. The entire decrease in central city population can
1310 be accounted for by the construction of Interstate highway rays through these cities. If we think that the main effect of highways was to reduce transportation costs, then this is broadly consistent with the prediction of the monocentric city model. Gonzalez-Navarro and Turner [2018] finds something similar for subways. The gradient of light intensity flattens out in cities after they build subway systems.

1315 We need a caveat here. Baum-Snow [2007] finds that the population of the constant boundary central city falls with the construction of radial interstate highways. This means that population density in the center must fall. As we formulated the monocentric city model in Chapter 1, each household is constrained to consume a constant amount of land \bar{l} , and so population density cannot change. This is an
1320 obvious shortcoming of the model, and we address it in Chapter 3.

The monocentric city model predicts that the population of a city will grow with a reduction in commute costs. Here, the evidence is for a small effect, at most. Gonzalez-Navarro and Turner [2018] can find no effect on city populations from the construction of subways, and Duranton and Turner [2012] find a small effect of interstate highways on the growth of employment. There is nothing in the monocentric city model that requires that reductions in commute costs have a large effect on population, but it does require it to be positive. The available evidence is not conclusive
1325 on this point. The effects of transportation infrastructure on city population are small or zero, and it is not clear which.

1330 If wages, w , increase by one dollar, the monocentric city model predicts that

expenditure on housing will also increase by one dollar for almost all households. In fact, in the US expenditure on housing increases by about one dollar for each three or four dollars of additional income. This is the right sign, but the magnitude is too small. However, this comparison is not quite fair. The data describes expenditure on
1335 “housing” that consists of both house and land, while in the model “housing” is just land. Just as we need a richer model of housing in order to think about population density, we also need a richer model of housing if we are going to predict the share of wages that are capitalized into housing prices.

Property taxes are also capitalized into asset prices much in the way the model
1340 suggests. Because the model requires that tax revenue leave the model, when in reality at least some of it will provide valuable services, the model is addressing an unusual special case. In the rare real world analogs of this special case, a good guess would be that 60-100% of a household’s property tax bill is capitalized into the price of housing. That is, if the property tax on a house goes up by one dollar per year forever, then we expect the sale price of the house to fall by 60-100% of the discounted
1345 present value of this stream of payments.

If we generalize the basic model to allow for locations to contribute directly to utility through amenities, then the model predicts that cities will be larger as amenities improve. We have evidence in support of this. US cities with milder winters grow
1350 faster than those with harsher winters.

We can now score the contest between the monocentric city model and the data. Almost all of the predictions the model makes about the slope of the rent gradient and about the way that a city responds to changes in commuting costs, wages, and amenities are qualitatively correct. Some of the details and magnitudes are wrong,

¹³⁵⁵ however. Rent goes up only 25 to 35 cents for every one dollar increase in wages, not one dollar as the model predicts. Central city population density falls as commuting costs fall but the model does not allow population density to change. The model predicts that land rent fall about linearly with distance to the center, while the data shows a much faster decrease.

¹³⁶⁰ Finally, the model predicts that cities grow as commute costs fall, and in spite of the fact that the data shows that changes in commute costs have large effects on the way cities are organized, we see at most a tiny effect of changes in commute costs on city size. Given how successful the monocentric model is otherwise, this finding is puzzling.

Table 2.1: Median expenditure on rent divided by median wage, in a sample of MSAs

MSA	1980	1990	2000
Atlanta, Sandy Springs, Marietta	0.24	0.25	0.25
Austin, Round Rock	0.27	0.25	0.25
Boston, Cambridge, Quincy	0.24	0.26	0.24
Buffalo, Niagara Falls	0.20	0.22	0.23
Charlotte, Gastonia, Concord	0.23	0.24	0.24
Chicago, Naperville, Joliet	0.21	0.23	0.23
Cincinnati, Middletown	0.21	0.22	0.20
Cleveland, Elyria, Mentor	0.21	0.22	0.23
Columbus	0.22	0.23	0.23
Dallas, Fort Worth, Arlington	0.24	0.24	0.24
Denver, Aurora	0.25	0.24	0.26
Detroit, Warren, Livonia	0.21	0.22	0.22
Grand Rapids, Wyoming	0.19	0.24	0.21
Houston, Sugar Land, Baytown	0.23	0.22	0.23
Indianapolis, Carmel	0.21	0.23	0.23
Jacksonville	0.27	0.24	0.25
Kansas City	0.21	0.22	0.22
Los Angeles, Long Beach, Santa Ana	0.25	0.29	0.27
Louisville, Jefferson County	0.22	0.23	0.21
Miami, Fort Lauderdale, Pompano Beach	0.27	0.29	0.29
Milwaukee, Waukesha, West Allis	0.20	0.23	0.22
Minneapolis, St. Paul, Bloomington	0.24	0.25	0.23
Nashville, Davidson, Murfreesboro, Franklin	0.23	0.24	0.24
New Orleans, Metairie, Kenner	0.24	0.25	0.24
New York, Northern New Jersey, Long Island	0.22	0.24	0.24
Orlando, Kissimmee	0.26	0.27	0.27
Philadelphia, Camden, Wilmington	0.22	0.24	0.23
Phoenix, Mesa, Scottsdale	0.28	0.26	0.26
Pittsburgh	0.21	0.21	0.22
Portland, Vancouver, Beaverton	0.27	0.24	0.25
St. Louis	0.22	0.23	0.22
Salt Lake City	0.24	0.23	0.27
San Antonio	0.22	0.24	0.24
San Diego, Carlsbad, San Marcos	0.29	0.30	0.28
San Francisco, Oakland, Fremont	0.26	0.28	0.25
San Jose, Sunnyvale, Santa Clara	0.24	0.26	0.25
Seattle, Tacoma, Bellevue	0.25	0.25	0.26
Tampa, St. Petersburg, Clearwater	0.26	0.25	0.25
Washington, Arlington, Alexandria	0.23	0.26	0.24
US Average	0.24	0.25	0.24
Standard deviation	0.02	0.02	0.02

Note: Table excerpted from Davis and Ortalo-Magné [2011].

1365 **Problems**

This problem will examine the change in the rent and purchase price gradients from Gupta et al. (2021).

1. Before the pandemic, the rental price gradient was described by:

$$\ln R_0(x) = 7.6 - 0.04 \ln(x + 1)$$

1370 where x is distance from the city center. This is shown in the left panel of figure 2.5. During the pandemic, the rental gradient changed to:

$$\ln R_1(x) = 7.5 - 0.004 \ln(x + 1)$$

What are the monthly rental prices at $x = 0$, before and during the pandemic?

What is the percent change in rent at $x = 0$?

2. As shown in the right panel of figure 2.5, the asset price gradient before the pandemic was described by:

$$\ln P_0(x) = 13.2 - 0.127 \ln(x + 1)$$

1375 During the pandemic, this gradient changed to:

$$\ln P_1(x) = 13.15 - 0.115 \ln(x + 1)$$

What are the asset prices at $x = 0$, before and during the pandemic? What is

the percent change in asset price at $x = 0$?

- 1380 3. Suppose that the pandemic-related changes in rental prices are permanent. Use the results from problem 1 to find the implied asset price of rental properties at $x = 0$ before and after the pandemic, using interest rate $\rho = 0.03$. What is the percent change in these implied asset prices?
- 1385 4. Compare this implied change in asset prices, which assumed that the change in rental prices due to Covid would be permanent, to the actual change in asset prices from problem 2. Which is larger? What does this suggest about how long people expect the pandemic to last?
5. Throughout the pandemic, people have speculated that Covid would be “the death of cities”. What does your work above suggest about this sort of speculation?

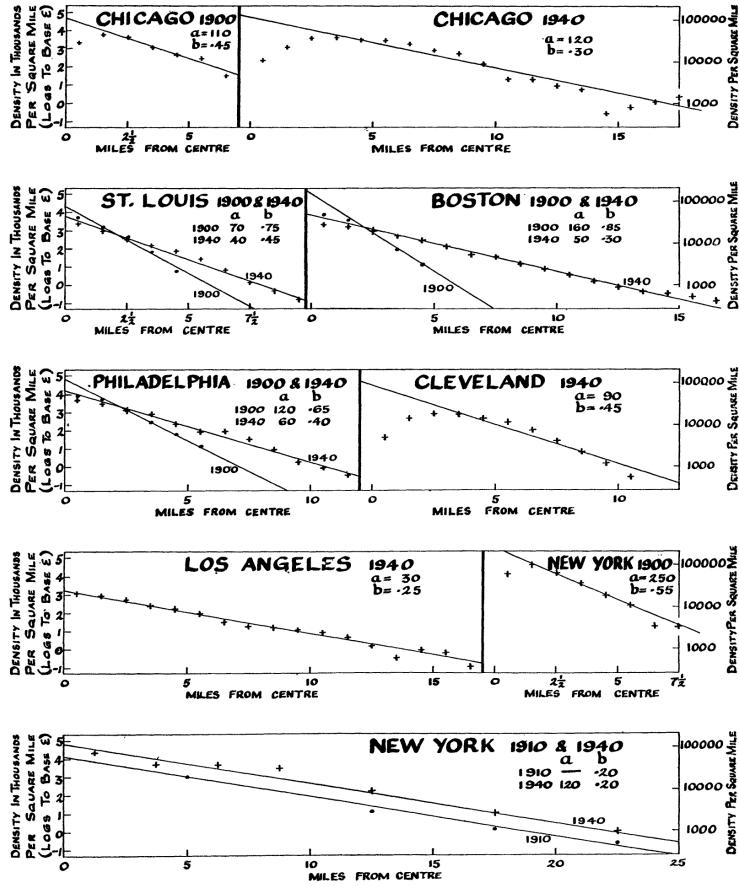
Chapter 3

¹³⁹⁰ The Monocentric City Model with Housing

3.1 Population density gradients in real life

In a classic study, Clark [1951] looks at the evolution of population density in cities from early in the industrial revolution until the mid-20th century. He starts by estimating gradients, but this time for population density rather than land prices or rents. To do this, let y be population density in a census tract. The size of a census tract varies a little bit from county to county, but in the US they usually range from 4000-5000 people, and (in urban areas) a few city blocks. Let x be distance from the center of the tract to the CBD. Then for each city and census year in his sample,
¹³⁹⁵ Clark uses all census tracts, i , in that city to estimate the population density gradient
¹⁴⁰⁰

Figure 3.1: Population density gradients



Note: Reproduced from Clark [1951], ©Journal of the Royal Statistical Society.

for that city,

$$\ln y_i = \ln A - Bx_i + \eta_i. \quad (3.1)$$

This is similar to what we've done before, but not exactly the same. Comparing to the land rent gradients we estimated in equation (1.1), we see that we previously looked at the relationship between the logarithm of the outcome and the *logarithm*

¹⁴⁰⁵ of the dependent variable. Clark is tweaking this by comparing the logarithm of the outcome and the *level* of the outcome. This still lets us ask similar questions. Does y increase or decrease as x increases? How fast? However, we lose the ability to interpret B as an elasticity, so we need to pay attention to the units we use.

Clark finds that A and B decrease over time for almost all the cities in his sample.
¹⁴¹⁰ To interpret a decrease in A , note that when $x = 0$, equation (3.1) reduces to $\ln y_i = \ln A$, so A is population density at the center. To interpret a decrease in B , suppose B changes from B^0 to B^1 , but we hold A constant. If we let $y^0(x)$ and $y^1(x)$ indicate the corresponding densities, we have

$$\ln y_i^1 - \ln y_i^0 = (B^1 - B^0)x_i,$$

or

$$\frac{y_i^1}{y_i^0} = e^{(B^1 - B^0)x_i}. \quad (3.2)$$

¹⁴¹⁵ For example, consider what happens one mile from the CBD of London, where $x = 1$, between 1901 and 1921, when B decreases from 0.45 to 0.35. Using equation (3.2) we have that density one mile from the CBD decreases by a factor of $e^{(B^1 - B^0)x_i} = \exp(-0.1) \approx 0.90$ between 1901 and 1921. Two miles from the center, population density decreases by a factor of $e^{(-0.2)} \approx 0.82$.

¹⁴²⁰ Figure 3.1 presents Clark's results for seven US cities. In each of these plots, the x -axis is miles to the CBD and the y -axis is log mean population density for the tracts at distance x . The solid lines describe the regression line resulting from an estimation of equation (3.1) for 1900 or 1940, and the dots are a histogram showing

mean population density by distance.

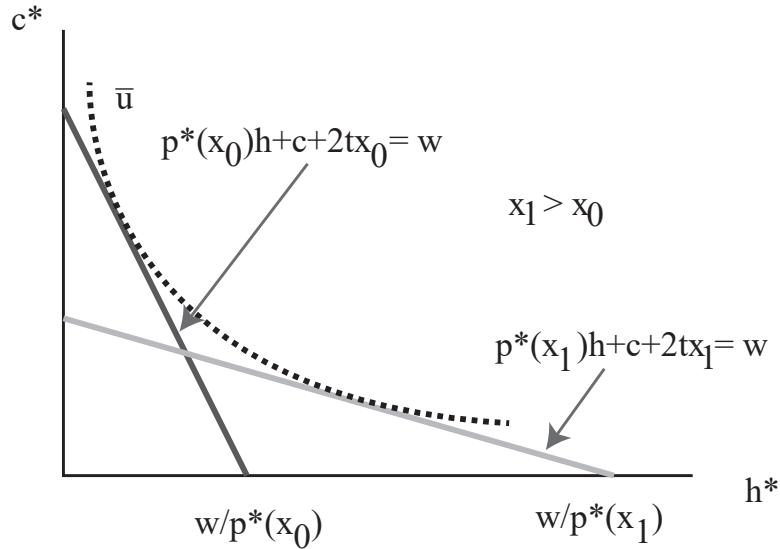
Several features of these graphs are noteworthy. First, the population density gradient is downward sloping in every case. Population density falls rapidly as we move away from the CBD in every city Clark looks at. Second, the density gradient gets flatter over time for every case reported in the figure, except for New York City between 1910 and 1940 where it is constant. Third, the intercept of the density gradient falls over time for almost every case illustrated in the figure, so central population density almost always falls over time. Fourth, the histograms show that the regression line and the histograms usually diverge close to the center. Population density is lower near the CBD than the estimated density gradient predicts. This almost surely reflects the fact that the centers of these cities are given over to employment and not residence. Fifth, central city population densities were sometimes as high as 100,000 per square mile in the early 20th century. This is much higher than in most modern developed world cities.

Finally, figure 3.1 was made around 1950. This is before computers. Not only was the figure drawn by hand, all of the regression results that it reports were probably calculated by hand. While the paper looks pretty crude by today's standards, for its time, it was a remarkable accomplishment.

3.2 Households and housing

As we developed it in Chapter 1, the monocentric city model assumes that household land consumption is the same for all households, fixed at \bar{l} . This means population density is constant by assumption. This is contradicted by the data we've just seen

Figure 3.2: Graphical analysis of the monocentric city model with housing



Note: *x* and *y* axes are housing and composite consumption. The dashed curve is an indifference curve describing the choices of housing and consumption that satisfy the free mobility condition, $u(c, h) = \bar{u}$. The dark gray line is a budget line for a household nearer the CBD, and the light gray line is a budget line for a more remote household. If households are rational, their budgets must be tangent to their indifference curve, as drawn. Because the more remote household spends more on commuting, and so the *y*-intercept of their budget line is lower, the budget line for more remote household can only be tangent to their indifferent curve if it is less steep. A flatter budget line for the more remote household requires that housing prices fall with distance to the center.

in section 3.1, not to mention common sense.

We now generalize the model to allow people to adjust their housing consumption. This will let us explain the three main facts about cities that we see in figure 3.1. That is; negatively sloped population density gradients, decreasing central population over time, and flatter population density gradients over time.

1450

In our initial formulation of the monocentric city model, households only consumed

the composite consumption good, c . Together with the free mobility condition, this required that all households end up with the same consumption level, c^* . When we introduce housing into the model, household utility levels will still be fixed by the outside option, but households achieve the reservation utility level with a mix of housing and consumption.¹⁴⁵⁵

How will this work? As households move closer to the center, land is more expensive to offset lower commuting costs, and so households substitute away from it. This can happen in two ways. First, households consume less housing. Second, households use less land, and more capital, to produce their housing (here, capital is all of the stuff other than land that goes into housing). This will mean that households nearer the CBD consume less land, or conversely, that population density is higher.¹⁴⁶⁰

This logic requires that cities have higher population densities near the CBD and that population density falls with commuting distance. That is, that population density gradients be negatively sloped. This is the first of the three facts we learned from the population densities in figure 3.1, two to go. How can we explain the flattening of density gradients that we see in figure 3.1? This is a comparative static result. When transportation costs fall, the model will require a flattening out of population density gradients. The model does not explain the drop in density right at $x = 0$.¹⁴⁷⁰

We need some new notation to set up the monocentric city model with housing. Let $c(x)$ be the consumption level of a household at location x , let $h(x)$ be housing consumption, and let $p(x)$ be the housing price gradient. This is not generally going to be the same as the land price gradient, $R(x)$, that we defined earlier. However, if we think that “housing” is just “land”, as it was in Chapter 1, then $p(x)$ and $R(x)$ ¹⁴⁷⁵

are the same. For the rest of this section, this is what we're going to assume. That is, for now, housing just consists of land, no capital. In section 3.3, we will generalize and require that the housing that households consume is made from both land and capital. Once we do this, land and housing will be different things in the model, and
¹⁴⁸⁰ they will each need their own price gradient.

We assume that households are “rational” and choose their favorite from among the alternatives available to them. We also made this assumption in the monocentric city model without housing. In that case, because utility depends only on the composite consumption good, rationality just requires that households spend everything
¹⁴⁸⁵ on consumption that does not go to rent or commuting, i.e., they don’t throw anything away. Once we add housing to the model, the household problem becomes a little more complicated. A rational household must choose the best combination of housing and consumption (and location).

If we assume that household preferences across consumption and location are given
¹⁴⁹⁰ by

$$u(c, h) = c(x)^\alpha h(x)^{1-\alpha}, \quad (3.5)$$

then at all occupied locations a household solves,

$$\max_{c,h,x} c(x)^\alpha h(x)^{1-\alpha} \quad (3.6)$$

$$\text{s.t. } w = c(x) + p(x)h(x) + 2t|x|. \quad (3.7)$$

This is a little cumbersome to write, so I'll often drop most of the x 's, and write it

like this,

$$\begin{aligned} & \max_{c,h,x} c^\alpha h^{1-\alpha} \\ \text{s.t. } & w = c + ph + 2t|x|. \end{aligned}$$

Here, the fact that c , h , and p vary with x is implicit.

¹⁴⁹⁵ To make this look a little simpler, let $\tilde{w} = w - 2t|x|$, income net of commuting. Using this notation, the household's problem becomes,

$$\begin{aligned} & \max_{c,h,x} c^\alpha h^{1-\alpha} \\ \text{s.t. } & \tilde{w} = c + ph. \end{aligned} \tag{3.8}$$

You may recognize this problem. It is a standard statement of the problem of a household that must choose how to allocate its endowment, here \tilde{w} , across two goods, here c and h .

¹⁵⁰⁰ The solution to this problem is two demand functions that describe a household's utility maximizing choice of consumption and housing for each possible price of housing and income. Solving for these demand functions, the details are in box 3.2.1, we have,

$$c^* = \alpha\tilde{w} \tag{3.9}$$

$$h^* = \frac{(1 - \alpha)\tilde{w}}{p}. \tag{3.10}$$

Notice that α is the share of income (net of commuting) used for consumption, and

1505 $1 - \alpha$ is the share that is used for housing. Recalling the results we saw in section 2.8, the share of income used for housing ought to be somewhere between 25% and 35%. That is, α is somewhere between about 2/3 and 3/4.

For an equilibrium, we need households to optimize. This is guaranteed if equation (3.9) and (3.10) hold. An equilibrium also requires that no one wants to move. 1510 To impose this condition, we need to calculate the “indirect utility function” by substituting h^* and c^* back into the utility function,

$$\begin{aligned} V(p, \tilde{w}; x) &= (c^*)^\alpha (h^*)^{1-\alpha} \\ &= \alpha^\alpha \tilde{w}^\alpha \left(\frac{(1-\alpha)\tilde{w}}{p} \right)^{1-\alpha} \\ &= \alpha^\alpha (1-\alpha)^{1-\alpha} \left(\frac{\tilde{w}}{p^{1-\alpha}} \right). \end{aligned}$$

For each x , $V(\cdot)$ is the highest utility that a household can achieve given housing price p and income net of commuting \tilde{w} .

We can now state the “no one wants to move” condition for the monocentric city model with housing as, 1515

$$V(p, \tilde{w}; x) = \alpha^\alpha (1-\alpha)^{1-\alpha} \left(\frac{\tilde{w}}{p^{1-\alpha}} \right) = \bar{u}. \quad (3.11)$$

This is the analog of equation (1.11) in the model without housing. The difference is that households achieve the reservation utility with an optimal bundle of housing and consumption, instead of just buying as much composite consumption as they can afford.

1520 By rearranging equation (3.11) we get,

$$p^* = \left[\frac{\alpha^\alpha (1-\alpha)^{1-\alpha} \tilde{w}}{\bar{u}} \right]^{1/(1-\alpha)}.$$

Recalling the definition of \tilde{w} , this becomes

$$p^* = \left[\frac{\alpha^\alpha (1-\alpha)^{1-\alpha}}{\bar{u}} \right]^{1/(1-\alpha)} (w - 2t|x|)^{1/(1-\alpha)}. \quad (3.12)$$

This is the housing price gradient.

1525 This expression for $p(x)$ lets us completely solve the model. If we substitute this expression into the expressions for housing and consumption, equations (3.9) and (3.10), we get consumption and housing gradients, $c(x)$ and $h(x)$. If we substitute $p(x)$ into equation (4.7), we can write the indirect utility level entirely in terms of known quantities. These are all of the quantities that the model is supposed to determine.

1530 The monocentric city model *without* housing predicts a linear land rent gradient. In fact, as we have seen, land and house prices decline more slowly as they are further from the center. We can check whether the monocentric city model *with* housing does any better.

If we take $(1-\alpha) = 1/4$ as a rough guess for the housing share of household expenditure, then equation (3.12) becomes

$$p^*(x) = \left[\frac{\frac{3}{4} \frac{3}{4} \frac{1}{4}}{\bar{u}} \right]^4 (w - 2tx)^4.$$

¹⁵³⁵ Using the chain rule, we can evaluate the first and second derivatives of $p^*(x)$ as,

$$\frac{dp^*(x)}{dx} = \left[\frac{\frac{3}{4} \frac{3}{4} \frac{1}{4}}{\frac{4}{\bar{u}}} \right]^4 (w - 2tx)^3 (-8t) < 0, \quad (3.13)$$

$$\frac{d^2p^*(x)}{dx^2} = \left[\frac{3}{4} \frac{3}{4} \frac{1}{4} \right]^4 (w - 2tx)^2 (24t^2) > 0.$$

Recalling that $w - 2tx$ has to be positive, households can't spend more than they have on commuting, then as we expect, the first derivative of $p^*(x)$ is negative. This means that the price of housing decreases with x . This is just what we found in the model without housing. We also see that the second derivative of $p^*(x)$ is positive. ¹⁵⁴⁰ This means that the price of housing decreases more slowly as x increases and we get farther from the CBD. This is just what we observed for land prices in figure 1.1. Figure 2.5 shows the same pattern for house price and pre-Covid rental gradients in the US.¹

So far, so good. By adding housing to the monocentric city model, we have a ¹⁵⁴⁵ non-linear house price gradient that decreases quickly near the CBD, and then more slowly further out, more like what we observe than the linear land rent gradient from the model without housing.

Substituting the expression for p^* from equation (3.12) back into the expression for h^* from equation (3.10), then we get an expression for the quantity of housing demanded at each location entirely in terms of quantities determined outside our ¹⁵⁵⁰

¹Recall from the discussion surrounding figure 1.2 that transforming a loglinear function like the ones reported in figure 2.5 to levels gives a gradient that decreases more slowly as x increases.

model. That is,

$$\begin{aligned} h^* &= \frac{(1-\alpha)\tilde{w}}{p^*} \\ &= (1-\alpha)\tilde{w} \left[\frac{\alpha^\alpha(1-\alpha)^{1-\alpha}\tilde{w}}{\bar{u}} \right]^{\frac{-1}{1-\alpha}} \\ &= \left[\frac{\alpha^\alpha}{\bar{u}} \right]^{-1/1-\alpha} \left[\frac{1}{\tilde{w}} \right]^{\alpha/1-\alpha}, \end{aligned}$$

and using the definition of \tilde{w} , we have,

$$h^* = \left[\frac{\alpha^\alpha}{\bar{u}} \right]^{-1/1-\alpha} \left(\frac{1}{(w - 2tx)} \right)^{\alpha/1-\alpha}. \quad (3.14)$$

By inspection of equation (3.14), we see that the equilibrium consumption of housing increases in x . This is just how households should respond to the decrease in the price of housing as they move farther from the CBD. If, as we're maintaining in this section, housing is just land, then increases to housing consumption at more remote locations are equivalent to reductions in population density, and this model gives us one of the main facts that we took away from figure 3.1; population density declines with distance from the center.

Also by inspection of equation (3.14), we see that the equilibrium consumption of housing increases as t decreases. This means that if transportation costs fall (say with the advent of the car or the construction of the interstate highway system) we should see the population density gradient increase everywhere (except right at $x = 0$), which means that the gradient has to flatten out. Thus, this model gives us the second main fact that we took away from figure 3.1; population density gradients get flatter over time.

It is hard to reconcile the monocentric city model with the fact that population density falls near the center of the city over time. Because commute costs don't affect the household at $x = 0$, changes to commute costs don't affect housing consumption at the CBD at all. Worse still, there is no other obvious comparative static that explains this phenomena. Income increased dramatically over this period. Inspection of $h^*(x)$ shows that housing consumption at $x = 0$ falls as income increases: as people have more to spend on consumption, all else equal, they require less housing to hit the reservation utility level. This is the opposite of what we see in the data.

Our analysis of the monocentric city model with housing has so far relied on a particular functional form for preferences over housing and consumption, equation (3.5). At best, this will be approximately right, so it's natural to wonder what happens if preferences are a little, or a lot different. It turns out that most of our conclusions are general. In particular, the housing price gradient is decreasing and the housing consumption gradient is increasing for any preferences over housing and consumption that can be represented by concave indifference curves.

Because it requires a solid understanding of vector calculus, we're going to skip working this out.² However, it turns out that we can do most of the analysis of the more general case graphically. Figure 3.2 is the main picture. The x and y axes are housing and composite consumption. Notice that, unlike most of the other graphs we've seen so far, *distance to the CBD is not one of the axes*. The curve is an indifference curve describing the choices of housing and consumption that satisfy the free mobility condition, $u(c, h) = \bar{u}$. Every household in the city must choose a combination of c and h on this line.

²If you are interested in the details, look at Brueckner's classic exposition of this model [Brueckner, 1987].

1590 Now consider two households, the first located at x_0 and the second at x_1 further from the CBD than the first, i.e., $x_1 > x_0$.

Because $x_1 > x_0$, if both households consume zero housing then the more remote household must have less of the composite good. With incomes the same, the longer commute means less left over. This means that the y -intercept of the budget line for the more remote household, the light gray line, must be lower than that for the more central household, the dark gray line. We also know that, given x and prices, both households are choosing the combination of c and h that maximizes their utility. That is, they are both solving equation (3.6). This requires that both budget lines be tangent to the indifference curve. This is just what figure 3.2 shows, and the only way 1600 this can happen is if the budget line for the more remote household is flatter than for the more central household. This in turn requires that the price of housing be lower and, housing consumption higher at the more remote location.

These are exactly the same conclusions we arrived at by solving the model analytically with a particular utility function; the price of housing declines and housing 1605 consumption increases for more remote households. However, in the graphical analysis, we can see that the result is more general. We get the same qualitative result as long as preferences can be represented by concave indifference curves like the one in figure 3.2.

3.3 The construction sector

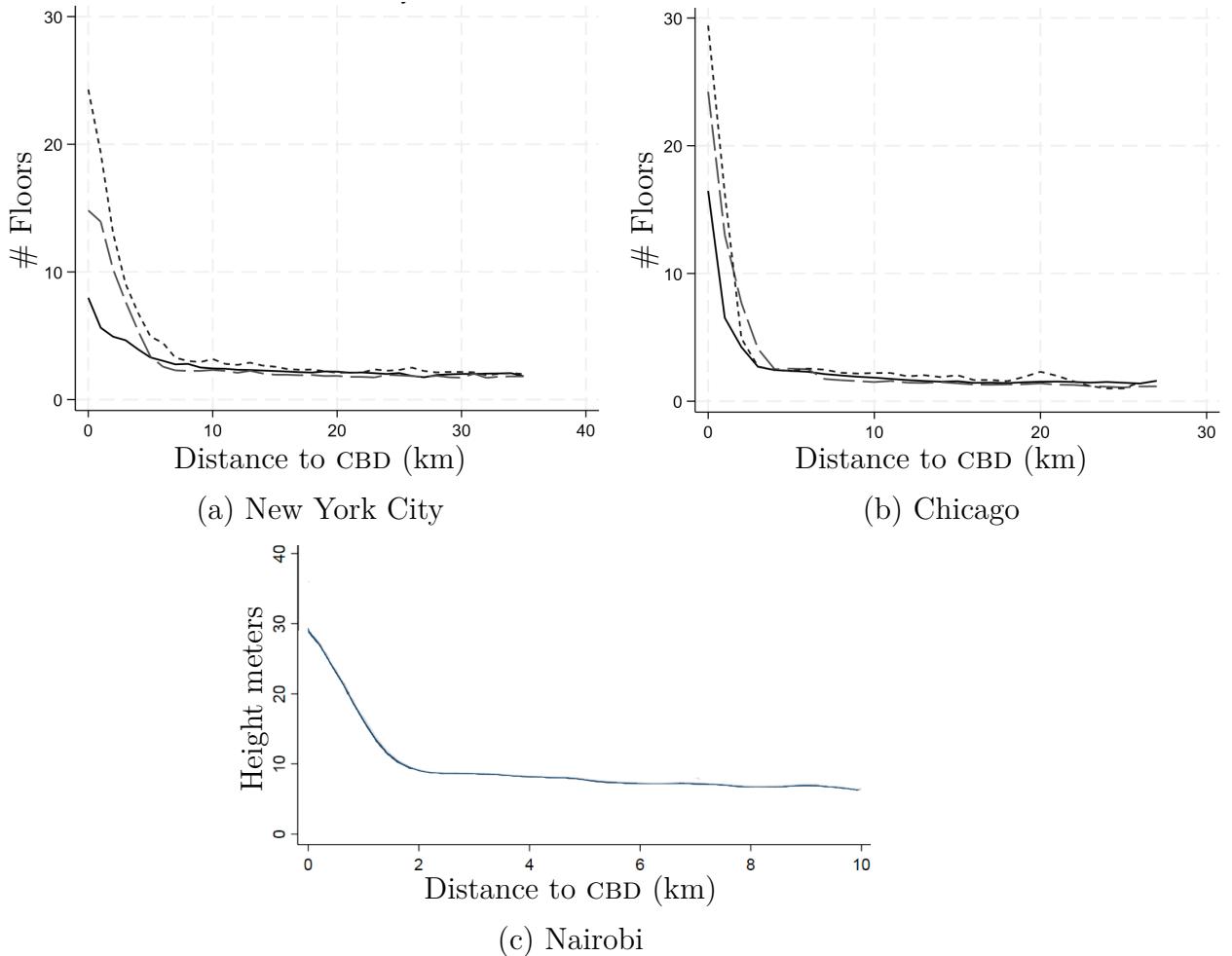
1610 While the monocentric city model with housing makes better predictions about the shape of the house price gradient than the model without housing, and does pretty

well at predicting the behavior of population density gradients, it has a couple of obvious failings. First, the model doesn't really describe "housing" at all. That is, so far there is not really a distinction between housing and land. Because people in cities live in buildings, a model of cities without buildings is obviously deficient.

A quick look at figure 1.4 makes the second problem clear. One of the defining features of cities is their skyline. However, building heights are much harder to study systematically than population density and little systematic information about building heights is available. With this said, the past few years, we have seen some progress, and figure 3.3, describes some of these results.

Figure 3.3 reports how building height varies with distance to the CBD for three cities, New York, Chicago and Nairobi. For Chicago and New York, the top two panels also show how the building height gradient changed over the course of the 20th century. Figure 3.3 shows us two things. First, it looks like cities everywhere are taller in the center than at the edges. To the extent that there are exceptions to this rule, they are places where land use is not organized by markets, e.g., Moscow during Soviet times. Second, cities are getting taller over time. This is at least partly due to improvements in the technology of building tall buildings. Third, building height gradients appear to be even steeper than population density gradients. The first skyscrapers appeared in the late 19th century when people first began building with iron and steel. Prior to this, buildings were all either brick or wood, and these technologies have a height limit of about six stories (although I have read of ten story wooden buildings in Rome that occupants navigated with ladders). The steep slope of building height gradients and the increase in building height are both confirmed in data describing a larger set of cities in Ahlfeldt and Barr [2022].

Figure 3.3: Building height gradients



Note: Two top panels show average building height in floors as a function of distance to the CBD in km for New York (a) and Chicago (b). Black is 1900-1939, dashed line is 1940-1989, and dotted line is 1990-2017. Building height declines rapidly with distance to the CBD, and buildings near the CBD of both have gotten taller over time. Bottom panel shows mean building height in meters as a function of distance to CBD for Nairobi around 2015. For reference, a building 10m tall is about three stories. Panels (a) and (b) courtesy of G. Ahlfeldt based on data in Ahlfeldt and Barr [2022], ©G. Ahlfeldt and J. Barr. Panel (c) based on Henderson et al. [2018a], ©J. V. Henderson, T. Regan, and A. Venables.

Box 3.2.1: Solving the household problem

This problem has exactly the same form as a common teaching example of the “household problem” that is one of the main topics in a microeconomics course.

To solve the household problem given in equation (3.8), reduce it to one variable by solving the constraint for c and substituting this into the objective. This gives us the equivalent, unconstrained maximization problem in one variable,

$$\max_{h,x} (\tilde{w} - ph)^\alpha h^{1-\alpha}$$

To solve, set the first order condition equal to zero and solve for h^* ,

$$\begin{aligned} \frac{d}{dh} (\tilde{w} - ph)^\alpha h^{1-\alpha} &= 0 \\ \implies -\alpha p(\tilde{w} - ph)^{\alpha-1} h^{1-\alpha} + (\tilde{w} - ph)^\alpha (1 - \alpha)h^{-\alpha} &= 0 \\ \implies -\alpha p(\tilde{w} - ph)^{\alpha-1} h + (1 - \alpha)(\tilde{w} - ph)^\alpha &= 0 \\ \implies -\alpha p(\tilde{w} - ph)^{-1} h + (1 - \alpha) &= 0 \\ \implies -\alpha ph + (1 - \alpha)(\tilde{w} - ph) &= 0 \\ \implies -ph + (1 - \alpha)\tilde{w} &= 0. \end{aligned}$$

Rearranging, we have the household’s demand function for housing

$$h^* = \frac{(1 - \alpha)\tilde{w}}{p}. \quad (3.3)$$

Substituting h^* back into the budget constraint, we get

$$\begin{aligned} c^* &= \tilde{w} - p \left[\frac{(1 - \alpha)\tilde{w}}{p} \right] \\ &= \alpha\tilde{w}. \end{aligned} \quad (3.4)$$

This is the household’s demand function for composite consumption.

Box 3.3.1: Constant returns to scale

A constant returns to scale function has the property that if you increase the input(s) by some factor, then output increases by the same factor. The Cobb-Douglas function that is often used as a teaching example, and that appears throughout this book, is a good example.

If $H_s(k, l) = k^\beta l^{1-\beta}$, and I scale both k and l by any factor $\lambda > 0$, then

$$\begin{aligned} H_s(\lambda k, \lambda l) &= (\lambda k)^\beta (\lambda l)^{1-\beta} \\ &= \lambda^\beta k^\beta l^{1-\beta} \\ &= \lambda H_s(k, l). \end{aligned}$$

For example, if $\lambda = 2$, then this means that doubling inputs exactly doubles outputs. Hence the name, “constant returns to scale”.

In order to think about buildings explicitly, we assume that housing, still h , is built from land, l and physical capital k . The price of housing is p , the price of land is R (like before), and the price of capital is i . Both p and R vary with x , but i does not.

1640 A housing sector transforms land and capital into housing according to

$$H_s(x) = k(x)^\beta l(x)^{1-\beta}, \quad 0 < \beta < 1.$$

$H_s(x)$ describes the amount of housing available at each distance from the center, the housing supply. To lighten notation, as we did when discussing household behavior, I

will often suppress the location argument when I write the housing technology. This is going to give us,

$$H_s = k^\beta l^{1-\beta}, \quad 0 < \beta < 1.$$

¹⁶⁴⁵ This function is constant returns to scale (see box 3.3.1 for definition). Because $H_s(k, l)$ is constant returns to scale, we have

$$H_s = k^\beta l^{1-\beta} = lk^\beta l^{-\beta} = l \left[\frac{k}{l} \right]^\beta$$

So we can write “housing supply per unit of land” as

$$h_s(S) = S^\beta, \tag{3.15}$$

where $S = k/l$, the capital to land ratio, i.e., building height.

Constant returns to scale functions make our analysis a lot easier. It also seems ¹⁶⁵⁰ to be a pretty good description of how the construction of housing actually happens. Our best evidence on this point is that housing production is actually close to constant returns to scale, and that $\beta \approx 2/3$ [Combes et al., 2020].

Note that we will ultimately require that housing markets clear at every location. This requires that the number of households at x (really the density of households ¹⁶⁵⁵ at x), times the housing consumption per household, $h(x)$, equals the total housing supply, $h_s(x)$.

Assume that developers maximize profits, and that the housing production sector is perfectly competitive. This seems like an easy to defend assumption about how this

market works. There are lots of developers and builders, and developers are nothing
₁₆₆₀ if not profit maximizes.

Restating the developer's problem in math, we have,

$$\max_{k,l} pk^\beta l^{1-\beta} - ik - Rl,$$

where p and R vary with x , but i does not. This is easier to tackle if we divide through by l and write the problem in terms of profit per unit of land, rather than profit. Recalling that $S = k/l$ we can write the developer's problem as,

$$\max_s pS^\beta - iS - R. \quad (3.16)$$

₁₆₆₅ We can use this trick for any housing production function that satisfies constant returns to scale. We're working through a particular example to avoid using a lot of vector calculus, but our results extend to any constant returns to scale housing production function. To see the details, look at Brueckner [1987].

The developer's problem, equation (3.16), is an unconstrained maximization problem in one variable. Solve by taking the derivative with respect to the choice variable, S , setting this derivative equal to zero, and then solving for the value of S that satisfies this equation.

The first order condition for equation (3.16) is

$$p\beta S^{\beta-1} - i = 0.$$

Rearranging, we have

$$S = \left(\frac{i}{\beta} \right)^{1/\beta-1} p^{1/1-\beta} \quad (3.17)$$

- ¹⁶⁷⁵ This is the first of two main equations that govern the behavior of the construction sector.

Because the market is competitive, developers enter the market until profits are driven to zero. This gives us the second equation that governs the behavior of the housing construction sector,

$$pS^\beta - iS - R = 0.$$

- ¹⁶⁸⁰ Notice the similarity between this condition and the free mobility condition for households. This free entry condition is really a free mobility condition for firms. We are assuming developers will always change where they build for a tiny deviation from zero profits, just as households will move anywhere for a small change in utility.

Rearranging, we have

$$R = pS^\beta - iS. \quad (3.18)$$

- ¹⁶⁸⁵ Remember that profit maximization and free entry, that is, equations (3.17) and (3.18) have to hold at all locations x , even though we haven't written the x 's explicitly.

In an equilibrium, we need a land rent gradient such that both equations (3.17) and (3.18) hold at all x . To find such a gradient, substitute S from equation (3.17)

into equation (3.18). This lets us write land rent as a function of housing price,

$$R = p \left[\left(\frac{i}{\beta} \right)^{1/\beta-1} p^{1/1-\beta} \right]^\beta - i \left[\left(\frac{i}{\beta} \right)^{1/\beta-1} p^{1/1-\beta} \right].$$

1690 After a lot of algebra,³ this gives,

$$R = \left(\frac{i}{\beta} \right)^{\frac{1}{\beta-1}} \left[\left(\frac{i}{\beta} \right) - i \right] p^{1/1-\beta}. \quad (3.19)$$

We would like to know is whether the land rent gradient is decreasing. But we already know the housing price gradient from equation (3.12), and that it is decreasing. By inspection, (3.19) is increasing in p (as long as $\beta < 1$). Thus, R increases in p and p decreases in x , so R must also decrease in x . We can use similar reasoning to extend 1695 the other comparative statics that we derived for the house price gradient to the land price gradient.

Recalling equation (3.17), we can write building height $S(x)$, in terms of $p(x)$. Again using the fact that $\frac{dp(x)}{dx} < 0$, we conclude that $S(x)$ is decreasing in x . That is,

³For the skeptics:

$$\begin{aligned} R &= p \left[\left(\frac{i}{\beta} \right)^{1/\beta-1} p^{1/1-\beta} \right]^\beta - i \left[\left(\frac{i}{\beta} \right)^{1/\beta-1} p^{1/1-\beta} \right] \\ &= p \left[\left(\frac{i}{\beta} \right)^{\beta/\beta-1} p^{\beta/1-\beta} \right] - i \left[\left(\frac{i}{\beta} \right)^{1/\beta-1} p^{1/1-\beta} \right] \\ &= \left[\left(\frac{i}{\beta} \right)^{\beta/\beta-1} p^{1/1-\beta} \right] - i \left[\left(\frac{i}{\beta} \right)^{1/\beta-1} p^{1/1-\beta} \right] \\ &= p^{1/1-\beta} \left[\left(\frac{i}{\beta} \right)^{\beta/\beta-1} - i \left(\frac{i}{\beta} \right)^{1/\beta-1} \right] \\ &= \left(\frac{i}{\beta} \right)^{\frac{1}{\beta-1}} \left[\left(\frac{i}{\beta} \right) - i \right] p^{1/1-\beta}. \end{aligned}$$

in equilibrium, the monocentric city model predicts that building height will decrease with distance to the CBD, just as we see in figures 3.3. Because the housing supply at each location, $h_s(x)$ is an increasing function building height, it follows that the housing supply is also decreasing with distance to the CBD. As was the case for land rent, $R(x)$, $S(x)$ and $h_s(x)$ also inherit the same qualitative comparative statics as $p(x)$. For example, all three get flatter as unit transportation costs fall.

We're now in a position to see whether the monocentric city model with housing can predict the behavior of the population density gradient, the problem that we began this Chapter with. To do this, we need to use one last assumption about equilibrium, housing markets clear. This assumption is really just bookkeeping. It says that everyone lives somewhere, and there are no empty houses. A little more formally, define $D(x)$ to be population density, though it is probably more precise to call it household density. It is the number of households per unit area at distance x from the CBD.

If the housing market clears, it must be that the number of houses available at x is equal to the number of households at x , or in math,

$$D(x)h(x) = h_s(x). \quad (3.20)$$

That is, the number of households per unit of land times housing per household per unit land equals the supply of housing per unit of land. If we rearrange equation (3.20), we get

$$D(x) = \frac{h_s(x)}{h(x)}.$$

This is our population density gradient.

Recalling equation $h_s(x) = S(x)^\beta$ is housing per unit of land. In the language of zoning regulations, this is pretty close to “floor to area ratio”, the amount of floor area per unit of parcel area. $h(x)$ is housing per household. The amount of housing consumed by a household located at x . It follows that the units of $D(x)$ are

$$\frac{\text{housing/unit land}}{\text{housing/household}} = \frac{\text{households}}{\text{unit land}}.$$

That is, population density.

We know that h_s decreases with distance to the CBD and h increases. It follows that $D(x)$, population density, decreases with distance to the CBD. Thus, the monocentric city with housing also correctly predicts the behavior of the population density gradient.

3.4 Conclusion

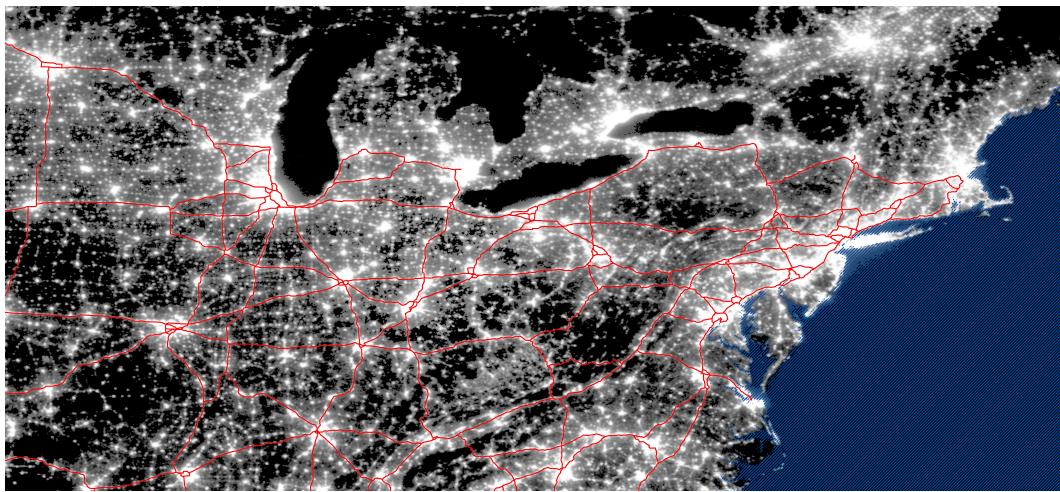
The monocentric city model makes the following assumptions about households: commuting is costly; households optimize; in equilibrium no one wants to move; people work in the center; and, the indifference curves that describe preferences for housing and consumption are concave. The model makes the following assumptions about the production of housing: developers maximize profits; there is free entry of developers; and, the housing production function can be written in terms of the capital-land ratio (because it is constant returns to scale). Finally, the model assumes that housing markets clear. Except for “everyone works in the center”, which (in the words of Mark Twain) seems like a bit of stretcher, these assumptions all line up with common

sense.

Here are some of the stylized facts we've established about cities: housing price gradients are downward sloping and get flatter over time; land rent gradients are downward sloping; building heights decrease with distance to the center; and, population density decreases with distance to the center. The monocentric city model with housing predicts every one of them. That is, it provides us with a way to think about how some of the most obvious features of cities emerge as a consequence of individual optimizing behavior.

With this said, the monocentric city model is far from perfect. Here are two of its more important failures, and with a little thought, you will probably be able to add to this list. To start, cities are not monocentric. Figure 3.4 shows lights at night and the interstate highway network for the Midwestern and Northeastern US. This figure is complicated and deserves a careful look. There are many small, uniformly round cities that seem consistent with the symmetric round geography required by the monocentric city model, but the larger cities, for example the Chicago and New York metropolitan areas, clearly do not. In Chapter 6 we will present a model capable of explaining this sort of complexity (although this model will rely on much stronger assumptions about how people behave). The figure also suggests that there might be a pattern in how cities are organized, with smaller cities surrounding larger ones. We will investigate what is known about this in Chapter 9. Second, much of the decrease in density with distance to the center occurs because there is more open space, not because people have bigger yards, but because tracts of undeveloped land are more common as we get further from city centers [Mieszkowski and Smith, 1991]. This sort of discontinuous development is often called 'Leapfrog' development; development

Figure 3.4: Lights at night and the interstate highway network



Note: 2007 *Lights at night* image of the Midwestern and Northeastern US. Cities are pretty clearly not all monocentric, though some of them are.

jumps over undeveloped land. This is pretty hard to explain with the monocentric city model (see Turner [2005] for an effort to make this work out).

While the monocentric city allows us to organize and explain many of the obvious features of cities, there are many questions that it cannot address. For example, what if people are not all identical? Why do people go to the center? How many cities are there? Why do cities specialize in different things? These are important questions that the model doesn't really speak to, and that we'll consider in the following Chapters.
1765

₁₇₇₀ **Problems**

1. This problem will examine the change in population density and its gradient over time based on Clark's 1951 study of population density in major metropolitan areas in Australia, the US and Europe.

₁₇₇₅ (a) Using Table 1 from Clark (1951), calculate the population density of London at the CBD ($x=0$), and 3 miles from the CBD, in 1801, 1841, and 1939.

(b) How does the ratio of population density at the CBD to the population density three miles from the CBD change between 1801 and 1939?

2. This problem will examine the relationship between population density and transportation costs. In 1, you saw that the population density gradient flattened out (there is relatively more population at $x = 3$ compared to $x = 0$ over time). This problem examines if decreased transportation costs could explain this flattening of the density gradient.

₁₇₈₀ Assume we have the setup of the monocentric city model with housing, as in the lecture. Assume as well that housing production is perfectly competitive. Let $\bar{u} = 3$.

(a) Let the household's problem be given by:

$$\max_{c,h,x} c^{1/2}h^{1/2} \text{ subject to } w = c + ph + 2tx$$

Let $\tilde{w} = w - 2tx$. Use the first-order condition of the household's problem with respect to h to find h^* in terms of p and \tilde{w} .

- 1790 (b) Use the fact that utility is $\bar{u} = 3$ everywhere to solve for p^* in terms of \tilde{w} .
- (c) Substitute your expressions for p^* and \tilde{w} into your expression for h^* to write h^* in terms of w , t , and x .
- (d) Let the developer's problem be given by

$$\max_S pS^{2/3} - iS - R$$

1795 where S is the capital to land ratio, and p , i and R are the costs of housing, capital, and land, respectively. For the remainder of the problem, let $i = \frac{1}{33}$.

Comment: The technology for producing housing is constant returns to scale and can be written as $h_s(S) = S^{2/3}$. Here, h_s is housing supplied, and is (with constant returns to scale) units of housing per constant area. This is NOT the same as h in the household problem, which is housing units per person.

1800 Use the first-order condition of this problem with respect to S to solve for h_s^* in terms of p .

- (e) Substitute in your expression for p^* from earlier to obtain an expression for population density, $\frac{h_s^*}{h^*}$, in terms of w , t and x .
- (f) Solve for the population density at $x = 1$ and $x = 2$ for $t = 1$ and $t = 0.5$. How does population density outside of the city center change when transportation costs fall?
- (g) In order to obtain a more general result about the population density gradient and transportation costs, take the derivative of your expression

1810

for population density with respect to t .

- (h) Evaluate the derivative from the previous part at $x = 0$. Does your expression for population density at $x = 0$ depend on t ?
- (i) Based on your answers to the previous two parts of this question, could this model of falling transportation costs explain the decreasing population density at the CBD and flattening of the population density gradient that you examined in the previous problem?

Chapter 4

Urbanization in the Developed ¹⁸²⁰ (mostly US) and Developing World

The scale of urban development since the industrial revolution has been geological.

In 2025, at the time of this writing, about 55% of the world's population of 8.2 billion is urban, about 4.5 billion people. That is, more people live in cities now than there were in the world in 1980. In this Chapter we investigate whether the same basic logic that we used to describe the internal structure of cities can help us to understand the process of urbanization.¹⁸²⁵

Because the developed world, mainly North America and Europe, was substantially urbanized by early in the 20th century, and the process is still ongoing in much of the developing world, I'll treat the two regions separately. In fact, there is an ongoing debate whether the economic forces behind urbanization in both regions are the same. To the extent that understanding why people in the developing world are moving to cities (or not) can affect urban policy in these places, arriving at a correct¹⁸³⁰

understanding of the forces behind the process of urbanization can affect the lives of billions of people.

¹⁸³⁵ 4.1 Urbanization in the US

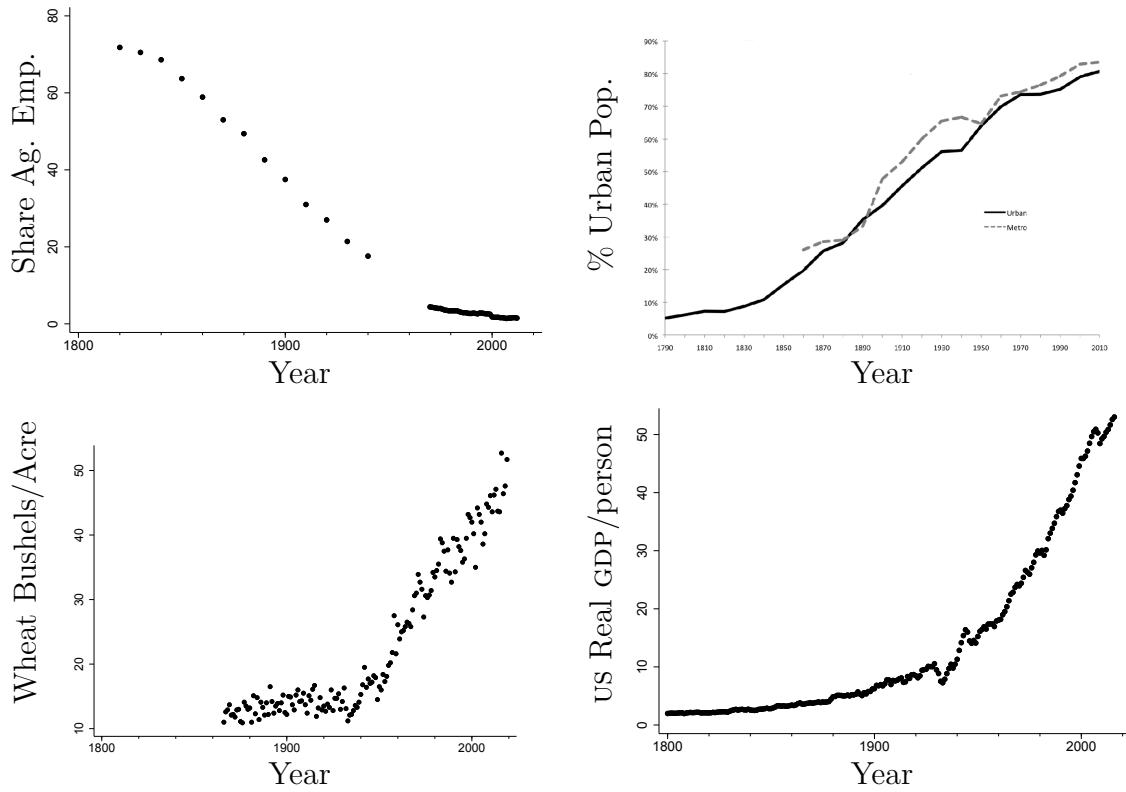
The US saw a dramatic increase in its urban population over the course of the 19th and 20th centuries. This section lays out the evidence that this process of urbanization reflected a tradeoff between rising urban productivity on the one hand, and on the other, innovations in public health that improved the disease environment in cities. ¹⁸⁴⁰ This done, I'll show that the monocentric city model and the idea of spatial equilibrium seems to fit these facts pretty nicely.

Figure 4.1 reports on the development of the US in the 19th and 20th centuries. The top left panel shows the national share of agricultural employment. The data are by decade from 1800 to 1940, and then annually from 1972 on. This figure shows that agricultural employment fell steadily from about 70% of the workforce in 1800 ¹⁸⁴⁵ to 1.5% in 2012. We have a lot fewer farmers than we used to.

The top right panel shows the share of people living in urban areas from about 1780 until about 2010, the solid black line, and the number of people living in MSAs from about 1870 to 2000, the dashed gray line. However you measure it, the share of ¹⁸⁵⁰ people living in cities has increased rapidly, from about 8% in 1800 to about 80% in 2010. In 1800, rural Americans outnumbered those living in cities by about 13 to 1. By 2000, urban residents outnumbered rural by about 4 to 1.

The bottom left panel shows how this was possible. This panel of figure 4.1 reports mean US wheat yields per acre from about 1860 until the early 21st century. Yield per

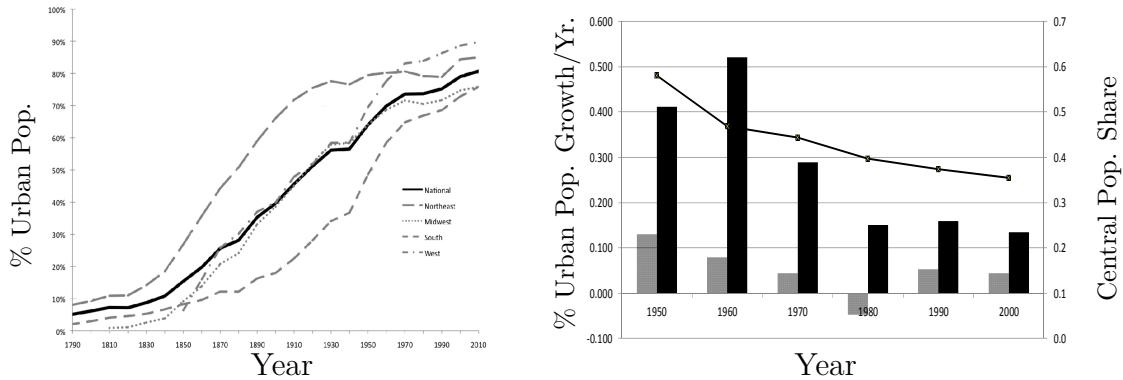
Figure 4.1: US development in the 19th and 20th centuries



Note: *Top left: Percent of Employment in Agriculture in the United States, Annual, FRED Graph Observations, Economic Research Division Federal Reserve Bank of St. Louis. The agricultural share of employment has declined from about 72% in 1820 to about 1.5% in 2012. Top right: Urban share increases as agricultural share decreases. Bottom left: Wheat yields 1866-2019 from the US Historical Census. Agricultural yields have increased more than fast enough to keep everyone fed. Bottom right: Real per capita GDP in constant 2011 dollars from Bolt and Van Zanden [2014]. From 1800 to 2016, US incomes increased from 1980\$ to 53015\$, a factor of about 27. Top right panel reproduced from Boustan et al. [2018] ©Oxford Publishing Limited.*

acre increased from about 10 bushels per acre to above 50. This increased agricultural productivity allowed the ever smaller share of farmers to feed the ever larger share of

Figure 4.2: Spatial patterns of US urbanization in the 19th and 20th centuries



Note: *Left panel breaks out urban share of population reported in figure 4.1 by region. A majority of the population in the South was rural until well into the 20th century, the Northeast crossed this threshold nearly a century earlier, while the Midwest and West split the difference. Right panel shows urban population growth by central city and suburb. Consistent with what we saw in figure 3.1 for an earlier period, most of the growth of US cities was at their edges throughout the second half of the 20th century. Both figures from Boustan et al. [2018], ©Oxford Publishing Limited.*

non-farmers.

Finally, the bottom right panel reports on US per capita GDP, loosely, a measure of material abundance. In terms of constant 2011 dollars, US GDP per capita increased from about 1980\$ in 1800 to about 53,000\$ in 2016, a factor of about 27. Notice that, as rapidly as agricultural yields increased, overall output increased even faster.

Figure 4.2 provides a more detailed description of how and where US cities grew. The left panel reports the urban population share from figure 4.1 by region. This figure shows that urbanization occurred much more slowly in the South than in the more industrialized North, with the Midwest and the West in between. A majority of the population in the South was rural until well into the 20th century and the

Northeast crossed this threshold nearly a century earlier. The right panel shows urban population growth by central city and suburb. This figure shows that the long history of urban decentralization that we saw in figure 3.1 persisted in the US, at least through the end of the 20th century.

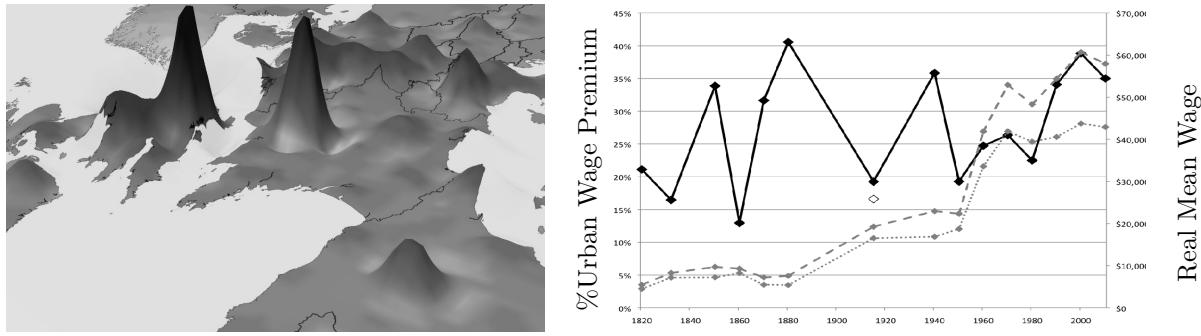
At the beginning of the 19th century, the US was a rural nation with most people employed as not very productive farmers. Over the course of the next two centuries the US became a nation dominated by more and more productive city dwellers with a progressively smaller and more productive farming sector. The Northeast urbanized a little earlier than the rest of the country, and most of the growth of cities in the late 20th century happened in the suburbs.

4.2 Urban productivity premium

The more rapid increase of GDP than wheat yields that we see in figure 4.1 suggests that cities may be more productive than the countryside, while the increasing share of urban population suggests that economic activity is increasingly concentrated in cities.

Figure 4.3 provides more evidence about urban and rural productivity. The left panel is a map of Western Europe that represents the level of economic activity in a grid cell as height. Economic activity is extremely concentrated in cities, and the data on which this figure is based suggests a similar urban concentration of economic activity in most countries of the world. The right panel of figure 4.3 is complementary. The gray dashed and dotted lines reports average urban and rural wages in the US between 1820 and 2010. In the US, wages are higher in cities. Because wages reflect

Figure 4.3: Urban versus rural productivity



Note: *Left panel illustrates the concentration of economic activity in Western Europe. Height reflects the level of economic activity in a small grid cell. The two highest peaks are London and Paris. Cities occupy a small fraction of the landscape and economic activity is concentrated in the cities. Even within cities, economic activity is concentrated close to their centers. Figure based on the G-Econ data for 2005 [Nordhaus, 2006]. Right panel reports the urban wage premium in the US from 1820 to 2010. The dotted line reports the mean rural wage, the dashed line reports the mean urban wage, and the black line reports the percentage by which the urban wage exceeds the rural wage. Urban wages consistently exceed rural wages. This gap fluctuates dramatically around an about 30% average premium and there is no obvious trend. Right panel reproduced from Boustan et al. [2018], ©Oxford Publishing Limited.*

the marginal (revenue) product of labor, this tells us that people in cities have been
 1890 more productive than people in the country throughout the history of the US. The
 solid black line shows the urban wage premium. This line fluctuates dramatically
 over time, but shows no obvious trend. On average, over two centuries, an urban
 worker in the US earned about 30% more than a rural worker. The concentration of
 economic activity and higher labor productivity in cities appears to be a feature of
 1895 economic activity almost everywhere in the modern world and throughout US history.

Looking carefully at Figure 4.3 we see that economic activity is more concentrated

in bigger cities than in smaller cities. The spikes for London and Paris are bigger and steeper than the spikes for Madrid and Milan. This suggests that people are more productive when they work in larger cities than smaller.

1900 The left panel of figure 4.4 provides more systematic evidence for the effect of city size on productivity using data on US MSAs in 2000. The y -axis in this figure is $\ln(\text{Gross Metropolitan Product per person})$ and the x -axis is $\ln(\text{Metropolitan Population})$. The upward trend in the data means that US cities are more productive as they are larger.

1905 The trend line plots the regression,

$$\ln(\text{Gross Metropolitan Product per person}) = A + B \ln(\text{Metropolitan Population}).$$

The slope of this line is about 0.13. Recalling Box 1.2.1, this means that the elasticity of income per person to city population is 0.13. Increasing city population by 10% increases the output of each urban worker by about 1.3%. This increasing relationship between city size and productivity is usually called an “agglomeration economy”. We’ll 1910 return to thinking about this in Chapter 8.

These sorts of agglomeration economies give a reason for people to be in cities. They are central to the study of urban economics, and are the subject of Chapter 8. Summarizing results from Chapter 8, to the extent that we can determine, and data going back much before the middle of the 20th century is scarce, agglomeration 1915 economies are a persistent feature of cities. Moreover, and also to the extent that we can check, agglomeration economies occur in all modern cities and their magnitudes are probably about the same almost everywhere.

A third fact is also important. The effect of agglomeration economies on worker output is at least partly causal. Big city workers are not more productive than small city workers entirely because they are harder working, more educated, or more ambitious. Similarly, workers in big cities are not more productive just because big cities are located in inherently more productive places like a good natural harbor or important natural resources. Big city workers are more productive than small city workers, in part, because they are in bigger cities. We postpone a discussion of the evidence for this claim until Chapter 8.

The right panel of figure 4.4 shows that the relationship between city size and productivity is probably more complicated than the discussion above allows. This figure, reproduced from Hsieh and Moretti [2019], reports dispersion of the MSA mean of the logarithm of mean city wages across in 1964 and 2009. Conceptually, this graph is simple, but its details are complicated.

The graph describes wages for 220 MSAs in two years. Calculating the x coordinate for each MSA in each year is done as follows. First, calculate the average wage in each MSA year, this gives 220 MSA average wages each year. Second, take the logarithm of each MSA-year average wage. Third, calculate the average log average wage (not a typo!) for each MSA year. Fourth, subtract this average log average wage from each MSA year log average wage. This process is called “de-meaning” and normalizes a population of observations so they have mean zero. In this case, it gives us a sample of 220 de-means MSA average log average wages. Because we’ve demeaned the data, the distributions for both years have the same zero mean. This makes it easy to compare how much dispersion there is, just what the figure does.

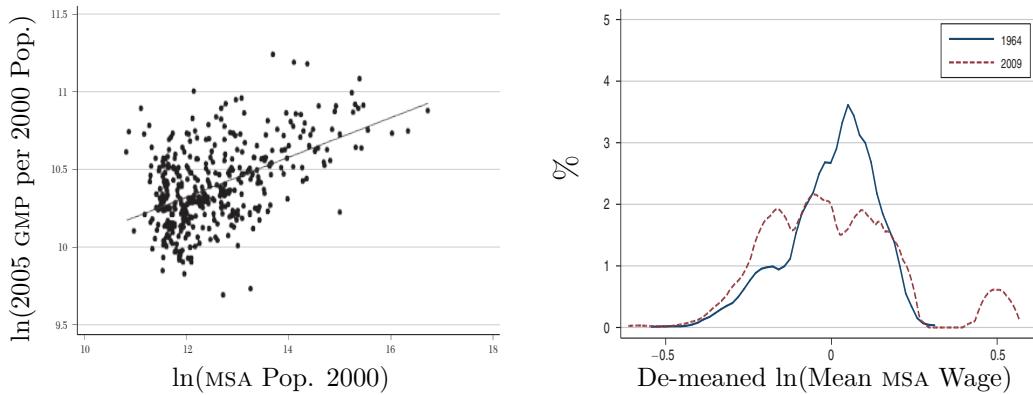
In jargon, the dashed line that describes a sample of 220 MSAs in 2009 is a prob-

ability density. The height of this line tells us the percentage of MSAs with the corresponding x value. Thus, about 0.02, or 4.4 of the 220 MSAs have de-meaned log average wage around zero in 2009. About a quarter as many have de-meaned log average wage around 0.5. For the MSA with $x = 0.5$, demeaned log average wage is about 0.5 larger than for an MSA with $x = 0$. Recalling exponentiation is the inverse operation of the logarithm, this means that the average wage for the MSA at $x = 0.5$ is about $e^{0.5} \approx 1.6$ times as large as for an MSA with $x = 0$.

If we compare the solid line describing average log average wage for 1964 to the dashed line describing 2009, we see that in 2009 a greater share of MSAs had log average wages far from zero than in 1964. This means that more MSAs were far from the average, or that there was more wage dispersion across cities in 2009 than 1964. In light of what we have learned about the relationship between city size and productivity, you might wonder if the increased cross-MSA dispersion of wages simply reflects increased dispersion of city sizes. Hsieh and Moretti [2019] show that nothing of the sort seems to have occurred. Rather, something about US cities changed around the turn of the 21st century so that cities, or the people in them, have become more different from each other.

Even though labor productivity is systematically higher in larger cities, the relationship between city size and productivity is more complicated than a simple rule relating population size or density and wages.

Figure 4.4: Urban productivity in the late 20th century US



Note: For the left panel, the y-axis is $\ln(\text{Gross Metropolitan Product per person})$ and the x-axis is $\ln(\text{Metropolitan Population})$. The upward trend in the data means that US cities are more productive as they are larger. The regression line plots the line $\ln(\text{Gross Metropolitan Product per person}) = A + B \ln(\text{Metropolitan Population})$. The slope of this line is about 0.13, so the elasticity of income per person to city population is 0.13. The right panel shows distributions of de-means log average wages across 220 MSAs in 1964 (solid) and 2009 (dashed). Wage dispersion is increasing over time. Left panel reproduced from Glaeser and Gottlieb [2009]. Right panel reproduced from Hsieh and Moretti [2019]. Both panels ©American Economic Association.

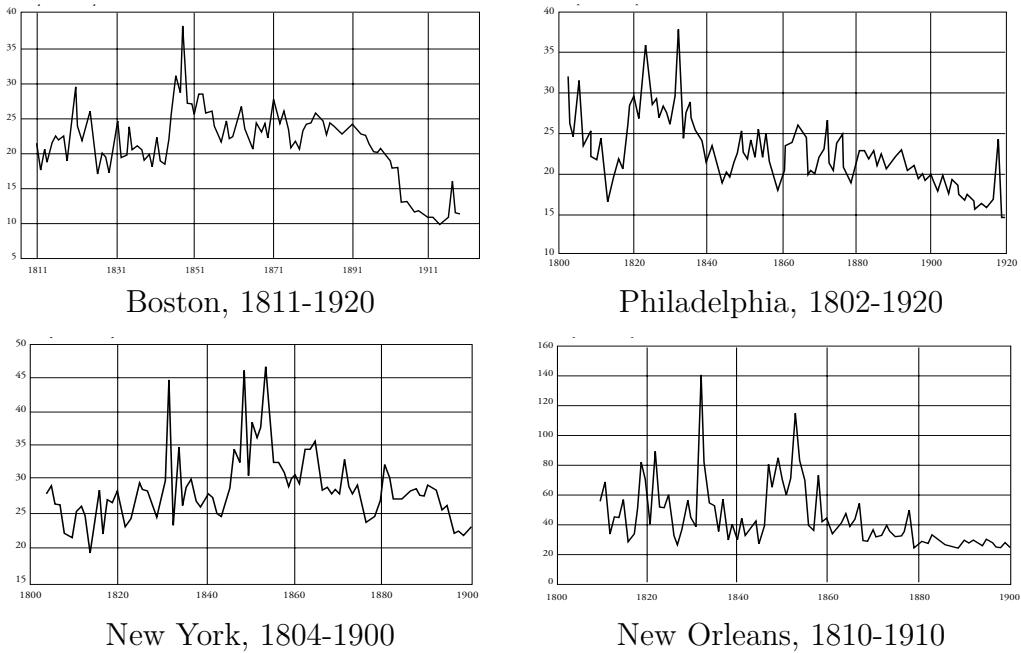
4.3 Excess urban mortality

In 1960, about 9.5 out of every 1000 people in the US died. This statistic, called the crude death rate, decreased slowly to about 8.5 by 2010, before spiking to 10.5 during Covid. At the time of this writing (in late 2024), it has fallen back to about 9.5.¹

Historically, crude death rates have been much higher, and in particular, they have been much higher in cities. Figure 4.5 shows the history of the crude death rate in Boston, Philadelphia, New York and New Orleans during the 19th and early 20th

¹United Nations Population Division. World Population Prospects: 2022 Revision

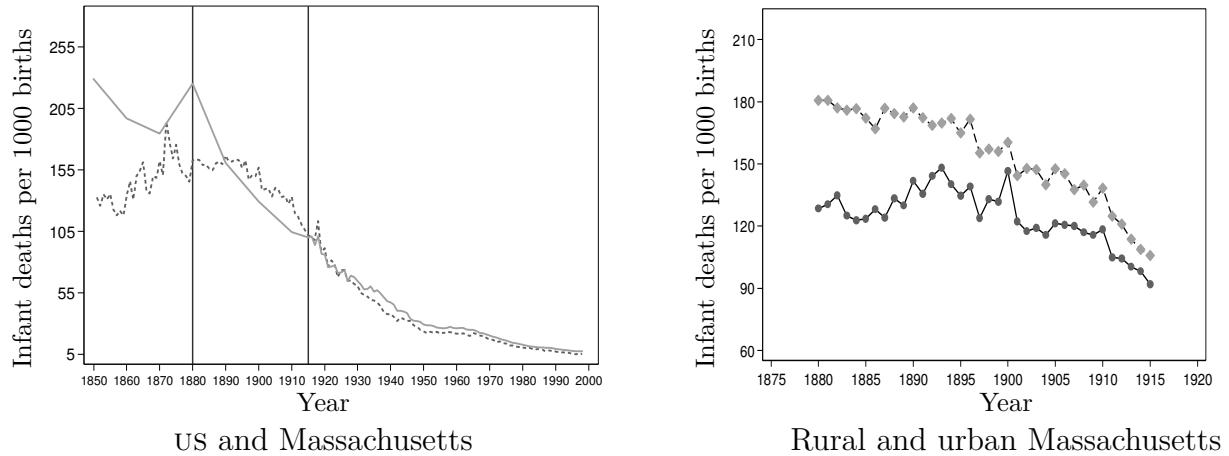
Figure 4.5: Crude death rates in four US cities in the 19th century



Note: *Crude death rates over the course of the 19th century in Boston, Philadelphia, New York and New Orleans. Crude death rates usually ranged between 20-80 until about 1860, and then began to decline. The frequency of epidemics also fell during this time. New Orleans was dramatically more dangerous than cities in the Northeast (note the different y-axis scale for New Orleans). Figures reproduced from Haines [2001].*

centuries. Three features of these graphs are noteworthy. First, crude death rates
 1970 are much higher than modern norms. In a good year the crude death rate in 19th
 century Boston was 20, double the modern level. In a year with an epidemic it could
 reach 40 per thousand, or about one in 250. Philadelphia and New York were similar,
 although the crude death rate in New York was a little higher. New Orleans was even
 more dangerous. In a good year, the crude death rate in New Orleans was close to
 1975 40, a bad year in Boston. The worst epidemic in 19th century New Orleans killed 140
 of 1000 people, about one person in seven.

Figure 4.6: Infant mortality rates in the US and Massachusetts



Note: *Left panel reports infant deaths per thousand births from 1850-2000. Solid line reports on the whole US, dashed line is Massachusetts. Infant mortality in the US and Massachusetts in the 19th century was terrifyingly high. Current US rates are about 5 per 1000. Right panel reports infant deaths per thousand births for rural (black line) and urban (gray line) parts of Massachusetts between 1880 and 1915. Urban infant mortality is dramatically higher than rural, mortality rates fall in both regions, and mortality rates fall faster in urban than in rural parts of Massachusetts.* Figures reproduced from. Alsan and Goldin [2019], ©University of Chicago Press.

Second, something happened in all four of these cities towards the end of the 19th century, the crude death rate began to decline and the frequency of epidemics decreased. By the early 20th century, the crude death rate in Boston and Philadelphia was comfortably under 20, and close to 25 in New Orleans. In New York, the level declined less dramatically, although the frequency of epidemics clearly decreased.

It is natural to wonder whether late 19th century reduction in the crude death rate was something special to cities, or if it just reflected a national trend. This was a time of rapid economic growth, and we should not be surprised if some of the innovation that expanded peoples' consumption choices also made them healthier. Figure 4.6

begins to answer this question. The left panel of this figure plots infant mortality rates, the number of babies born who do not survive until their first birthday, from the whole US (solid) and Massachusetts alone (dashed) from 1850 until 2000. The US data shows a rapid and continual decline from about 205 per thousand to the modern value of around 5 per thousand. That is, from about one in five, to about one in 200. The Massachusetts data is similar for most of the period, although rates in Massachusetts seem to have been lower in the early part of the sample. This suggests that at least some of the decline in crude death rate that we see for cities is part of a national trend, and not specific to cities.

The right panel of figure 4.6 gives us a little more detail. This figure shows the infant mortality rates in Massachusetts for rural regions (black) and urban regions (gray) between 1880 and 1915. Both rural and urban infant mortality rates fall over this period, but urban rates fall more rapidly. Although this graph covers a much shorter period than we have looked at so far, it suggests that the decrease in urban mortality is partly a reflection of a national trend, but also partly a reflection of a relative improvement of the urban disease environment.

Data reported in Haines [2001] makes the relative improvement in cities precise. Using data for the whole US, Haines [2001] finds that the urban mortality premium declines from about 40% in 1870 to about 20% in 1920. That is, the crude death rate in an average US city is about 40% higher than for an average rural region in 1870, but only about 20% higher by 1920. By the end of the 20th century, urban and rural crude mortality rates are the same for both regions. By 2020 the urban mortality premium actually reverses and the crude death rate is 20% higher in rural than urban regions [Thomas et al., 2024].

2010 4.4 The monocentric city model and urbanization in the developed world

Urbanization accelerated with the beginning of the industrial revolution. This was accompanied by dramatic increases in income and helped along by increases in agricultural productivity that were probably not quite as large.

2015 Modern cities are more productive as they are bigger and denser, and this was almost surely true of cities early in the industrial revolution. The larger cities that emerged during the industrial revolution were probably both a cause and a consequence of improving technology.

2020 On the other hand, developed world cities were unhealthy places in the 19th century, and they were more dangerous than the countryside. Moving from the countryside into a mid-19th century city meant a significant increase in the risk of illness and death. Urban mortality, and the gap between rural and urban mortality, both began to decrease in the late 19th century and early 20th century. By the end of the 20th century, the difference disappears and begins to reverse in favor of cities.

2025 These facts suggest that we think of urbanization in the developed world as reflecting the trade-off between income and illness that came with life in larger 19th century cities.

2030 As a way to work through this idea more carefully, let's see whether we can explain these phenomena with the monocentric city model. We introduced the idea of a city specific or urban amenity in Chapter 1. As when we thought about how Covid affected the land market, we can use urban amenities to think about the dangerousness of the city. However, to think about the process of migrating to and building cities, what

matters is not the absolute level of disease in the city, but whether cities are better or worse than the countryside. If the monocentric city model is going to have anything 2035 to say about urbanization in the 19th century US, we need to extend it so that it, somehow, also describes the level of disease in the countryside.

With this in mind, let A_R , c_R and w_R denote a rural amenity level, consumption level, and wage, respectively. Assume that rural households work where they live, and to make things easy, assume that land rent in the countryside is zero, that is, 2040 $\bar{R} = 0$. Rural households have nothing to spend money on except the consumption good, and so we have $c_R = w_R$.

Next, define the outside option as the level of rural utility. That is, $\bar{u} = u(A_R c_R)$. Thus, the reservation utility level for city dwellers is determined by rural income and amenities. This change aside, everything is the same as the monocentric city model 2045 with amenities developed in Chapter 1, except that to prevent confusion, we will denote the urban amenity level by A_U instead of just A as we did earlier.

Each urban household chooses their location, commutes to work and divides w between commuting and consumption, c . This means that a household's problem is

$$\begin{aligned} & \max_{c,x} u(A_U c) \\ \text{s.t. } & w = c + R(x)\bar{\ell} + 2t|x|. \end{aligned}$$

If no one wants to move, it must be that utility in the city and the countryside is 2050 the same. That is, $u(A_U c^*) = u(A_R c_R) = \bar{u}$. Solving this for c^* , we have

$$c^*(A_U) = u^{-1}(\bar{u})/A_U. \quad (4.1)$$

That is, everyone in the city must get just enough consumption that they do not want to leave for the countryside. It follows that for all locations x in the city, we must have

$$w - c^*(A_U) = R(x)\bar{\ell} + 2t|x|.$$

Let \bar{x} denote the most remote urban location to the right of zero. At this location,
2055 we must have

$$w - c^*(A_U) = \bar{R}\bar{\ell} + 2t\bar{x}.$$

Reorganizing, and using $\bar{R} = 0$, we have

$$\bar{x} = \frac{w - c^*(A_U)}{2t}.$$

Because the city extends from $-\bar{x}$ to \bar{x} and each household consumes an exogenously fixed amount of land $\bar{\ell}$, the total population of the city is

$$N^* = \frac{2\bar{x}}{\bar{\ell}} \tag{4.2}$$

$$= \frac{2}{\bar{\ell}} \left[\frac{w - c^*(A_U)}{2t} \right] \tag{4.3}$$

$$= \frac{w - c^*(A_U)}{t\bar{\ell}}. \tag{4.4}$$

So N^* is increasing in the wage, and *decreasing* in c^* . Why would population decrease
2060 as reservation consumption increases? All else equal, an increase in c^* means that the reservation utility level has gone up, this makes it harder for the city to attract

residents. In equilibrium, if consumption in the city is higher, commutes must be shorter, and so the city is smaller.

Substituting equation (4.1) into equation (4.2), we can write the size of the city
2065 as a function of urban and rural amenity levels,

$$\begin{aligned} c^*(A_U) &= \frac{u^{-1}(\bar{u})}{A_U} \\ &= \frac{u^{-1}(u(A_R c_R))}{A_U}. \end{aligned}$$

By inspection, c^* is decreasing in A_U . As urban amenities improve, urban dwellers match rural utility levels with less consumption.

Differentiating the expression for $c^*(A_U)$ with respect to rural amenities and rural consumption, we get (using the chain rule),

$$\frac{d}{dA_R} c^*(A_U) = \frac{1}{A_U} (u^{-1})' u' c_R$$

2070 and

$$\frac{d}{dc_R} c^*(A_U) = \frac{1}{A_U} (u^{-1})' u' A_R$$

With $u' > 0$ by assumption, we must have $(u^{-1})' > 0$, so c^* is increasing in rural income and amenities. That is, cities must offer their residents higher consumption when rural amenities or consumption improves. Using the result above, this means that city sizes decline as rural income and amenities improve.

2075 A similar argument shows that city population, N^* , is increasing in the urban wage and amenity level, w and A_U , and decreasing in rural consumption and amenities,

c_R and A_R . In equilibrium, as urban income increases, cities get larger and, as deaths decrease the urban amenity increases and city size increases. This is broadly consistent with the data presented above. Once again, the monocentric city model
2080 seems to make qualitatively correct predictions.

With this said, we've omitted a few things. Transportation costs fell dramatically over the course of the 19th century. This period saw the arrival of the horsecar (horse drawn trolley cars on rails), commuter railroads, and in a few cities, subways.² We worked out the comparative statics of falling transportation costs in Chapter 1, and
2085 they do not change with the addition of amenities. Falling transportation costs should lead to larger cities. This must surely be part of the story of US urbanization in the 19th century, although little systematic evidence is available.

It would also be interesting to know the relative importance of the urban wage premium and the mortality premium. Which was more important? Do cities grow
2090 faster when we increase their productivity, or when we make them less dangerous? To my knowledge, there is no systematic evidence on this question.

Notice that we also cut a few corners. First, we are using a model of a single city to analyze the migration of people all over the continent into many cities. Thus, the problem we've analyzed does not quite match the one we set out to solve. In particular, by restricting attention to a single city, we preclude considering the possibility
2095 that cities are different from each other, even though we see that even cities of about the same size have different levels output per person in the left panel of figure 4.4. We return to this issue when we talk about systems of cities in Chapter 9.

²The first subway seems to have opened in Hamburg in 1865, followed by Munich in 1871. Subways in the following other cities opened (in order) prior to 1900; Boston, Frankfurt, London, Berlin, New York, Chicago, Budapest, Glasgow, and Oslo. From [Gonzalez-Navarro and Turner, 2018].

Second, the evidence in support of our story, that the development of US cities
2100 reflected the tradeoff between higher urban productivity and mortality, is pretty in-
formal. The basic features of the data are all correct; productivity is up, disease is
down, urban share is up, but it is not clear if the details are right. Does the timing
of the reductions in disease match the timing of growth of cities exactly as it should?
Healthy people are more productive, so maybe much of the increase in productivity
2105 is not because more people moved to cities, but because cities became healthier? At
the time of this writing, we do not have good answers to these questions. The story
I've told in this section is a theory, and it may be rejected as more data becomes
available.

Finally, our description of the urbanization process does not recognize the possi-
2110 bility that city size was capped by the productivity of farmers. If the work of nine
farm households can feed ten, then the availability of agricultural surplus caps the
urban population at 10% of the population. The urban share of population in Europe
in 1500 and 1750 was about 10.7 and 12.2%, not much different than the US in 1800,
and in 1500 there were only 154 cities in all of Europe with populations above 10,000
2115 [De Vries, 2013].

It is natural to suspect that this persistently low urban share partly reflects the
available agricultural surplus. There is good support for this story. The introduction
of the potato to Europe dramatically increased agricultural productivity in places
where the soil was suitable for its cultivation. Nunn and Qian [2011] shows that such
2120 areas saw an increase in urban share after the potato was introduced. By assuming
that consumption is available at a constant price, no matter the quantity demanded,
our specification of the monocentric city model rules out the possibility of an aggregate

Table 4.1: Urbanization rates around the world

	% Urban 2018	Urbanization rate %/year, 2010-15
South Asia	35.8	1.2
Sub-Saharan Africa	41.5	1.4
Southeast Asia	48.9	1.3
Latin America and the Caribbean	80.7	0.3
Europe	74.5	0.25
North America	82.2	0.21

Note: *Latin America and the Caribbean, Europe and North America are all highly urbanized and the urban share is stable. South Asia, Southeast Asia, and Sub-Saharan Africa are less than half urbanized and the urban share is growing rapidly. This is where the world is building cities.* Table reproduced from [Henderson and Turner, 2020]

constraint on consumption. This is probably appropriate for the late 19th and early 20th century US when agricultural productivity had risen to well above pre-industrial levels, but perhaps not in earlier periods.

4.5 Urbanization in the developing world

Basic facts about urbanization in the 19th century developed world support the hypothesis that the process of urban migration and construction was governed by the tradeoff between higher urban productivity and disease. This hypothesis is consistent with the logic of spatial equilibrium in the monocentric city model.

Urbanization in the modern *developing* world seems different. On the one hand, the urban wage premium is probably at least as large as in modern or historical developed countries. On the other hand, developing world cities are not obviously as dangerous as were early developed world cities. Moreover, other “amenities” in developing world cities look better than in the countryside. Taken together, these

Table 4.2: Income at urban share

	Year > 40% Urban	1990 GDP/person
East Asia	2010	3537
Sub-Saharan Africa	2018	1481
Latin America and the Caribbean	1950	2500
US	1900	6250

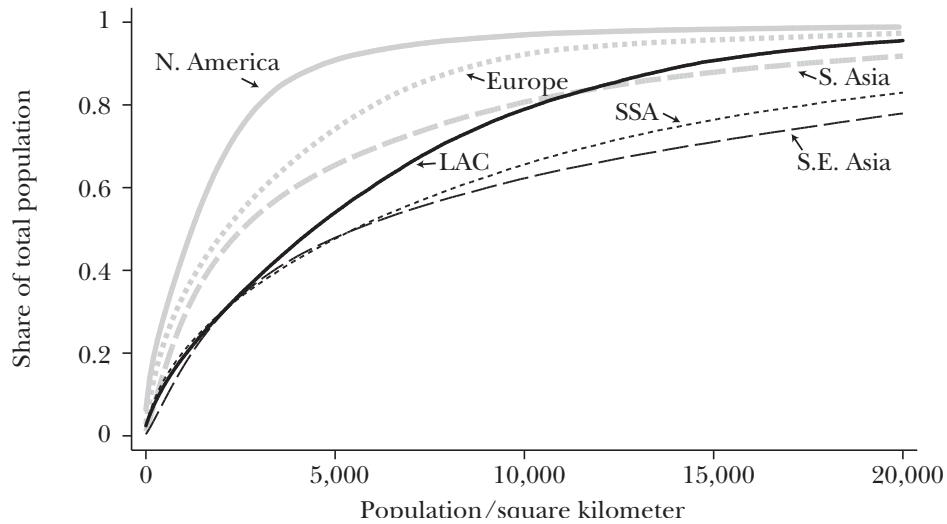
Note: *Latin America and the Caribbean, Sub-Saharan Africa, East Asia began building cities at much lower levels of income than that of the US when it was at the same urban share. This makes it harder to pay for infrastructure and state capacity to make the cities work.* Table reproduced from [Henderson and Turner, 2020].

facts make it hard to understand why everyone in the developing world doesn't live in a city. Wages are higher, cities are not obviously less healthy, and, many other things look better in cities, too.

For spatial equilibrium and the monocentric city model to explain why people stay in the countryside, we need some cost of urban migration or residence (what Edward Glaeser picturesquely calls the “demons of density”) to offset the obvious benefits of cities.²¹⁴⁰

We can eliminate some candidate urban costs. It is probably not high urban unemployment, mortality, or that rural residents don't know about urban opportunities.²¹⁴⁵ Other urban costs look more important, like the cost of lost rural social networks, exposure to higher urban crime rates, or an increase in less obvious health risks. That is, it looks like there are enough problems with developing world cities to plausibly that they offset urban benefits for a marginal migrant (but the case is not a slam dunk). That is, the evidence does not let us reject the hypothesis that same logic of spatial equilibrium explains urbanization in the developed world and the developing world. This section lays out this case, and shows how the monocentric city model²¹⁵⁰

Figure 4.7: Cumulative share of population by population density



Note: *x-axis is population density per square kilometer. y-axis gives the share of people living at or below any given population density. Thus, in SSA, about 60% of people live at population densities below 10,000 per square kilometer. We see a clear and somewhat surprising pattern. In North America, almost 90% of people live at densities below 5,000 per square kilometer. In Sub-Saharan Africa and South East Asia, this share is below one half. Thus, in spite of the fact that these places are less urban than North America, a greater share of their people live at high population densities.* Figure reproduced from [Henderson and Turner, 2020], ©American Economic Association.

can apply to urbanization in the developing world, too.

Table 4.1 reports the urban share of population and its growth rate, by region, in the early part of the 21st century. As of 2018, Latin America and the Caribbean (LAC), Europe and North America have urban population shares around 80%. In Asia and Africa, this number is between 35% and 50%. During the period 2010-15, the annual growth rate of the urban population was about 0.3% in Latin America, Europe and North America, and about four times as high in Africa and Asia. A much lower share of the population of Asia and Africa lives in cities than in the rest of the

²¹⁶⁰ world, and this is where the world is building cities most rapidly.

While developing world cities are growing more rapidly than developed world cities now, developed world cities probably grew as fast when they were younger. During the 20 years between 1880 and 1900, the mean urban share across a sample of developed countries increased by about 7%. Between 1985 and 2005, the corresponding statistic ²¹⁶⁵ for the developing world is about the same [Scott, 2009, Ch. 4].

As in the developed world, people in developing world cities earn more money than people in the countryside. The top panel of figure 4.8 plots the ratio of urban to rural per capita GDP by country, as a function of country per capita GDP. In almost all countries, people in cities have higher incomes than in the countryside, usually at ²¹⁷⁰ least half again as much, but sometimes by two or three times as much. Beyond this, the urban income premium seems to be declining in country income.³

The bottom panel of figure 4.8 suggests that the story may be a little more subtle than the top panel suggests. This figure presents the ratio of urban to rural *consumption* per capita rather than income per capita, again as a function of country ²¹⁷⁵ per capita GDP. Note that the scale on the *x*-axis is different than the top panel; the bottom panel focuses attention on poorer countries. This figure also shows an urban premium, people in cities consume about twice as much as in the countryside, but the negative relationship between the urban premium and country per capita income is less clear. The difference between consumption and income is made up by own ²¹⁸⁰ production. Subsistence farmers in the countryside may not have any money, but

³Note that this second conclusion is not obviously consistent with the about constant urban wage premium we saw for the 19th and 20th century US in figure 4.3. US GDP per capita increased dramatically over this period, so if the urban wage premium declines with country income, it should have declined over time in the US, too. This is a puzzle with no definitive resolution. I suspect that the US data presented in figure 4.3 is simply so noisy we can't see the trend that shows up so clearly in figure 4.8.

they can eat what they grow, so the rural-urban gap in welfare is probably not as large as the urban wage premium suggests.

Henderson and Turner [2020] refines our understanding of the urban wage premium by looking at individual level data describing urban and rural outcomes in the developing world. They rely on two main types of data. First, they use gridded population data from the Global Human Settlements project for 2015. These are “best guess” estimates of population in every one square kilometer cell on a regular grid covering the whole world. Second, they use geocoded survey data describing the demographic characteristics of respondents and economic outcomes across much of the developing world, Africa in particular. Putting these two types of data together shows how outcomes vary with nearby density.

Figure 4.9 illustrates these results. For all panels density is on the x -axis. This is the count of people within 5km of a survey respondent. On the y -axis, all panels show a measure of income. The dots are a histogram and the line is a plot of a regression of

$$\ln(y) = A_0 + A_1 \ln(\text{Density}) + \varepsilon. \quad (4.5)$$

The top left panel has the logarithm of total household annual income on the y -axis. We see a clear upward trend. The elasticity of household density is positive and about 0.3. That is, a 10% increase in city size gives a 3% increase in wage. Chapter 8 will report a number of similar estimates, however, this is a really large effect, much larger than we will see for the cities in the developed world.

Why is this effect so large? Perhaps it is because the households moving to cities

are different than those who stay in the country side? The top right panel of figure 4.9 attempts to address this possibility by controlling for household demographics; the object is to compare people whose age, gender, and education are alike, as we vary their residential density. The positive household income elasticity of density persists and, surprisingly, its magnitude is practically unchanged. The increase in household income that comes from moving to denser places is probably not primarily a reflection of the sorting of productive people into cities.

The top two panels of figure 4.9 show the relationship between the logarithm of *household income* and density. The bottom panels are similar, but show the relationship between the logarithm of the *hourly wage* and density. Household income is an obvious measure of the benefit to a household from urban migration. Wage is also useful as a measure of the benefit of urban migration and also reflects labor productivity.

The bottom two panels of figure 4.9 also show a clear upward trend. In the bottom left panel we see that the elasticity of wages to density is clearly positive, and is about 0.12. Thus a 10% increase in density gives about a 1.2% increase in wages. Unlike household income, however, controlling for demographics reduces the urban wage premium dramatically. In the bottom right panel we compare similar workers and see that increasing density by 10% gives only about a 0.5% increase in wages. We will see in Chapter 8, that this number is about the same everywhere it has been measured.

We can only guess at why household incomes grow so much more rapidly than wages as density increases. Increased female labor force participation is a likely suspect. Women are probably more likely to be in the workforce in cities than in

the countryside, and this boosts household income much faster than the increase in individual wages that comes with density.

Putting figures 4.8 and 4.9 together suggests a large rural-urban income gap, possibly as large as a factor of three in the poorest countries. This gap probably overstates
2230 the gap in welfare: the poorest countries are where the rural-urban income gap looks largest and also where eating what you grow is most important. Disaggregated data shows that the increases in household income that come with higher density are large, positive, and much larger than the corresponding increases in wages.

Developing world cities, like their developed world counterparts, are more productive than rural areas. This gives people a reason to migrate to them, just as
2235 for 19th century developed world cities. We now investigate factors that could be counterbalancing this incentive.

People in the developing world are migrating to, and building cities when they are relatively poor. Table 4.2 reports the year and income at which different regions
2240 reached a 40% urban share. The US crossed this threshold in 1900 when per capita income (denominated in 2005 dollars) was 6250\$ per year. The developing world reached a 40% urban share at much lower levels of income. In Sub-Saharan Africa, on average, per capita income was less than one fourth that of the US when the US crossed the 40% urban threshold.

2245 To understand why this difference in income is important, consider the construction of sewers in Chicago. Chicago began the construction of its sewer network around 1855. Chicago is notoriously flat and sewers only operate if there is a sufficient grade to allow sewage to flow. This meant that building Chicago's sewer network involved raising and regrading Chicago streets. The scale of this project was enormous. Con-

2250 temporaneous plans state that “[i]t will be necessary to raise the grades of streets an average of eighteen inches per 2500 feet going West.” Some simple calculations let us guess that raising a 2,500 foot segment of a 20 foot-wide street by 18 inches requires 8300 cubic yards or about 12,450 tons of fill, all moved with animal power [COURY et al., 2021]. Because modern day developing world cities are poorer than was late
2255 19th century Chicago, they are less well able to afford big public works projects to tame the demons of density.

The geography of developing world cities is also different from their developed world counterparts. Figure 4.7 reports the fraction of people living at particular population densities by region. In this figure, the *x*-axis is population density per
2260 square kilometer and the *y*-axis gives the share of people living at or below any given population density. For example, in SSA, about 60% of people live at population densities below 10,000 per square kilometer.

In North America, almost 90% of people live at densities below 5,000 per square kilometer. In Sub-Saharan Africa and South and South East Asia, this share is below
2265 one half. In spite of the fact that these places are less urban than North America, a greater share of their people live at high population densities. One reason that this could occur is these developing world cities are much denser than the their North American counterparts. That Europe is nearer to North America, and Latin America (despite it's high urban share) is closer to Asia and Africa suggests that developing
2270 world cities are systematically denser than those of the developed world.

We can use the results of figures 4.7 and 4.9 to be more precise about the way income changes when people migrate to denser places. Eyeballing figure 4.7, the 20th percentile for population density in Sub-Saharan Africa is about 1000 and the 80th is

about 20,000. Using equation (4.5), and the 0.3 estimate of the elasticity of household income to density reported in the note for figure 4.9, we can calculate the change in income that results from migrating from the 20th to the 80th percentile of density, from the bottom to the top fifth, in Sub-Saharan Africa,

$$\begin{aligned}
 \ln(Y_1) &= A_0 + 0.32 \ln(20000) + \varepsilon \\
 -\ln(Y_0) &= -A_0 - 0.32 \ln(1000) - \varepsilon \\
 \ln(Y_1) - \ln(Y_0) &= 0.32(\ln(20000) - \ln(1000)) \\
 \implies \ln\left(\frac{Y_1}{Y_0}\right) &= 0.32 \ln(20) \\
 \implies \ln\left(\frac{Y_1}{Y_0}\right) &= \ln(20^{(0.32)}) \\
 \implies \ln\left(\frac{Y_1}{Y_0}\right) &= \ln(2.6) \\
 \implies \frac{Y_1}{Y_0} &= 2.6
 \end{aligned} \tag{4.6}$$

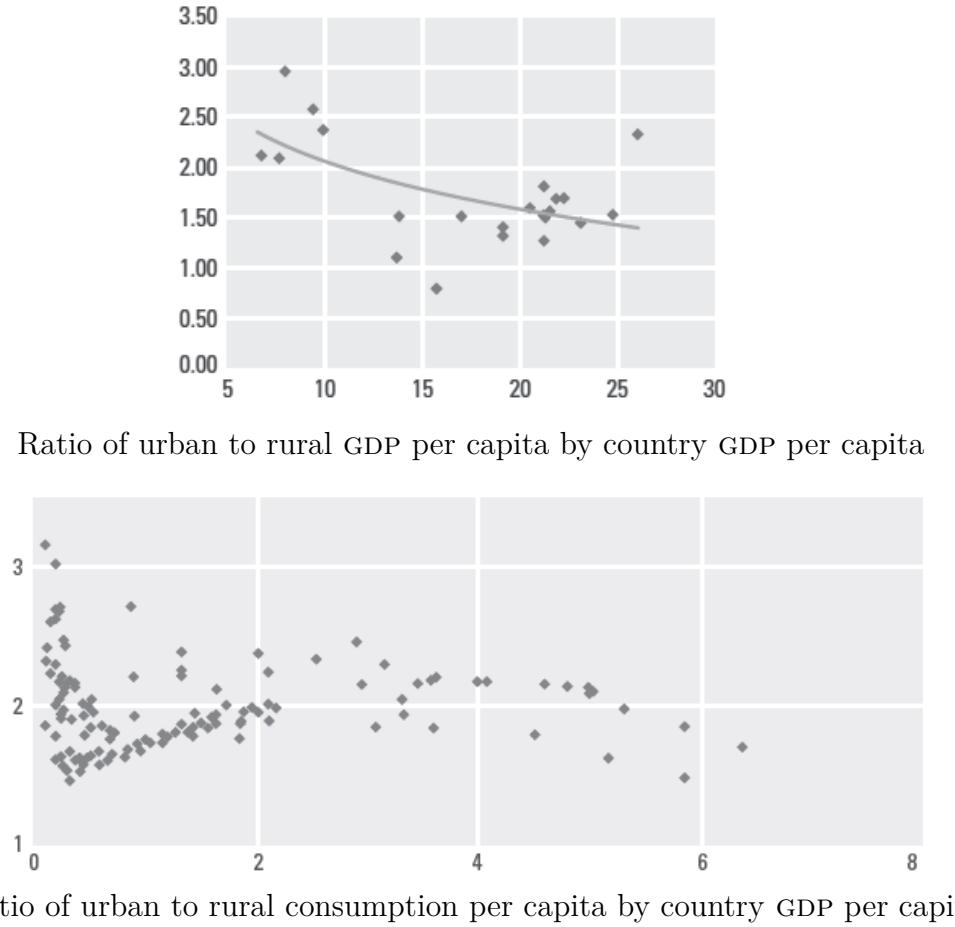
So an average household in Sub-Saharan Africa increases its income by a factor of 2.6 by moving from the 20th to the 80th percentile of density.

Not only are developing country cities denser, they are also more likely to be slums. The x -axis of the left panel of figure 4.10 is the annual growth rate of urban population and the y -axis is annual growth rate of slum population. Each dot indicates a single country, with the size of the dot reflecting the size of the country. Looking carefully, we see that the slope of the trend line is close to one. Roughly, in developing countries every new urban dweller moves to a slum.

The right panel of figure 4.10 suggests that slums may be a temporary phenomena. The x -axis of this figure reports the urban share of a country's population and the

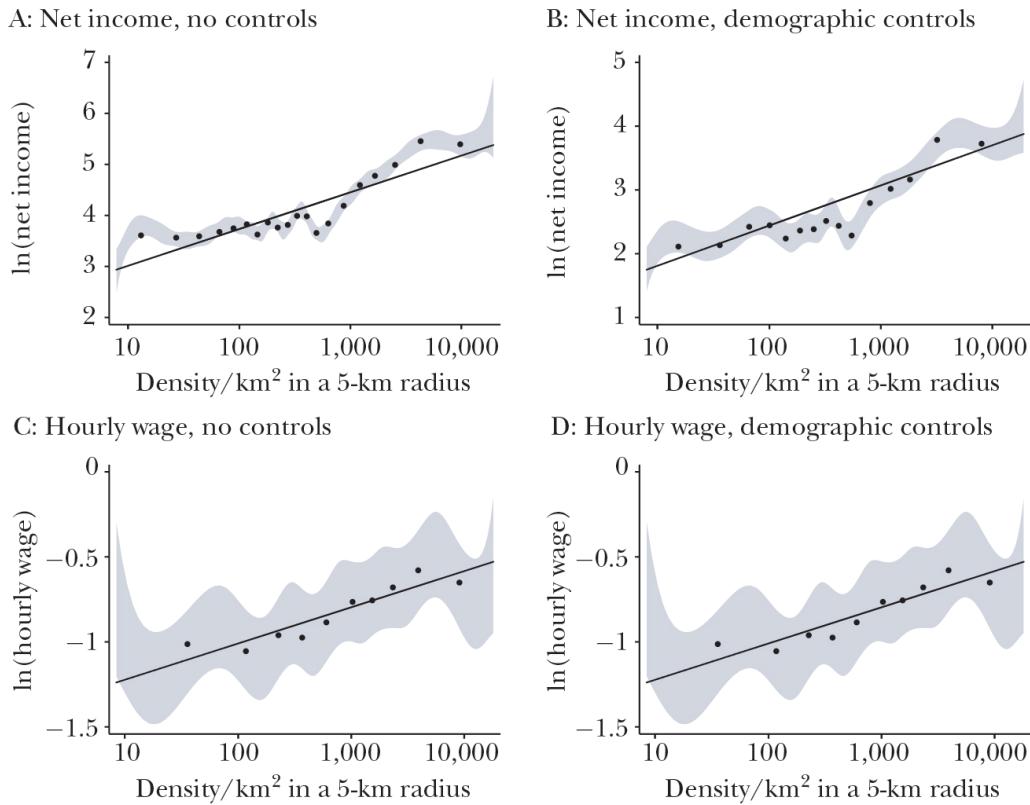
y-axis is the share in slums. As in the left panel, the size of the dots reflects country size. Here we see a strong negative slope, just the opposite of what we saw when
2290 we looked at changes in the left panel. Looking carefully, the trend line declines from about 70% to about 30% as the urban share increases from 0 to 100%, so each 10% increase in urban share gives us about a 4% decrease in the slum share. Taken together, these figures suggest that as cities grow they first house migrants in slums, and then, slowly, build better neighborhoods.

Figure 4.8: Rural versus Urban income and consumption by country income



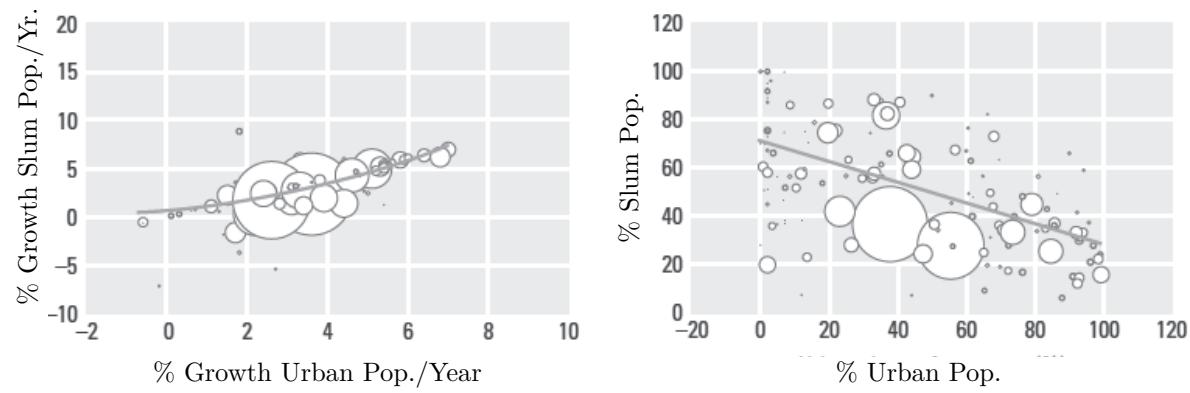
Note: Top panel shows the ratio of average urban to average rural income on the y-axis, and country income per capita (in thousands of 2009 US dollars) on the x-axis. The urban-rural income ratio shrinks with country income, from about 3 for the poorest countries to less than 1.5 for some of the richer countries. Bottom panel is the same as the top, except that the y-axis shows the ratio of the value of urban to rural per capita consumption. When we at consumption levels, the urban-rural ratio does not decrease dramatically with income, but stays about constant at 2 as income varies. Presumably, poor rural households are consuming their own production and poor urban households are not. Note that the range of incomes covered by the x-axis on the bottom panel is much less than the top panel. The bottom panel describes only the poorest countries. Figures reproduced from Scott [2009].

Figure 4.9: Household income and wages as a function of density in the developing world



Note: Plots of household income and hourly wage, against the log of population density in a 5km disk around the survey respondent's location. Left panels have no controls. Right panels include demographic controls. Countries included in the survey are Ethiopia, Ghana, Malawi, Nigeria, Tanzania and Uganda. Slope coefficients best linear fit lines are; (A) 0.313 (0.016) (B) 0.317 (0.014) (C) 0.118 (0.015) (D) 0.049 (0.009). Figure reproduced from Henderson and Turner [2020], ©American Economic Association.

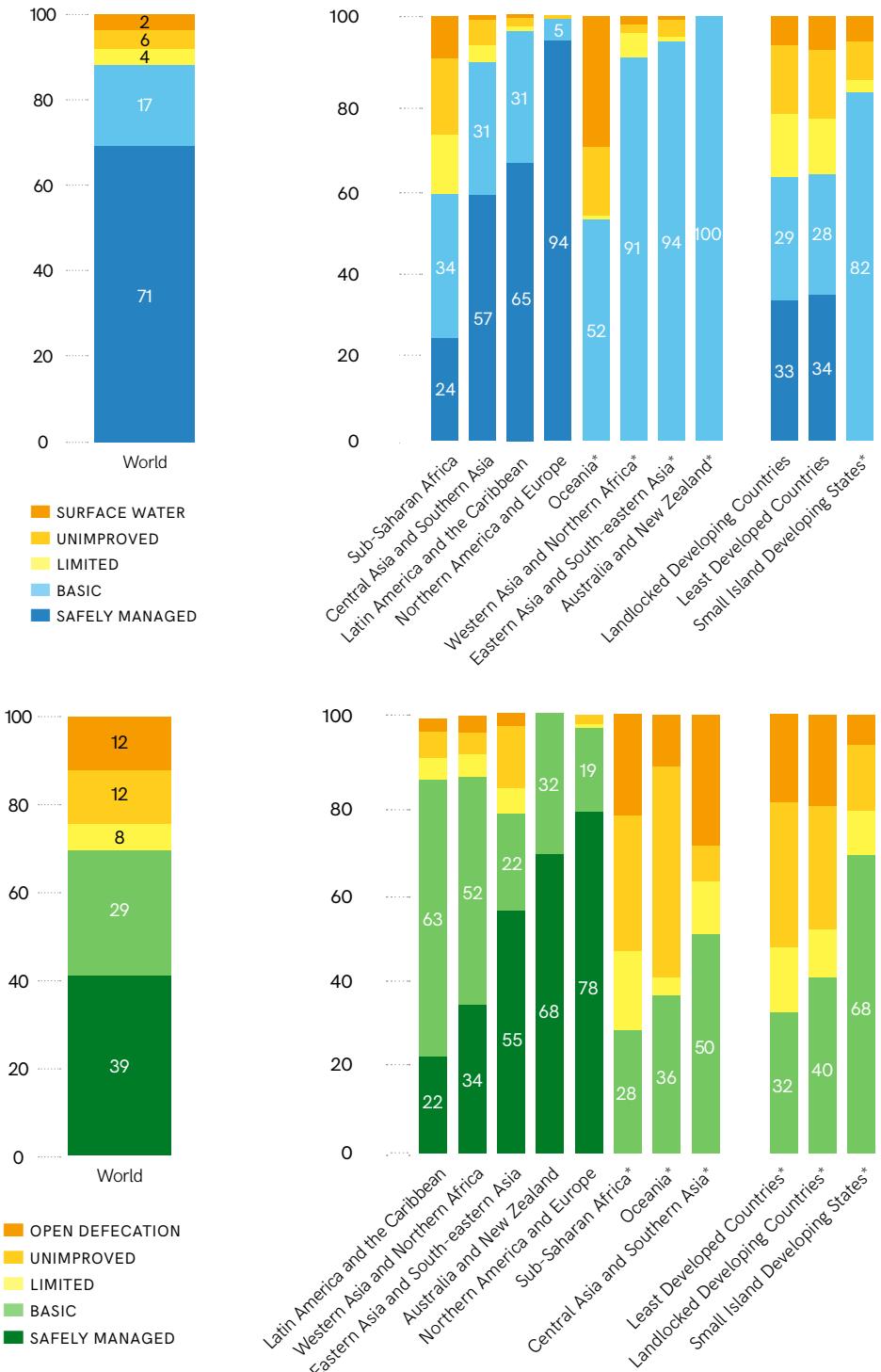
Figure 4.10: Urban share and slums



Note: *Left: When the urban population grows rapidly, the prevalence of slums increases.*

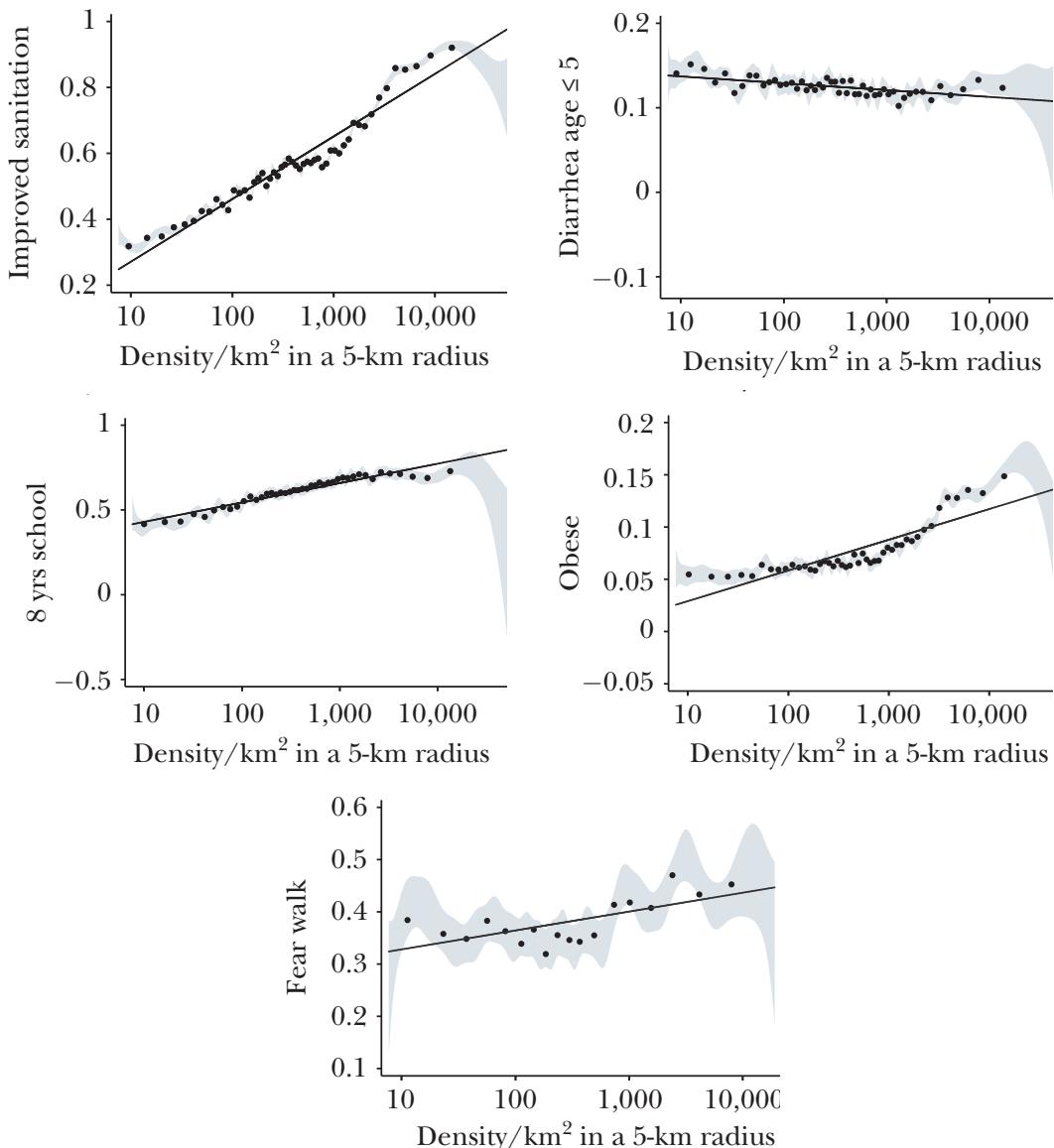
Right: As the urban share of a country's population increases, the share of urban population in slums falls. Figures from Scott [2009].

Figure 4.11: Developed world availability of water and sewer



Note: Figures reproduced from World Health Organization [2017], ©the World Health Organization.

Figure 4.12: Amenities and density in the developing world



Note: Plots of household access to improved sanitation (top left), share of mothers responding positively when asked if any of their children under age five had diarrhea in the past two weeks (top right), share of household children with 8 or more years of schooling (middle left), share of survey respondents who are obese (middle right), and share of respondent reporting that they are afraid to walk unaccompanied (bottom). In the top four panels the x-axis reports the number of people living within 5km of the survey respondent. All plots are based on the Development and Health Surveys for Angola, Bangladesh, Benin, Burkina Faso, Burundi, Cambodia, Cameroon, Chad, Colombia, Comoros, Congo Democratic Republic, Cote d'Ivoire, Dominican Republic, Ethiopia, Gabon, Ghana, Guatemala, Guinea, Haiti, Honduras, India, Kenya, Lesotho, Liberia, Malawi, Mali, Mozambique, Myanmar, Namibia, Nepal, Nigeria, Rwanda, Senegal, Sierra Leone, Tanzania, Timor-Leste, Togo, Uganda, Zambia, and Zimbabwe between 2010-6. The bottom panel is based on 26 African countries that are part of the Afrobarometer survey.

2295 Developing world cities are more likely to lack public health infrastructure that residents of developed world cities take for granted, so much so that even measuring the availability of sanitation infrastructure requires the introduction of new terminology. Abbreviating a little, the World Health Organization (WHO) defines “safely managed drinking water” as water from an improved source on the premises and free
2300 from dangerous contamination. “basic” water service is the same, but allows an up to 30 minute walk each way to get water. The definition of an “improved source” includes piped water, boreholes or tube wells, protected wells and springs, and delivered water. “limited”, “unimproved” and “surface water” involve progressively longer walks and inferior quality. By the WHO definitions, even “safely managed drinking water” is a long way from the multiplicity of taps running safe, hot and cold water
2305 that this author is accustomed to.

The definition of safely managed sanitation is similar. Again abbreviating a little, the WHO defines “safely managed” sanitation as the availability of improved facilities, not shared with other households. “Basic” sanitation involves only the use of improved facilities. As for safe water, the definition of “improved” sanitation facilities is key.
2310 Improved facilities include flush or pour toilets (a toilet that is flushed by pouring a bucket of water in the bowl rather than pushing a lever) connected to sewers or septic tanks or a pit latrine, along with ventilated pit latrines and other versions of carefully constructed outhouses. Even “safely managed sanitation” includes conditions that
2315 the developed world reserves for campgrounds. The inferior categories “limited”, “unimproved” and “open defecation”, involve the use of shared facilities, or none at all.

Figure 4.11 reports on world availability of water and sewer service in the devel-

oped world, both overall and by region. There is better availability of water than
2320 sanitation service. About 70% of the world had access to safely managed water in
2015, and this share rises to about 90% if we include basic water service. There is,
however, wide variation by region. Only one in four residents of Sub-Saharan Africa
have access to safely managed water, but it is available to about 2/3 of those in Latin
America and is almost universal in North America and Europe. Access to safely man-
2325 aged sanitation is worse. Only about 40% of the world has access to safely managed
sanitation, again with wide variation by region. Only about 1/4 of the population of
Latin America has access to safely managed sanitation, and as of the 2015 date of
these data, it safely managed sanitation was basically non-existent in Sub-Saharan
Africa.

2330 Figure 4.11 describes the availability of water and sanitation to everyone, not just
to urban residents. However, comparing the shares with water and sanitation service
from figure 4.11 to the urban shares in table 4.1, it is clear that in parts of the world,
Sub-Saharan Africa in particular, many urban residents are living under primitive
and unsanitary conditions.

2335 Putting all of this together, on the basis of aggregate data, the same basic con-
ditions seem to be present in the modern developing world as in the 19th century
developed world. There is a large income premium for moving to the cities. Devel-
oping world cities are dense. Unsanitary conditions are pervasive and probably lead
to higher rates of disease and mortality.

2340 Looking at disaggregated data, however, suggests that something more subtle is
going on. Figure 4.12 is like figure 4.9 in that it plots an outcome of interest against
population density. It is unlike figure 4.9 in that it considers a much larger set of

countries (41 versus 6) and reports on outcomes that measure quality of life, amenities, rather than wages or income.

2345 The top left panel of figure 4.12 reports on the share of population with access to improved sanitation. It shows that at low densities, access to improved sanitation is rare, but at higher densities around 5000 per square kilometer it is pervasive. Looking back at figure 4.7, we see that nearly half the population of Sub-Saharan Africa lived at densities at least this high, which suggests that much of the urban population in
2350 Sub-Saharan Africa has at least rudimentary sanitation facilities. This suggests a rosier picture than the WHO data reported figure 4.11. Henderson and Turner [2020] also find that access to safe water and electricity improves with density. If you live in a developing country and want a nice outhouse or a toilet, safe water and electricity, you should move to the densest city you can find.

2355 The top right panel of figure 4.12 complements the top left. This figure shows the share of mothers reporting that one of their children under age 5 had diarrhea in the past two weeks. That this line slopes slightly downward tells us that children are not sicker in cities than in rural areas. The increased availability of improved sanitation seems to be sufficient to keep the risk of childhood illness about constant as density increases.
2360 Children are also more likely to be vaccinated and childhood mortality rates are about constant as density increases. This is definitely not what happened in the 19th century developed world where, recalling figure 4.6, cities were much more dangerous than the countryside.

2365 The remaining three panels of figure 4.12 muddy the waters still further. The middle left panel reports on the share of children who receive at least eight years of schooling. This is unambiguously increasing with density, children are more likely to

receive at least a primary school education as they live in more urban environments. The middle right panel reports on the share of the population that is obese. Obesity is worse in more urban environments, although it is less clear if people who might face
2370 hunger in the countryside would regard this as a problem. High blood pressure and diabetes are also more common in denser places, though asthma is not. The bottom panel reports on the share of respondents reporting that they feel fear when walking on the streets around their homes. This share increases with population density. On average, people are more afraid in denser cities than in the countryside.

2375 In all, the relationship between quality of life in the city and the countryside in developing countries looks like a complicated one. Cities are crowded. Sanitation is more available in cities, but is often rudimentary. Urban wages and income are much higher, the health of children is probably about the same, access to education is better, and crime is probably worse.⁴

2380 Where the decision to migrate from the country to the city in the 19th century US required a trade-off between an obviously higher wage and a 40% higher crude death rate, the same decision in a modern day developing country looks much more nuanced. Does this fit with our story for the developing world? Maybe. Clearly, life in developing world cities involves costs not present in the countryside. While these
2385 costs could be large enough to offset a doubling or tripling of income, they could also be smaller. In this case, how might we explain why not everyone moves to cities to double or triple their income?

⁴Curiously, attitudes towards women also vary with density. One question on the Development and Health Survey asks women “Is wife beating justified for any reason?”, and the frequency of positive responses declines with density, as does the total reported births per woman. Complementary to this, the use of contraception increases with population density, as does the risk of spousal abuse, albeit slightly.

There are three candidate explanations. As ever, maybe the idea of spatial equilibrium is just wrong, and like Ptolemy plotting progressively more complicated paths for earth orbiting planets, we are waiting for some Copernicus to explain how things work. Second, maybe the problems we see in developing world cities are worse than they look, and they actually about offset the income gains that come with density. Finally, people in the developing world are really attached to their rural homes, or in jargon, rural amenities are more important than we've considered. The explanation based on rural amenities is the current favorite of researchers. In the next section, we use the monocentric city model to make a guess at how important rural amenities must be in order make sense of what we have observed so far.

4.6 The Monocentric city model and rural amenities

Recall how we set up the monocentric city model with rural amenities in section 4.4, A_R and A_U measure rural and urban amenities, and c_R and c_U are rural and urban consumption. Reservation utility for urban residents is $\bar{u} = u(A_R c_R)$, so the reservation utility level for urban residents is determined by rural income and amenities. Each urban household chooses their location, commutes to work and divides wages between commuting and consumption. This means that a household's problem is

$$\begin{aligned} & \max_{c,x} u(A_U c) \\ \text{s.t. } & w = c + R(x)\bar{\ell} + 2t|x| \end{aligned}$$

Everyone, urban and rural, should get the same utility in equilibrium so that no one wants to move. Thus,

$$\bar{u} = u(A_U c_U) = u(A_R c_R).$$

Rearranging gives,

$$c^*(A_U) = u^{-1}(\bar{u})/A_U. \quad (4.7)$$

To make things easy, we maintain our earlier assumption rural rent is zero and
²⁴¹⁰ that farmers don't commute. This means that

$$c_R = w_R. \quad (4.8)$$

Finally, taking the estimate of the urban wage premium that we made in equation (4.6) seriously, suppose that

$$w_U = 2.6w_R. \quad (4.9)$$

For urban residents, their budget must hold, and they must obtain the same level of consumption at all locations in the city. This requires that,

$$c_U = w_U - R(x)\bar{\ell} - 2tx. \quad (4.10)$$

²⁴¹⁵ At the edge of the city, this becomes,

$$c_U = w_U - 2t\bar{x}. \quad (4.11)$$

Substituting equations (4.8) and (4.11) into equation (4.7), we get

$$A_R w_R = A_U [w_U - 2t\bar{x}].$$

Finally, using equation (4.9) and rearranging, we get

$$A_R = \left(1 - \frac{2t\bar{x}}{w_U}\right) 2.6 A_U.$$

Let's call $\frac{2t\bar{x}}{w_U} = \alpha$. This is the fraction of total household resources devoted to commuting. Redding and Turner [2014] survey roundtrip commute times for several developed countries and find that the largest value is less than an hour. Doubling this to two hours a day, we have that two hours of commuting costs about 20% of a 10 hour work day in lost wages. If we take this as an upper bound on the cost of commuting, we have $\alpha < 0.2$ and

$$\begin{aligned} A_R &= (1 - \alpha) 2.6 A_U \\ &\gtrsim 2.0 \times A_U \end{aligned}$$

That is, in order for us to have a spatial equilibrium where rural residents don't want to move to the city, and where doing so increases income by a factor of 2.6, we need rural amenities to be a lot larger than urban amenities.

What could be going on? There are three main hypotheses. First, that unemployment is high in developed world cities. Second, that people don't know about the opportunities available in the city. Third, that people value the social networks where they live too highly to leave them behind. Let's look at the case for each.

The first candidate is that urban unemployment is high, and in particular it is higher than rural unemployment, an idea that was proposed in Harris and Todaro [1970]. If this is right, then we are mismeasuring the wage that people use to make their location decisions. We are observing the wage conditional on being employed,
²⁴³⁵ but people decide where to live on the basis of their expected income, factoring in the possibility that they could end up unemployed.

A little more formally, if people who move to the city are unemployed with probability p , but they are never unemployed in the countryside, then we could observe $w_U > w_R$. But because we are not observing wages for the unemployed, this would
²⁴⁴⁰ not be the wage people were using to make migration decisions. Rather, people would be comparing the expected urban wage to the rural wage. That is, spatial equilibrium would be based on a comparison of pw_U and w_R , not w_R and w_U .

But household income is not subject to this reporting problem. Household surveys report incomes for all households, not just those where people are employed. We saw
²⁴⁴⁵ in figure 4.9 that household incomes go up with urbanization, even after allowing for the possibility some households face unemployment. This suggests that urban unemployment is not playing an important role as a deterrent to urban migration.

The second possibility is that rural residents simply don't know that better jobs are available in the cities. They would move if they knew, but they don't. In this
²⁴⁵⁰ case, if we could just teach them about these opportunities, then we could shift a lot of people out of rural poverty into the more productive urban economy. Bryan et al. [2014] do an important experiment to assess this hypothesis.

Parts of rural Bangladesh are subject to regular famines each year during the months leading up to the harvest. These are poor places and households. In 2010

²⁴⁵⁵ an average four person household in the study sample has an income of about 60 dollars per month. Seasonal employment is available in the cities, and it is possible to migrate to the city during the famine season, work, and send money home. On average, these remittances allow recipient households to increase their daily calories from about 2000 to about 2800 per person per day.

²⁴⁶⁰ To learn about why more households don't send a family member to work in the city, Bryan et al. [2014] selected households from rural villages at random to receive one of three treatments,

1. A cash transfer to the household conditional on a household member migrating to the city to work during the famine season. The transfer was about 8.50USD, about equal to the cost of a round trip bus ticket and a little less than a week of income for an average household.

2. A loan of the same amount, to be repaid at the end of the famine season, again conditional on a family member migrating to the city. This is the same as the first treatment, but the money is a loan rather than a gift.

- ²⁴⁶⁵ 3. Information about the opportunities available in the city.

There is also a randomly selected group of control households, whose behavior is surveyed in the same way as the treated households, but who do not otherwise interact with the researchers.

If households are randomly assigned to the different treatment and control groups, ²⁴⁷⁵ then comparing outcomes for treated and control households gives us the effect of the treatment, with no econometrics required. Bryan et al. [2014] explicitly randomize households into the different treatment groups.

Table 4.3: Migration rates by treatment and year in bus ticket experiment

	Incentivized	Not Incentivized
Migration rate in 2008	58.0%	36.0%
Migration rate in 2009	46.7%	37.5%
Migration rate in 2011	39%	32%

Note: *Table based on results in Bryan et al. [2014].*

Bryan et al. treated villagers in 2008, and then returned to survey treated and control households about their consumption and migration behavior in 2008, 2009,
2480 and 2011.

The study found that people who received the gift and the loan behaved about the same way, and that people who received the information only treatment and the control behaved in about the same way. This means we can understand the main results of the study by just looking at the behavior of these two groups,
2485 “Incentivized” and “Not Incentivized”.

Table 4.3 presents these results. 36% of non-incentivized households send someone to work in the city during the famine season. On the other hand, 58% of incentivized households do so. The gift of a bus ticket, or a loan to purchase one, increases the share of households with a migrant worker by 22%. In the year after the intervention,
2490 the share of incentivized households with a migrant worker is 47%, versus 38% in non-incentivized households, even though there is no incentive offered this year. Finally, in the final year of the study, the share of incentivized households with a migrant worker is 39% versus 32% for non-incentivized households. Again, no incentive in this year.

2495 This experiment shows a really big effect on migration behavior from a subsidy

the size of a bus ticket. But, the effect wears off after two years. After they learn about the urban labor market, most treated households revert to their pre-treatment behavior. In the last year of the experiment, only 7% of treated households were making a different decision than we would expect if they had not been treated. It
2500 looks like most households are making decisions they are happy with, even if the experiment pushes them away from this decision for a few years. This does not suggest that the rural poor are failing to move to the more productive cities because they don't understand the opportunities that are available there.

It is tempting to interpret this experiment as suggesting a possible path toward alleviating rural poverty. If we could scale this project and hand out free bus tickets to everyone, perhaps we could make a dent in the number of hungry people during famine season? These results do not permit this conclusion. Everything depends on the nature of labor demand in the cities. If urban labor demand is perfectly inelastic (we know very little about how the elasticity of urban labor demand in developing countries, so this is possible), that is, it does not change as the supply of workers changes, then each hungry holder of a subsidized bus ticket takes a job from a worker who migrated to the city unsubsidized. In this case, the experimental treatments have no impact on aggregate poverty, they just shuffle the identities of the fixed number of people who can migrate from the rural regions to find famine season jobs.
2510

The final explanation we'll consider is that people want to stay in the countryside because rural social networks are too important to leave behind. The rural population in developing countries is often poor. We just saw that many of them decided to migrate to the city during the famine season in response to a subsidy that would about cover lunch in the developed world. Perhaps for such very poor people, their
2515

Table 4.4: Percent of loans by source and purpose in India

<i>Purpose:</i>	<i>Investment</i>	<i>Operating expenses</i>	<i>Contingencies</i>	<i>Consumption expenses</i>	<i>All</i>
<i>Source:</i>					
Bank	64.11	80.80	27.58	25.12	64.61
Caste	16.97	6.07	42.65	23.12	13.87
Friends	2.11	11.29	2.31	4.33	7.84
Employer	5.08	0.49	21.15	15.22	5.62
Moneylender	11.64	1.27	5.05	31.85	7.85
Other	0.02	0.07	1.27	0.37	0.22
Total	100.00	100.00	100.00	100.00	100.00

Note: *Investment includes land, house, business, etc. Operating expenses are for agricultural production. Contingencies include marriage, illness, and others. Table reproduced from Munshi [2014].*

- ²⁵²⁰ social networks are important enough to offset the benefits of a higher urban wage, far from people who can help them get a loan (often for consumption) or find a job.

Table 4.4 provides some evidence about the importance of social networks in India. The rightmost column is probably most interesting. It shows that about 21% of loans, those from one's Caste and friends, are clearly related to one's social network.

- ²⁵²⁵ The opportunity to take such loans could easily be lost by moving away. Another 14% of loans come from an employer, a money lender, or "other". The interpersonal connections required to make such loans could also be difficult to make in a new place. However, fully two thirds of loans come from a bank, and it is hard to imagine that all access to bank loans would be permanently lost if one moved to a new community.

- ²⁵³⁰ Even if we restrict attention to the third and fourth columns, loans from one's caste for contingencies and consumption expenses are more important, but even here, banks continue to provide a substantial share of loans.

This table supports the claim that community social networks can play an important role in the lives of the developing country poor. The possibility that the
2535 loss of these social networks is an important impediment to urban migration seems plausible. However, the case seems far from certain. We do not have a good sense for whether people are able to reestablish social networks in the city, or if they can easily find a market source of emergency loans, e.g., a pawnshop. It may also be that the increase in urban income renders these social networks less important. Doubling
2540 or tripling one's income likely reduces the need to borrow grocery money.

Summing up, the large rural-urban wage gap in developing countries is a challenge to the idea of spatial equilibrium. How can people be indifferent between locations when the income difference is so great?

There was a large rural-urban wage gap in developed countries, too, but not as
2545 large, probably about 30% in the 19th century US. There was also a large urban mortality premium, as high 40%. It is not hard to imagine people being indifferent between a 30% wage increase and a 40% increase in mortality risk. Developed world cities were productive but the demons of density were obviously dangerous when these countries were urbanizing. The situation is less clear in modern developing
2550 countries. Looking at how amenities vary with density in developing countries is a mixed bag. Some things, like access to improved sanitation and piped water, and primary education, are better. Others are worse. Children are a little sicker though they don't die at a much different rate, and crime is more salient. On the other hand, the wage and income gap looks much larger. Income in developing world cities
2555 is often larger than in the country by a factor of two or three. Are the demons of density scary enough to offset this wage increase? The location decision faced by the

modern resident of the developing world countryside looks more complicated than the one faced by his or her 19th century US counterpart.

But if the residents of the developing world are not simply trading off higher incomes against worse and more dangerous living conditions, what could be stopping them from moving to the city to increase their incomes? Researchers have proposed three possible answers; that urban unemployment is too high, that rural residents are unaware of urban opportunities, and that rural social networks are too important.

The evidence that higher urban unemployment rates are important seems unconvincing. We see a rural-urban gap in household income in a sample of households where we think the rate of unemployment should be representative.

It also does not seem that rural ignorance can be the explanation. We see that most of the extremely poor villagers of rural Bangladesh revert to their pre-experiment behavior after they are induced by researchers to go to the city and experience the lucrative urban labor market.

Could it be social networks? There is evidence that social networks help poor people. Could this be important enough to explain this wage gap?

The jury is still out. But it is hard to think of a more important question to answer. If we can speed the urbanization of the rural poor, it looks like we can easily double their incomes. To put this in perspective, even in its good years, China experiences income growth of about 10%, a rate that is rarely attained by other developing countries. At a 10% rate of growth, it takes about eight years to double incomes, an increase that it looks like many households could have immediately by moving to cities.

2580 4.7 Conclusion

Most urbanization today is happening in poor countries in Asia and Africa. The rest of the world, including Latin America, is already pretty highly urbanized. There are lots of problems with developing world cities. They are being built when their residents are much poorer than were the people who built developed world cities in 2585 the 19th century. Basic infrastructure is scarce and slums are common. But for all this, developing world cities do not seem to be as dangerous as developed world cities were while they were being built. Infant mortality is not obviously higher in developing world cities than in the countryside, and this was certainly not true in the 19th century US.

2590 On the other hand, developing world cities appear to be places of tremendous opportunity. Indeed, as grim as they are, we should probably see slums as successes, not failures. They are places where people can escape the still worse poverty of the countryside. Understanding why people stick to rural places rather than moving to cities is not a solved puzzle. It is hard to point to an explanation that seems important enough. Amenities and social networks are our leading candidates, but they are hard 2595 to observe and quantify.

We also do not have a good understanding of which of the onerous features of developing world cities are most onerous, much less, which can be most cost-effectively resolved. In particular, how can we allocate scarce dollars to city building in the ways 2600 that will lead to the greatest increase in opportunity and largest reductions in squalor and misery? These questions have only recently begun to attract the attention of urban and development economists and are areas of active research.

Problems

1. This problem will use the information presented in the text to estimate urban
 2605 and rural incomes in 1820 and 2000.

(a) First, use the bottom right panel of figure 4.2 to determine real per capita
 GDP in 1820 and 2000, in constant 2011 dollars.

(b) Then, use top right panel in figure 4.2 and the right panel of figure 4.3
 2610 to find the share of the population in urban areas, and the urban wage
 premium, in both 1820 and 2000.

(c) Combine the information you have collected above to estimate urban and
 rural incomes in both 1820 and 2000 (hint: GDP per capita is a weighted
 average of wages in rural and urban areas).

2. In this problem, we will combine assumptions about urban and rural ameni-
 2615 ties with the income figures you produced above to estimate urban population
 growth between 1820 and 2000.

Denote urban and rural amenities by A_U and A_R , respectively. Let $c_R = w_R$.
 Let $u(Ac) = \ln(Ac - 1)$. Finally, let A_R/A_U be proportional to the ratio of rural
 to urban death rates (you can assume that the ratio in 1820 is the same as it is
 2620 in 1870, and that in 2000 this ratio is 1).

(a) In a spatial equilibrium, $u(A_R c_R) = u(A_U c_U)$. Assuming we are in a spatial
 equilibrium, write an expression for c_U in terms of rural wages and the ratio
 of urban to rural amenities.

- (b) Write down the household's problem for the household living at \bar{x} (you do not need to solve it).
- (c) Assuming $\bar{R} = 0$, solve the constraint in the above problem to get an expression for \bar{x} .
- (d) Recall that the city extends from $-\bar{x}$ to \bar{x} , and that each household consumes an exogenous amount of land \bar{l} . This means that the city population is given by $N^* = \frac{2\bar{x}}{\bar{l}}$.
- Write an expression for N^* based on your expression for \bar{x} .
- (e) Assume that t in 2000 is 1/2 of what it was in 1820, and that \bar{l} is constant over time. Use the information about wages from the previous problem, as well what we know about the ratio of rural to urban amenities, to write expressions for N^* in 1820 and 2000.
- (f) How much does your model imply the urban population grew between 1820 and 2000? That is, what is N_{2000}^*/N_{1820}^* ?
- (g) How does the urban population growth you computed above compare with the actual growth in the urban population over that time period? (For your reference, the US population in 1820 was 9,638,453 and in 2000 was 281,421,906 according to the Census). Why do you think there is a discrepancy?

Chapter 5

White Flight, Gentrification, and

Bid-Rent

2645

Decentralization has been an important trend in cities in the US and developed countries over the past 150 years. We have seen several pieces of evidence describing this phenomena. In figure 3.1, we see the flattening of population density gradients in US and European cities that occurred during the 19th and early 20th centuries. In figure 4.3, we see the rapid growth of the suburbs relative to the central city in the second half of the 20th century. Baum-Snow [2007] shows that the share of MSA total population in constant boundary central cities falls dramatically from 1940-1990 even as MSA population increased, and figure 2.6 illustrates how population spreads out along the interstate for a particular city.

2650

Central city populations in the US fell so far by the 1970s that central cities were shells of their former selves and the issue of the day was urban blight and decay. In 1975, New York City was so close to declaring bankruptcy that its mayor wrote, but

Figure 5.1: New York City bankruptcy announcement

October 17, 1975

STATEMENT BY MAYOR ABRAHAM D. BEAME

I have been advised by the Comptroller that the City of New York has insufficient cash on hand to meet debt obligations due today.

The financing which was to be made available by the Municipal Assistance Corporation will not be forthcoming because the Teachers Retirement Fund failed to approve its participation in the State Financing Plan.

This constitutes the default that we have struggled to avoid.

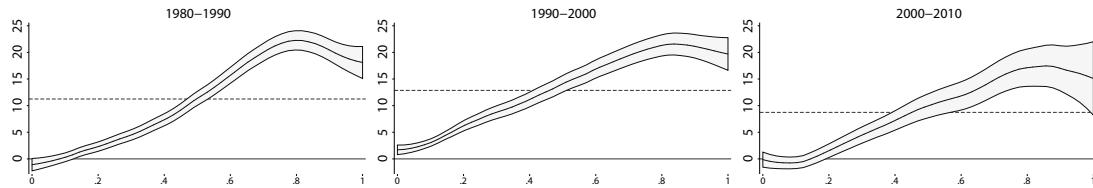
Note: *The 1970s and 80s were a pretty scary time for cities. New York City was so close to bankruptcy that its mayor prepared this speech. Image reproduced from Nussbaum [October 16, 2015].*

ultimately did not make the announcement whose first page is shown in figure 5.1.

Couture and Handbury [2020] show that the decentralization of US cities continued into the 21st century. Figure 5.2 reports their results. Unlike most of what we've discussed so far, the unit of observation in Couture and Handbury [2020] is a CBSA or "Core Based Statistical Area".¹ The y -axis of the figure shows the percentage increase in population over a decade, at each distance from the center. The x -axis is a measure of distance to the CBD, but it is exotic. The problem is that cities have different sizes, and traveling a mile from the center of New York is not the same as traveling a mile from the center of Boise. To get a standardized notion of distance, for each city they calculate the radius of circles, centered on the CBD, that contain 1%,

¹Like MSAs, CBSAs are cities built from counties, but unlike MSAs, which must describe a metropolitan area of at least 50,000, CBSAs can be as small as 10,000. All MSAs are CBSAs, but smaller CBSAs are not MSAs.

Figure 5.2: Decentralization of US CBSAs early in the 21st century



Note: *Percent change in census tract population at different distances from the city center between 1980 and 1990, 1990 and 2000, and 2000 to 2010. The dashed row in each plot shows the average population growth as a percentage for the relevant decade. The shaded region depicts the 95% confidence interval around the average. The suburbanization of population continued through 2010. Figure reproduced from Couture and Handbury [2017], ©Journal of Urban Economics.*

2%, ..., 100% of the city's population, and then use these percentages as the x -value in the figure. For example, in the left panel of figure 5.2, if we consider the donut around the center such that the inner circle contains about 50% of the population and the outer ring contains 51%, that is $x = 0.5$, we see that population increased by about 10% in this donut between 1980 and 1990. Overall, these figures show that the suburbs grew much faster than the central city between 1980 and 2010, and central city population stayed about constant or shrank.

2675 Thus, the end of the 20th century did not see the same absolute declines in central city population as occurred in the middle of the century, but the importance of central cities relative to their suburbs continued its long decline. Recalling Chapter 2, Covid appears to have accelerated the decentralization of cities around 2020, though it remains to see whether the Covid based suburban migration is transitory or permanent.

2680 While the decentralization of US cities has been going on almost since colonial

times, the process is not specific to the US. Clark [1951] shows decentralization in a small sample of European cities from 1850-1930, Baum-Snow et al. [2017] shows decentralization of Chinese cities from 1990-2010, and Garcia-López et al. [2024] shows something similar for European cities from 1961-2011.

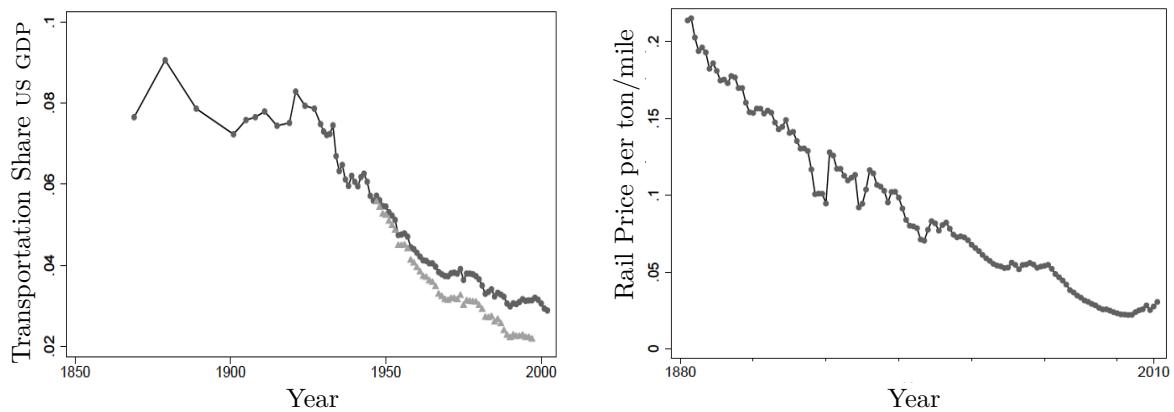
With more than 50% of the world's population now living in cities, and this share rising, this means that the larger cities, decreased density and longer commutes that come with decentralized cities, play an important role in shaping the lives of half the world's population. In this section, we consider the two main candidate explanations for this phenomena, falling transportation costs and, for the US only, "white flight".

5.1 Transportation costs and decentralization

We've seen that cities are decentralizing. We'd like to know the extent to which this can be explained by falling transportation costs. Because decentralization seems to have been happening steadily about 200 years, if transportation costs are responsible for decentralization, then we should probably see transportation costs trending down over the same period.

There is pretty good evidence that transportation costs have been falling for hundreds of years. The fact that there have been such dramatic innovations in transportation during this period is suggestive. The period from 1800 to 2000 saw the invention of the horsecar (a horse drawn trolley on rails), the steam powered locomotive and passenger train, the automobile, various sorts of public transit, airplanes, and most recently, telecommuting. It is hard to imagine that these technologies have not reduced the cost of moving people around.

Figure 5.3: 150 years of US freight costs



Note: Left panel shows the share of US GDP that was generated by activities related to transportation over time. The black line indicates all transportation sectors. The gray line is all sectors except for air travel. Right panel shows the cost of moving a ton of freight one mile by rail. Both declined dramatically. Figures reproduced from Redding and Turner [2014], ©Elsevier 2015.

In fact, transportation costs probably began falling even earlier than 1800. In the
 2705 18th century it was common for travelers in England to record their travels and travel times in their diaries. By collecting many of these diaries (about 100) and reading carefully to figure out how long different journeys took, Bogart et al. [2024] document that the speed of travel by stagecoach in Britain about doubled between about 1730 and 1850, most likely because of improvements in the quality of roads.

The cost of travel continued to fall in the 19th century. The left panel of figure
 2710 5.3 reports share of US GDP that was produced by industries that move people or goods. In 1850, about 6 cents of every dollar earned in the US went to pay for moving goods or people. By 2010, these same industries were responsible for less than 2 cents of every dollar. That is, people are using a smaller share of their budgets to move

²⁷¹⁵ themselves and their stuff around (this excludes expenditure on cars). The right panel shows the average cost to move one ton of freight one mile by rail (one “ton mile”) from 1850 until 2010. This price declines by about a factor of 10.

With this said, not all of the available evidence indicates declining transportation costs. The price per ton mile to ship goods by truck in the US stays approximately ²⁷²⁰ constant over the second half of the 20th century [Redding and Turner, 2014]. US travel survey data show that between 1995 and 2008 the average speed of a trip by car falls from 43KPH to 39KPH [Couture et al., 2018]. The cost of building a lane-mile of interstate increases by about a factor of seven between 1955 and 2008 [Brooks and Liscow, 2023, Mehrotra et al., 2024].

²⁷²⁵ Where does this leave us? Certainly, the cost of travel by certain modes has fallen pretty uniformly over the period when we see cities decentralize, but the constant cost of moving goods by truck, the increase in the construction cost of highways and the late 20th century reduction in US driving speed should give us pause.

With this said, my inclination is to discount this contrary evidence. During the ²⁷³⁰ period when truck freight rates were constant and driving speeds were falling, the amount of freight moved by truck increases by about 50% and the amount of automobile travel more than doubles. Seeing these big increases in demand for a service whose cost is constant or rising makes me wonder whether we’ve somehow mismeasured costs. For example, maybe the full cost of a slower trip in a new car, with air ²⁷³⁵ conditioning, power steering and a good stereo, is lower than a faster trip in an older, less comfortable car?

There is also direct evidence showing that changes in transportation costs are related to urban decentralization. As we saw in Chapter 2, Baum-Snow [2007] doc-

uments the relationship between the construction of radial interstate highways and
2740 decentralization of US cities. During this period, an average city in Baum-Snow's sample received 2.6 interstate rays. Baum-Snow shows that each (randomly assigned) ray causes about a 9% drop in central city population. Multiplying, interstate rays caused about a 23% decline in central city population. This is slightly more than was actually observed. Thus, highway construction can explain all decentralization of
2745 US cities. Moreover, the ability of highways to decentralize cities is not particular to US cities. Baum-Snow et al. [2017] shows that highways decentralize Chinese cities, too. Garcia-López et al. [2015] and Garcia-López [2019] provides further evidence for Europe. Gonzalez-Navarro and Turner [2018] provide similar evidence for the construction of subways in a sample of world cities.

2750 The steady ongoing decentralization of cities is incontrovertible. That transportation costs have fallen, at least for some modes of travel, also seems like a pretty safe conclusion. We also have good evidence that radial highways cause decentralization. To the extent that one of the main effects of highways is to lower transportation costs, that the construction of highways causes the decentralization indicates that
2755 cities decentralize in response to falling transportation costs.

5.2 The Great Migration and white flight

The period from 1890 to about 1980 saw a large black migration from the Southern US to the North. Between 1890 and 1940, the black population of Northern and Midwestern cities grew by about 4% per year, and about 2% per year in the West
2760 and South. This migration peaked between 1940 and 1970, when four million blacks

migrated from the South to the North. This increased the black population share of northern cities from 4% in 1940 to 16% in 1970 [Boustan, 2010]. By 1980, 78% of metropolitan blacks lived in central cities, while only 33% of metropolitan whites did so.

²⁷⁶⁵ Cutler et al. [1999] describes the evolution of US cities over this time. This period saw the rise of the black ghetto, primarily in northern US cities. Cutler and Glaeser [1997] analyzes the effects of these ghettos on their black residents, unsurprisingly, the answer is “not so good”. This suggest that black ghettos did not arise because southern migrants wanted to live near each other. On the contrary, these ghettos seem to have been difficult places that self-interested people would have moved away from if they could. Finally, Boustan [2010] analyzes the hypothesis that “white flight” was responsible for the decentralization of US cities we observe over this period. That is, urban whites moved away to the suburbs to avoid newly arrived black migrants, and then contrived to prevent the black migrants from following, leaving them stuck in inhospitable central city ghettos. This section elaborates and provides evidence for this argument.
²⁷⁷⁰
²⁷⁷⁵

5.2.1 The American ghetto, 1890-1990

Cutler et al. [1999] studies the rise and fall of the American Ghetto from 1890 to 1990. The first hurdle the paper must overcome is to figure where people of different races live. It accomplishes this by analyzing a 1% sample of decennial US census data from 1890-1990. Pre-1940, census units are “wards”, post 1940, they are tracts. A census tract is about four thousand people, wards are bigger. These data report the census tract or ward of residence and the race of the census respondent. Cutler
²⁷⁸⁰

et al. [1999] organize these data into a sample of 54 cities in 1890 and this increases
 2785 to 313 by 1990. Post 1940, “cities” are MSAs. Pre-1940, they are the corresponding
 municipalities.

The second main challenge the paper faces is to develop a way to measure seg-
 2790 reation and ghettos. It does this with two indexes of segregation, an “index of
 dissimilarity” and an “index of isolation”. Both indexes try to turn the difficult ques-
 tion of how segregated a city is into a number. I will discuss the dissimilarity index
 in detail. To see the gory details of the index of isolation, see Cutler et al. [1999].

To describe the dissimilarity index, we require notation to describe where people
 of different races live. Let i index census tracts (or wards) and j index cities. Define
 B_i^j and W_i^j as the count of black and white people in census tract i of city j . Also
 2795 define,

$$B^j = \sum_{i \in \text{city } j} B_i^j$$

$$W^j = \sum_{i \in \text{city } j} W_i^j$$

as the total population of black and white people in city j . Finally, define the dis-
 similarity index for city j ,

$$\text{dissimilarity index}^j = \frac{1}{2} \sum_{i=1}^N \left| \frac{B_i^j}{B^j} - \frac{W_i^j}{W^j} \right|. \quad (5.1)$$

The two main terms of this index are the share of the black population in each tract
 2800 and share of white population in each tract. If races are distributed symmetrically
 across tracts, then this index is zero. If all of the blacks are concentrated in a single

tract, then the index is 1.

The index is undefined when $B^j = 0$ or $W^j = 0$. The sample in the paper is restricted to cities with population more than 100k, and more than 10k black to avoid this problem. Empirically, 0.3 is considered “low”, 0.3-0.6 is “moderate”, and
2805 above 0.6 is “high”.

Figure 5.4 presents the main findings from Cutler et al. [1999]. The left panel describes the average value of the dissimilarity index by decade for four different sets of US cities. “All cities” is the average over all of the cities in their data in each year. This series is based on the largest possible sample, but not all cities are present
2810 in all decades, so some of the decade-to-decade changes may reflect changes in the composition of the sample. “Matched-sample” restricts attention to the about 60 cities present in every decade of their sample. This solves the problem of composition at the expense of much restricted coverage. A limitation of Cutler et al.’s data is that a “city” is a municipality up until 1950 and an MSA afterwards. It’s not clear the best
2815 way to fix this problem, but the central part of an MSA is probably a better match to the 1950 and before municipal boundaries than is the whole MSA. The “Central city” line is based on the same sample of cities as the “All cities” line, but restricts attention to the population of central cities post 1950. Finally, the “Weighted” line is the same as the “All cities” line, but each city in the average is weighted by the size
2820 of its black population rather than all cities being treated equally. The right panel of figure 5.4 is identical, except that it describes the evolution of the isolation index.

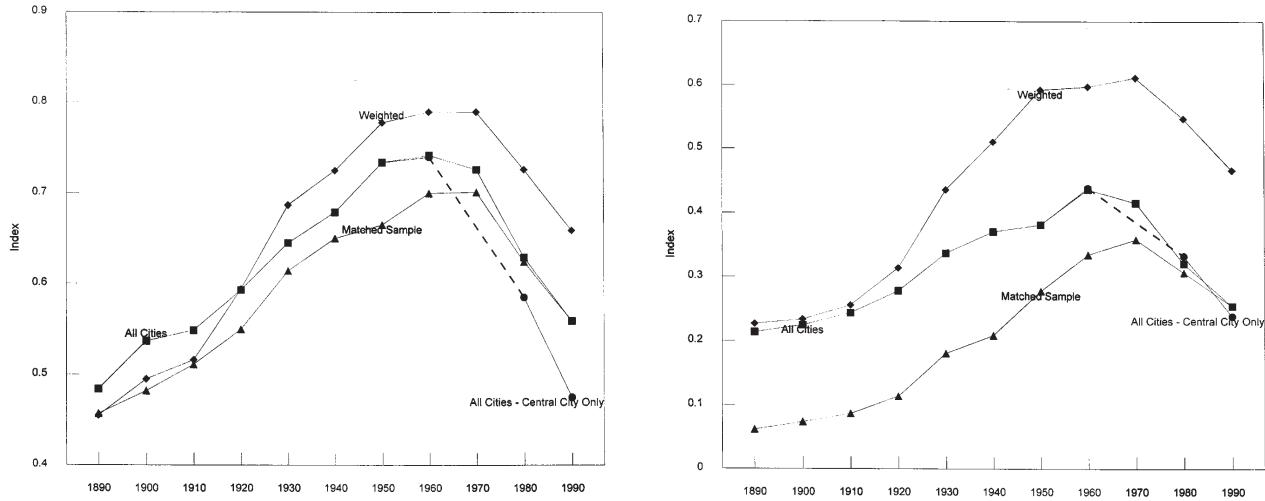
In every case, the trajectory of the dissimilarity index suggests the same story. Segregation was low at the beginning of the 20th century, rose steadily until about 1940, peaked in 1970 and then fell dramatically in the next two decades. The isolation

²⁸²⁵ index tells much the same story, although it does not show the same leveling off around 1940.

Looking at city-by-city values of the indexes within a decade suggests two additional features of segregation. First, it was worse in the Northern cities. Second, the value of either index in one decade is highly predictive of its value in the next decade.
²⁸³⁰ This is important and reassuring. The boundaries of cities in the Cutler et al. [1999] sample change over time and the building block of the indexes changes from ward to tract. The fact that the “Central city” line in both panels of figure 5.4 looks much like the others, along with the fact that patterns of segregation are persistent within cities suggests that changes in the index values are not simply artifacts of these problems
²⁸³⁵ with the data.

As a further check on the pattern shown in figure 5.4, Starting in 1940, Cutler et al. [1999] calculate a simpler statistic, though one that they can only evaluate with their more recent data. If we start from the premise that people like more-or-less the same things, regardless of the color of their skin, then there should be some black person who wants to live in every neighborhood that is attractive to a white person, and conversely. In this case, we should see few exclusively white or exclusively black neighborhoods. Letting census tract substitute for neighborhood, table 5.1 shows the share of urban census tracts where the share of black population was either exactly zero or was less than 1%, by decade. In 1940, fully 60% of urban census tracts had less than 1% black population, even though 37% of the urban population was black.
²⁸⁴⁰ This share is higher in 1960 and falls dramatically by 1990, just as for the indexes of segregation shown in 5.4. It is possible to perform this calculation for the central part of cities and their suburbs separately for 1960 and 1990. This measure of segregation

Figure 5.4: Segregation in US cities during the 20th century



Note: *Left panel reports on the history of the dissimilarity index described in equation (5.1). According to this index, us cities became home to populations of increasingly segregated blacks over the first half of the 20th century. This trend flattened out between 1940 and 1970, before falling rapidly until 1990. Right panel reports the history of a second, different index, the index of isolation, over the same period. This index shows about the same pattern as the dissimilarity index in the left panel. The different lines in each sample describe slightly different samples of cities, but all show the same basic pattern.* Figures reproduced from Cutler et al. [1999], ©University of Chicago Press.

was worse in the suburbs than in the central cities.

Cutler et al. [1999] say that a city has a ghetto if the dissimilarity index is above 0.6 and the isolation index is above 0.3. Table 5.2 reports on the evolution of black ghettos in US cities. In 1890, the black population of only one city was segregated enough to satisfy the definition of a ghetto, although 7.5% of the urban population was already black.²

²This is not to say that there was no segregation. In 1890, the average black person lived in a census tract where 20% of the population was black, even though only 7.5% of the population was black.

Table 5.1: Distribution of percentage black in census tracts

	1940		1960		1990	
	City	City	Suburbs	City	Suburbs	
# tracts	6,133	13,310	9,378	16664	27,183	
% tracts with black share:						
Exactly 0	21.2	19.6	22.3	7.3	14.7	
0-1	39.1	36.2	48.0	10.2	25.0	

Note: *Results reproduced from Cutler et al. [1999]*

Table 5.2: Demographic change and segregation

	Year				
	1890	1910	1940	1970	1990
Number of ghettos, all cities	1	5	55	127	98
% cities with ghettos	1.7	7.0	50.5	60.2	29.5
% black pop. in a city with a ghetto	1.7	4.6	72.4	93.1	72.4
% black population	7.5	7.1	10.8	13.9	16.2
% black in tract of average black	20.0	22.6	37.6	69.7	60.9

Note: *Results from Cutler et al. [1999]*

2855 The incidence of ghettos increased over time. By 1970, 127 cities had ghettos, 60% of all cities in the sample, 93% of the black population lived in a city with a ghetto, and the average black person lived in a census tract where almost 70% of the other residents were also black. The 1890 to 1970 increase in segregation occurred even though the black share of the total urban population only about doubled from 2860 its level in 1890. Each of these statistics improved between 1970 and 1990. In all, these data suggest that the black ghetto in US cities was a creation of the early part of the 20th century. It was most prevalent in the 1970s, and began to retreat by the end of the 20th century.

At the end of the 20th century and the beginning of the 21st century, in spite
 2865 of ongoing decentralization, the composition of central cities began to change in a
 different way. Table 5.3 reports on mean incomes within certain distance bands as
 distance to the center increases. In 1980, the average MSA resident who lived within
 one mile of the center had an income of just 89% of the MSA average, by 1990, this
 number had increased to 94%, while beyond three miles from the CBD, mean incomes
 2870 fell slightly relative to the MSA average. For the 10 largest MSAs, the bottom panel
 of table 5.3, the change is more dramatic. For an average city in the top 10, incomes
 increased within two miles of the CBD and decreased beyond five miles. After 1980,
 at least parts of these central cities began to attract wealthier and wealthier people.

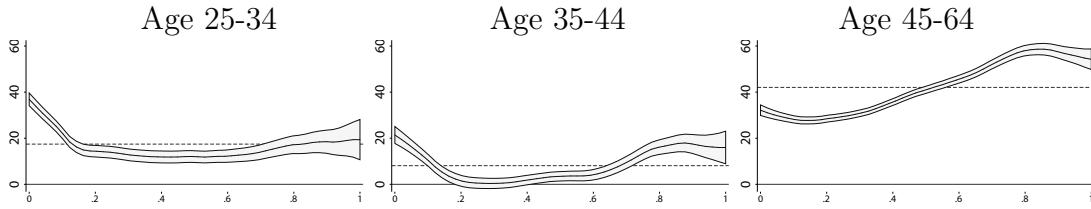
Figure 5.5 gives more detail on this trend and shows that it extended into the
 2875 first decade of the 21st century. This figure is like figure 5.2, but rather than showing
 the decentralization of all people, shows the centralization of college educated young
 people. These data suggest that the end of the 20th century saw not only a decline
 in the importance and prevalence of black ghettos, but also increasing gentrification
 of previously blighted and impoverished areas.

2880 5.2.2 Causes of ghettos

It is natural to ask why ghettos arose. Cutler et al. [1999] ask whether this was
 related to the arrival rate of black migrants. To proceed, as in equation (5.1) let j
 index cities and i index census tracts. Also define the variable,

$$\text{High segregation}_{1910}^j = \begin{cases} 1 & \text{if } j \text{ is in top half of cities for dissimilarity in 1910} \\ 0 & \text{if } j \text{ is in bottom half of cities for dissimilarity in 1910} \end{cases}$$

Figure 5.5: Central migration of educated young people, 2000-2010



Note: Percent change in the college-educated population in three age brackets at different distances from the city center, 2000-2010. Distance from the city center is measured as the cumulative share of CBSA population in the base year. The shaded region around kernel fit depicts the 95% confidence interval. The dashed row in each plot shows the average population growth rate for the relevant demographic group over the relevant decade. Data are from the 1980-2000 decennial censuses and the 2008-2012 ACS. Figure reproduced from Couture and Handbury [2017], ©Journal of Urban Economics.

Table 5.3: Increasing US central city incomes in the 1980's and 1990's

All MSAs		
Income relative to city average	1980	1990
Within one mile of CBD	89%	94%
Within two miles of CBD	95%	95%
Within five miles of CBD	101%	100%
Beyond five miles	109%	107%
10 Largest MSAs		
Within one mile of CBD	144%	163%
Within two miles of CBD	88%	97%
Within five miles of CBD	86%	86%
Beyond five miles	105%	100%

Note: The biggest central cities saw relative increases in income between 1980 and 1990. This was less relevant outside the very largest cities. Results reproduced from Glaeser et al. [2001].

Table 5.4: Difference in differences estimates of the effects of ghettos

	High School Graduate	College Graduate	Idle
Black:			
Low segregation	80.0%	10.7%	15.8%
High segregation	77.2%	12.0%	21.3%
Difference	-2.8%	1.3%	5.5%
White:			
Low segregation	88.1%	23.9%	9.9%
High segregation	89.3%	28.7%	9.4%
Difference	1.2%	4.8%	-0.5%
Difference in Differences	-4.0%	-3.6%	6.0%

Note: *Black outcomes at age 25-30 are worse than non-black outcomes in 1990 in 209 MSAs with 100k or more population and 10k or more black population. Results reproduced from Cutler and Glaeser [1997].*

That is, an indicator for the cities with high dissimilarity index values. Also, define
 2885 the corresponding indicator variable for the isolation index. Then, for each of the two city level indexes of segregation, isolation and dissimilarity, Cutler et al. perform the regression,

$$\begin{aligned} \text{Segregation index}_t^j &= A_0 + A_1 \Delta \ln B_t^j + A_2 \Delta \ln W_t^j \\ &+ A_3 \text{High segregation}_{1910}^j + A_4 \text{High segregation}_{1910}^j \times \Delta \ln B_t^j + \varepsilon_t^j. \end{aligned}$$

The parameters of interest are A_1 and A_4 . If $A_1 > 0$ then the index of interest increases when the rate of black migration is larger. If $A_4 > 0$, it tells that the index of interest increases more rapidly in response to black migration in places that are initially more segregated. Their results strongly support the conclusion that A_1 and A_4 are positive. That is, the segregation indexes increase with the rate of black
 2890

migration into the city, and the increase in the segregation index is more rapid in places that were initially more segregated.

2895 In all, this regression confirms the story suggested so far. The early part of the 20th century saw a large migration of blacks out of the rural south to cities. This period also saw the rise of the black ghetto in US cities. Ghettos were more likely in cities that saw more black migration. They were also more likely in places that were more segregated in 1910.

2900 The arrival of black migrants in otherwise predominantly white cities is necessary for the growth of black ghettos, but black migrants could have arrived without creating ghettos. This means that the question of why ghettos formed is still open. One possible explanation is that ghettos arise because black people prefer to live near other black people. To investigate this, Cutler et al. [1999] analyze the General 2905 Social Survey (from 1970 to 1990) which asks a series of questions about peoples' attitudes towards race. More specifically, the survey asks black people if they would rather live in a majority white neighborhood. They find that black responses are unrelated to the dissimilarity index of the respondent's city. This does not suggest that black people who do not want to live around white people are sorting themselves 2910 into highly segregated cities.

The survey also asked white people a series of questions about their attitudes towards blacks; "Do you believe there is a right to segregated housing?", "Do you support a ban on interracial marriage?", and, "Do you want to live in a 50% black neighborhood?" White people in more segregated cities were more likely to believe in 2915 a right to segregated housing. However, white people in more segregated cities were less likely to support a ban on interracial marriage, and there was no relationship

between the dissimilarity index in the respondent's city and their willingness to live in a 50% black neighborhood. These responses provide some support for the idea that white people who want to live in segregated cities have sorted themselves into these cities, though the index does not otherwise consistently predict white attitudes towards blacks.

Economists are usually distrustful of surveys. If people don't have a real choice to make, it's hard to have much confidence that they are telling us what they would really do. To learn about whether ghettos are places black people would seek out or shun, we would like to check whether or not they are harmful to their black residents. Cutler and Glaeser [1997] do exactly this. They look at census data describing demographic outcomes that can be matched to cities in 1990. For each city, they calculate the same dissimilarity index as in Cutler et al. [1999]. They can then ask whether the outcomes of black people living in highly segregated cities are worse than those who live in less segregated cities.

In general, black people had worse economic outcomes than white people in US cities in 1990. Because black people were also much more likely to live in black ghettos than white people (see table 5.1), this suggests the possibility that ghettos may be at least partly to blame.

However, it may also be the case that segregation is not the cause of the problem. Rather, segregation may occur in places that are otherwise inhospitable. To check this, Cutler and Glaeser [1997] look at whether blacks are harmed more by segregation than whites. That is, they calculate a "difference in differences" estimate of the effect of segregation, as measured by the dissimilarity index, on the economic outcomes of black people relative to white.

To begin, define $Y_L^B, Y_L^W, Y_H^B, Y_H^W$ to be black and non-black outcomes in high and low segregation cities, where “high segregation” is defined as above median dissimilarity index. This done, we can calculate the difference in differences estimator,

$$(Y_H^B - Y_L^B) - (Y_H^W - Y_L^W).$$

This statistic allows us to check whether segregated MSAs were more harmful to blacks than to whites. This is what we would expect to see if ghettos are specifically harmful to blacks, rather than just being in places where no one could thrive.

Table 5.4 evaluates this statistic for three outcomes using a sample of 25-30 year olds who live in one of the 209 MSAs with greater than 100k of population in 1990.

At the top of the first panel, we see that in cities with low values of the dissimilarity index, the high school graduation rate among black 25-30 year olds is 80.0% and in high segregation MSAs, it is 2.8% lower at 77.2%. For white 25-30 year olds, the high school graduation rate in low segregation MSAs is 88.1% and is 1.2% higher, 89.3% in high segregation cities. Thus, moving from low to high segregation MSAs increases the gap in the graduation rate of black and white 25-30 year olds by 4%. The next two columns of table 5.4 are similar, but consider college graduation rates for 25-30 year olds and the likelihood of being idle, that is, unemployed and not seeking work. Blacks are less likely to graduate from college, and are more likely to be idle relative to their white peers in segregated MSAs than not.

These data suggest that ghettos are harmful. They are places that self interested people would move away from if they could. Together with the survey data described above, this does not look good for the hypothesis that ghettos arise because blacks

want to live near other blacks (although this surely sometimes occurs).

We next consider whether ghettos arise because whites are fleeing neighborhoods populated by blacks, and whether this flight is responsible for the decentralization 2965 of US cities. Boustan [2010] examines this hypothesis by estimating the number of whites who leave the central cities per black arrival.

Boustan's data describe a subset of US MSAs between 1940 and 1970. The number of cities increases over time from 59 to 212.³ Suppose that we observe whites leaving when blacks arrive. How should we interpret this? It could mean that blacks are 2970 migrating to places whites were leaving anyhow, or it could mean that whites are leaving because the blacks arrive. To distinguish between these possibilities, we would ideally assign blacks to cities at random and see what happens.

Boustan [2010] tries to approximate such random assignment by focusing attention on blacks who migrate because of short run fluctuations in the economies of their 2975 southern origin counties. The results of this analysis suggest that whites leave central cities because blacks arrive.

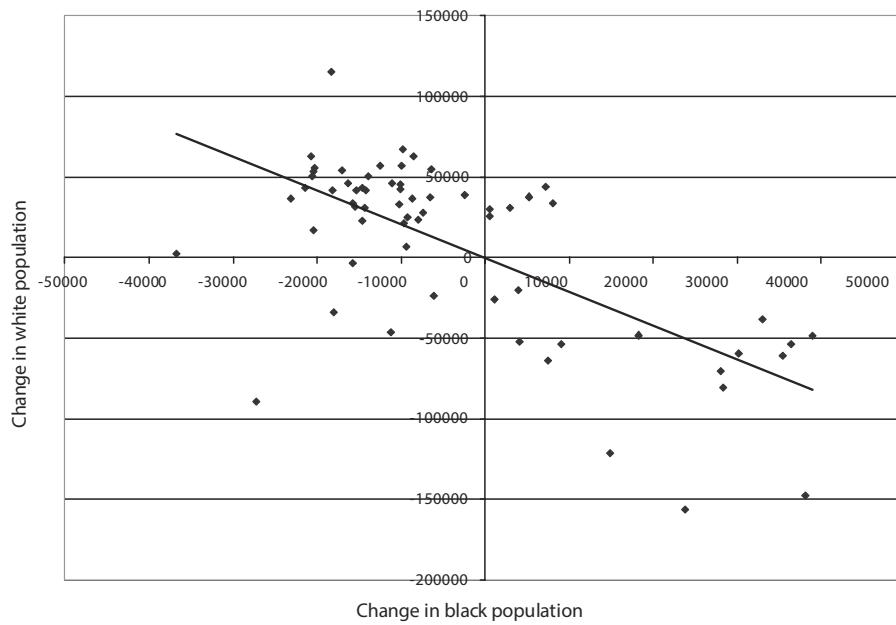
Figure 5.6 shows Boustan's main result. To understand this figure, let i index MSAs, and let B_i and W_i be the black and white population of the central city of MSA i . Next, conduct the regression,

$$W_i = A_0 + A_1 B_i + A_2 \Delta_{\text{MSA}} \text{pop}_i + A_3 \text{Region}_i + \varepsilon_i.$$

2980 The dependent variable in this regression is the white population of the central city. The independent variable of interest is the black population. Because the regression controls for changes in MSA population, the coefficient A_1 measures the change in

³The increase reflects an increase in the number of cities over time.

Figure 5.6: Change in black and white populations in central cities, 1950–1960



Note: *Each point in the scatter diagram represents the residual change in a city's black and white populations after controlling for region fixed effects and changes in the metropolitan area's population over the decade. The slope of the regression line through these points is -2.010. The four largest cities—Chicago, IL; Detroit, MI; Los Angeles, CA; and New York City, NY, are omitted from figure 5.6 to improve its legibility. Figure reproduced from Boustan [2010], ©Oxford University Press.*

central city white population that results from the arrival of one black central city resident, holding the population of the MSA constant. Depending on econometric details, Boustan estimates that A_1 is between about 2.0 and 2.7. That is, for each black person arriving in the central city, between 2.0 and 2.7 whites leave. This is white flight.²⁹⁸⁵

Figure 5.6 illustrates this result. The x -axis of this figure gives the change in the black population of the MSA between 1950 and 1960. The vertical axis gives the

2990 corresponding change in the white population. Each dot in the graph is an MSA. That
the central city white population decreases with the arrival of the black population is
clear in the negative trend in the data. The straight line plotted in the figure has a
slope of about -2.0 and is a simple example of the regression described above. That
this line has a slope of -2.0 indicates that, on average, two whites leave for each black
2995 arrival.

Consistent with the White Flight argument, white dominated governments, city,
state, local, created legal barriers to black residence in white neighborhoods. This
is documented in Rothstein [2017]. Some of these institutional barriers were simply
horrifying. Restrictive covenants in property deeds often prohibited black ownership
3000 or residence, except in the role of servant. The courts were willing to enforce these
covenants, and sometimes dispossessed black owners of deed restricted properties.
The federal government offered subsidies for mortgages to veterans, but would not
grant these subsidies to black veterans or to white households seeking to live in black
neighborhoods. The constitution of the state of Oregon flatly prohibited the residence
3005 of blacks in the state from 1857 until 1926.

These institutions compelled blacks to live in black neighborhoods. Empirically,
this will “look like” blacks want to live near other blacks, and in a strict sense, they
do. Living in black neighborhoods allowed them to avoid the penalties attached to
violating these laws, e.g., dispossession. These institutions were broadly disallowed
3010 with civil rights legislation in the late 1960s. Interestingly, the supreme court disal-
lowed restrictive covenants in 1948, just about the time that we see the segregation
indexes begin to flatten out in figure 5.4.

5.3 Transportation costs, white Flight, and decentralization

³⁰¹⁵ Boustan estimates that the median Northern and Western city received 19,000 black migrants between 1940 and 1970. At 2.7 white departures per black arrival, this means that about 51,000 whites left the median central city in response. Because Boustan's regressions condition on metropolitan area population, this is migration from the central city, holding metropolitan population constant. That is, it is white suburban migration. In her sample, this works out to a 17% decline in central city population. Baum-Snow [2007], discussed in Chapter 2, reports a decline in central city population of exactly 17% between 1950 and 1990. Baum-Snow's results suggest that almost all of this effect can be explained by radial interstate highways. This raises an obvious problem. Between highway construction and white flight, we can ³⁰²⁰ explain twice as much decentralization as actually occurred. How can we resolve this apparent contradiction?

³⁰²⁵ It could be that the estimates of the effects of highways on decentralization are wrong. However, the Baum-Snow estimates have been replicated using similar data and methodology in China and Europe. It could also be that Boustan's results are not right. This seems more plausible. We observe decentralization in cities around the world, in particular, in countries less obsessed with skin color than the US. Finally, it could be that both Baum-Snow and Boustan are right, but both overestimate effects to some degree. Recall that all regressions actually result in coefficients estimated with a certain precision. Really, they are returning likely ranges for the coefficients, ³⁰³⁰ rather than actual numbers. With this in mind, I note that the precision of Boustan's

estimates does not rule out that only 1.4 whites left the central city per black arrival. This means that an effect about half what she estimates is also plausible. This would cut her estimate of the white flight effect size about in half. The precision of Baum-Snow's estimates is about the same as Boustan's. So it is possible, if unlikely, that 3040 both estimates are actually at the bottom of the likely range. If this is the case, then highways would explain about half of the realized decentralization, and white flight the other half.

Finally, it might be Boustan and Baum-Snow are both about right, but Boustan's results need to be reinterpreted. The rate of decentralization is determined by highways and transportation costs. The identity/color of the people who decentralize is 3045 determined by black migration patterns. I think this one gets my vote.

5.4 Sorting and bid-rent

There has been a long history of decentralization, both in the US and around the world. There is good evidence that this decentralization was caused by reductions 3050 in transportation costs. In the US, decentralization was also caused by (or maybe contributed to) the concentration of blacks in the center city that followed from the Great Migration of rural blacks from the South into cities. While decentralization of cities is ongoing, the concentration of blacks into inhospitable central city ghettos has been on the decline since about 1980. The nearly opposite phenomena, the 3055 central migration of affluent or highly educated young people, gentrification, seems to have been underway since the late 20th century, particularly in big cities, up until the 2020 Covid pandemic. We now ask whether we can explain this process with

the monocentric city model using a version of the basic model based on LeRoy and Sonstelie [1983] and developed independently by Glaeser et al. [2008].

3060 Up until now, we have assumed that all households are the same. This is a whopper. Now, we will do a little better and suppose that there are two types of households, “rich” (r) and “poor” (p). The two types are the same, except that wages for the rich are higher than for the poor, $w_r > w_p$.

To explain patterns of decentralization and white flight, also suppose that the 3065 transportation technology is more complicated. There are two modes of transportation, car and bus, indexed by a (for auto) and b . Each mode involves a cost in minutes per unit distance, t^a and t^b . A minute spent commuting is treated as a minute less earning wages, and so commute time is valued at the wage rate. This means that the time cost of commuting is higher for the rich than the poor. There is also a 3070 money cost per unit distance, c^a and c^b (beware, these are commute costs, not to be confused with consumption c). Finally, cars also involve a fixed cost, f^a , that is the same regardless of commute distance. Assume $f^a > 0$, $c^b < c^a$ and $t^b > t^a$. That is, cars are more expensive than buses, but also faster. Putting this all together, the total cost of a commute of distance x by car is $f^a + (wt^a + c^a)x$, and the total cost 3075 for the same commute by bus is $(wt^b + c^b)x$.

Everything else is the same as the standard monocentric model. In particular, the reservation utility levels for the two types are the same, $\bar{u}_p = \bar{u}_r = \bar{u}$, and land consumption is the same for both types, $\bar{\ell}_p = \bar{\ell}_r = \bar{\ell}$.

A poor bus riding household solves,

$$\begin{aligned} & \max_{c,x} u(c) \\ \text{s.t. } & w_p = c + R_p^b(x)\bar{\ell} + (w_p t^b + c^b)|x|. \end{aligned}$$

- 3080 This is exactly the same problem that a representative household faces in the basic model, but with a more complicated expression describing commuting costs. Notice that to make this budget make sense, w_p is the wage for a whole day, so t has to be the fraction of a day spent commuting. As in the original analysis of the monocentric city model in Chapter 1, free mobility requires that $u(c) = \bar{u}$ for all locations x . Letting
 3085 $c^* = u^{-1}(\bar{u})$, this means that land rent is given by,

$$R_p^b(x) = \frac{w_p - c^* - (w_p t^b + c^b)|x|}{\bar{\ell}},$$

ignoring the corners that determine the edges of the city.

To accommodate the two types of households in the model, we need to restate this problem. In particular, an equivalent statement of a poor bus riding household's problem is,

$$\begin{aligned} \Psi_p^b(x) &= \max_c \frac{w_p - c - (w_p t^b + c^b)|x|}{\bar{\ell}} \\ \text{s.t. } & u(c) \geq \bar{u} \end{aligned}$$

- 3090 This is an ugly looking expression, but it is actually quite simple. We know that in any spatial equilibrium, we must have $u(c) = \bar{u}$, or equivalently, $c = c^*$. This means

we can rewrite this expression as,

$$\Psi_p^{b*}(x) = \frac{w_p - c^* - (w_p t^b + c^b)|x|}{\bar{\ell}}$$

The numerator of the fraction on the right is all income left over after the household pays for commuting and consumption. The denominator is land consumption. Thus,
3095 $\Psi_p^{b*}(x)$ is the most that a bus riding poor household can pay to live at x and still afford the reservation consumption level. The function $\Psi_p^{b*}(x)$ is called a “bid-rent” function.

Let’s do the same thing for a rich bus rider. A rich bus riding household’s problem is,

$$\begin{aligned} & \max_{c, x} u(c) \\ & \text{s.t. } w_r = c + R_r^b(x)\bar{\ell} + (w_r t^b + c^b)|x|. \end{aligned}$$

3100 Using exactly the same logic as for the poor bus riders, this gives us a bid-rent function for the rich bus riders,

$$\Psi_r^{b*}(x) = \frac{w_r - c^* - (w_r t^b + c^b)|x|}{\bar{\ell}}.$$

Suppose that the fixed cost of owning a car, f^a , is large enough that no rich household ever buys a car. This means that we can ignore both rich and poor driving households, and an equilibrium city contains only rich and poor bus riders. What
3105 does such an equilibrium city look like? Where do the two types of households end up and what does the land rent gradient look like?

To answer these questions, note that because $w_r > w_p$,

$$\frac{w_r - c^*}{\bar{\ell}} = \Psi_r^{b*}(0) > \Psi_p^{b*}(0) = \frac{w_p - c^*}{\bar{\ell}}. \quad (5.2)$$

This means that if we evaluate both bid-rent functions at $x = 0$, the bid-rent of the rich is larger than the bid-rent for the poor. The rich bus rider can pay more than
3110 the poor bus rider to live at $x = 0$ and still have enough left over to achieve the reservation utility level.

Now, notice that, also because $w_r > w_p$,

$$\frac{-w_r t^b - c^b}{\bar{\ell}} = \frac{d\Psi_r^{b*}}{dx} < \frac{d\Psi_p^{b*}}{dx} = \frac{-w_p t^b - c^b}{\bar{\ell}}, \quad (5.3)$$

when $x > 0$, and the reverse when $x < 0$. That is, bid-rent for the rich slopes down more steeply than does bid-rent for the poor, both to the left and right of the origin.
3115 Even though the rich value $x = 0$ more highly than the poor, as we move further and further from the center, and x gets bigger, eventually we arrive in a neighborhood where the poor can pay more than the rich and still achieve the reservation utility level. Travel is cheaper for the poor because their value of time, here the wage, is lower.

3120 In equilibrium, we should have each location x occupied by the type willing to pay the most for it. That is,

$$R(x) = \max \{\Psi_r^{b*}(x), \Psi_p^{b*}(x)\} \text{ for all } x.$$

In words, the rent gradient should be the upper envelope of the bid-rent functions for

the two types. Each location is occupied by the type of household that is able to pay the most and still afford c^* .

³¹²⁵ We need two comments about this expression. First, this statement of the rent gradient is a little bit imprecise. No landlord should ever pay a tenant to occupy a parcel, so land rent cannot go below zero. This is not explicit in the equation above. To fix this, we would want to write the expression for the rent gradient as,

$$R(x) = \max \{\Psi_r^{b*}(x), \Psi_p^{b*}(x), 0\} \text{ for all } x.$$

To keep notation a little simpler, I will stick with the first version, even though it ³¹³⁰ means we need to remember that negative rents are impossible.

Second, this is the second time we have used this condition, although it is the first time I am stating it explicitly. In the monocentric city model of Chapter 1, the edge of the city is determined by the location where farmers begin to outbid urban dwellers for land. If we were going to account for the fact that farmers also compete ³¹³⁵ for land and value all land for its productivity in agriculture, \bar{R} , as in Chapter 1, then equilibrium land rent would be determined by the highest value among these three types of households. That is,

$$R(x) = \max \{\Psi_r^{b*}(x), \Psi_p^{b*}(x), \bar{R}\} \text{ for all } x.$$

Throughout this section, I am also implicitly setting $\bar{R} = 0$, and ignoring it to keep notation simple.

³¹⁴⁰ Taken together, equations (5.2) and (5.3), require that the bid-rent for rich bus riders is above that of the poor bus riders near the center, but slopes down more

quickly. This means that once we are sufficiently far from the center, the bid-rent for the poor bus rider must be above that of the rich bus rider. Because the type that values the location most highly occupies it in equilibrium, this means that rich bus riders occupy a region near the CBD and the poor occupy a region further out.

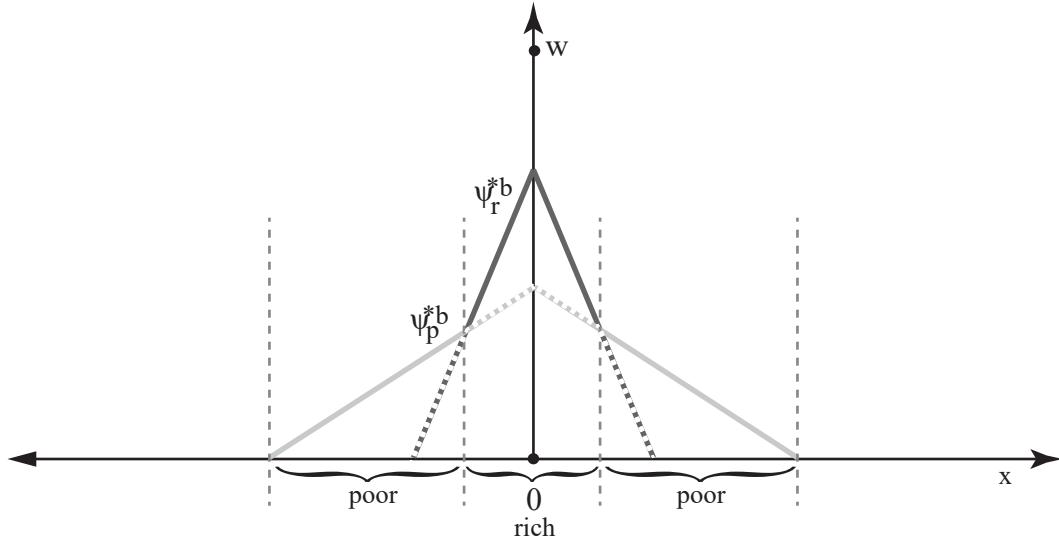
Figure 5.7 illustrates this equilibrium. The black line in the figure illustrates the bid-rent function for rich bus riders and the gray line illustrates the bid-rent function for poor bus riders. Bid-rent for the rich bus riders is above that of the poor bus riders for locations near the center, but not further away. The upper envelope of the two bid-rent functions is the solid kinked gray and black line. Regions where the black line is above the gray line house rich bus riders and regions where the black line is above the gray line house poor bus riders.

Notice that the edges of the city are determined by the intersection of the poor bus riders' bid-rent with the x -axis. This is an implication of the assumption that the value of land in its best alternative use, \bar{R} , is zero.

Now suppose that the fixed price of cars falls enough that rich people sometimes buy them (note that this is a decrease in the cost of transportation). In this case, a rich driving household's problem is,

$$\begin{aligned} & \max_{c,x} u(c) \\ \text{s.t. } & w_r = c + R(x)\bar{\ell} + f^a + (w_r t^a + c^a)|x|. \end{aligned}$$

Figure 5.7: Equilibrium with rich and poor bus riders



Note: The black line in the figure illustrates the bid-rent function for rich bus riders and the light gray line illustrates the bid-rent function for poor bus riders. Bid-rent for the rich bus riders is above that of the poor bus riders for locations near the center. The upper envelope of the two bid-rent functions is the solid kinked gray and black line. Regions where the black line is above the light gray line house rich bus riders and regions where the black line is below the light gray line house poor bus riders.

Alternatively, bid-rent for rich drivers is,

$$\begin{aligned} \Psi_r^a(x) &= \max_{c,x} \frac{w_r - c - f^a - (w_r t^a + c^a)|x|}{\bar{\ell}} \\ \text{s.t. } u(c) &\geq \bar{u}. \end{aligned}$$

³¹⁶⁰ By the same logic as we used for rich and poor bus riders above, this means that

$$\Psi_r^{a*}(x) = \frac{w_r - c^* - f^a - (w_r t^a + c^a)|x|}{\bar{\ell}}.$$

We now know the bid-rent functions for rich and poor bus riders, and for rich drivers. We would like to know what an equilibrium city housing these three types of households looks like. Generalizing from the case of two types, with three types, the rent gradient is the upper envelope of the bid-rent function for all three types of households, rich and poor bus riders and rich drivers, rather than just two types.
³¹⁶⁵ That is,

$$R(x) = \max \{ \Psi_r^{a*}(x), \Psi_r^{b*}(x), \Psi_p^{b*}(x) \} \text{ for all } x.$$

Similarly, the way that the types arrange themselves in the city depends on which type is willing to pay the most for each location. Therefore, figuring out what happens in an equilibrium city with three types boils down to figuring out the regions where each type is able to outbid the others. Not too surprisingly, this is a little complicated.
³¹⁷⁰

To begin, evaluate the intercept of a rich driver's bid-rent at $x = 0$. This gives,

$$\Psi_r^{a*}(0) = \frac{w_r - c^* - f^a}{\bar{\ell}}.$$

Using this expression, we can compare the intercept of a rich driver's bid-rent to that of rich and poor bus riders, and figure out which type is willing to pay the most to live at the center.

³¹⁷⁵ Next, evaluate the slope of a rich driver's bid-rent. For $x > 0$ this gives,

$$\frac{d\Psi_r^{a*}}{dx} = \frac{-w_r t^a - c^a}{\bar{\ell}},$$

with the opposite result when $x < 0$. With the intercept and the slopes of the bid-rent curves for all three types, we can plot their bid-rent curves and see which

type is willing to pay the most for each location. This will completely describe the equilibrium city.

³¹⁸⁰ It turns out that, depending on the parameters of the commute cost functions and the wages of the two types, many equilibrium configurations are possible. I want to restrict these parameters in ways that are consistent with what we have learned about the history of cities. First, suppose that cars are expensive enough that

$$w_r - f^a < w_p \quad (5.4)$$

³¹⁸⁵ That is, after paying the fixed cost for a car, a rich household's residual income is less than the income of a poor household. In this case, if we evaluate the bid-rent for rich drivers and poor bus riders at $x = 0$, we see that poor bus riders can pay more than rich drivers to live at $x = 0$ and still achieve the reservation utility level. More formally, this means that,

$$\frac{w_r - c^* - f^a}{\bar{\ell}} = \Psi_r^{a*}(0) < \Psi_p^{b*}(0) = \frac{w_p - c^*}{\bar{\ell}}.$$

³¹⁹⁰ This is intuitive. At $x = 0$ the big fixed investment in an automobile has no value because the commute is too short. Given that we've assumed that $w_r - f^a < w_p$, this means that rich drivers and poor bus riders have the same free commute at $x = 0$, but rich drivers have less money once they pay for the car that they don't use, and so poor bus riders can outbid them to live near the center. Recalling equation (5.2), this means that rich bus riders outbid both the poor bus riders and the rich drivers to live in the area adjacent to the center.

³¹⁹⁵ Finally, suppose that cars are “enough” faster than buses. More formally, suppose

t^a is enough smaller than t^b that,

$$\frac{d\Psi_p^{b*}}{dx} = \frac{-w_p t^b - c^b}{\bar{\ell}} < \frac{-w_r t^a - c^a}{\bar{\ell}} = \frac{d\Psi_r^{a*}}{dx}.$$

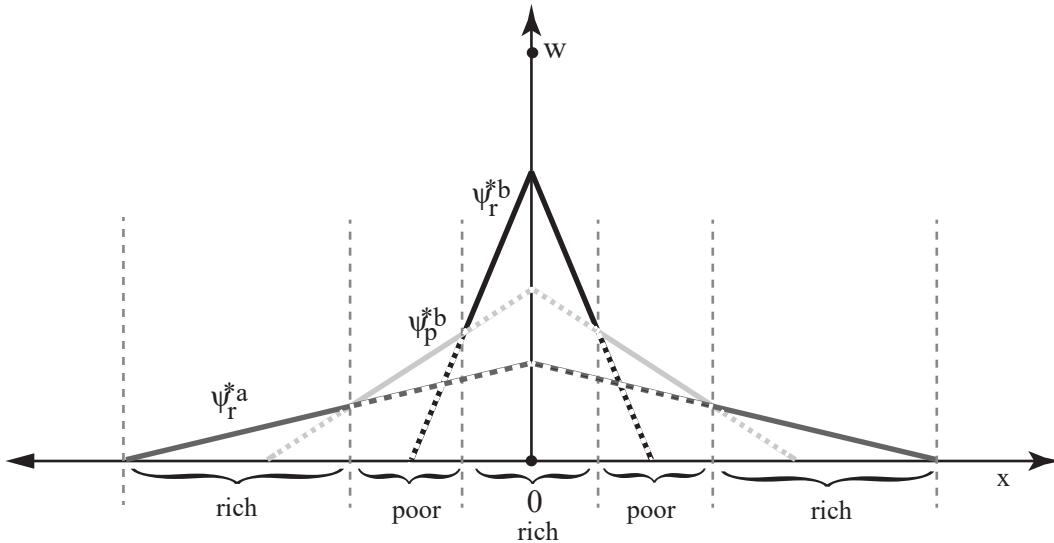
That is, the poor bus rider's rent gradient slopes down more steeply than the rich driver's. Using this condition, together with equation 5.3, we can rank the slopes of
 3200 the bid-rent functions for the three types of households. The bid-rent for the rich bus riders is steepest. Poor bus riders are next, and rich drivers have the flattest bid-rent curves.

With these restrictions on the intercepts and slopes of the bid-rent functions for the three types in place, we get an equilibrium like the one illustrated in figure 5.8.
 3205 In this equilibrium, the rich drive in from the suburbs, the poor live near the center and take the bus, and more rich people live even near the center and take the bus.

Comparing the equilibrium cities described by figures 5.7 and 5.8 we see the extent to which the monocentric city model is able to explain the patterns of decentralization and segregation that we see in cities in the 20th century. The main difference between
 3210 the two figures is transportation costs. The costs of car ownership are lower for the equilibrium in figure 5.8 than 5.7. As a consequence of this reduction in the cost of car ownership, we see an equilibrium city transform from one where poor people are peripheral and the rich occupy the center, to one where the central part of the city is substantially occupied by poor people, and the rich commute by car from more
 3215 remote suburbs.

This is a pretty good match to the first part of the 20th century. During this time, cities decentralized pretty continuously. This is more-or-less what we are seeing

Figure 5.8: Equilibrium with rich and poor bus riders and rich drivers



Note: The black line describes the bid-rent function for rich bus riders, light gray is for poor bus riders, and dark gray is for rich drivers, under the assumptions given in the text. The equilibrium rent gradient is the upper envelope of these three bid-rent functions, the solid black, light gray, and dark gray lines. In this equilibrium, rich bus riders live near the center, poor bus riders live just further out, and rich drivers live beyond the poor bus riders.

as we go from figure 5.7 and 5.8. During this time period, we also saw the rise of the black ghetto. That is, the increasing concentration of relatively poor blacks in the center city. Figures 5.7 and 5.8 show a similar change in patterns of segregation by income in response to a drop in transportation costs. If we conflate income and race, this is like what we saw in US cities in the early and middle of the 20th century.

It is important not to overstate the case here. The monocentric city model is an abstraction and it does not address the many institutional barriers that restricted the location choices of black people during this period, nor does it allow for the different

3220

3225

types to like or dislike proximity to each other. With this said, what this analysis does show is that even a simple description of income heterogeneity, together with a more realistic description of transportation costs, can go a long way towards explaining what we have observed.

3230 Our analysis so far fails to reflect two details of what we observe. First, even in the equilibrium with three types illustrated in figure 5.8, the rich occupy the very center of the city. This is probably not a good description of what happened in US cities in the middle of the 20th century. There was not a central rich enclave surrounded by poor neighborhoods surrounded by rich suburbs.

3235 Is there a way that we can adjust the model to change this? Looking at equation (5.2), we see that as long as the rich have higher wages than the poor, the rich will always outbid the poor for the very central location. Worse still, this is true for any description of transportation costs. Sufficiently close to the center, transportation costs are irrelevant no matter their costs. To stop the rich from outbidding the poor for the very central location, we require that rich people consume more land than poor people. In this case, recalling that $\bar{\ell}_r$ and $\bar{\ell}_p$ are land consumption for the rich and poor, if $\bar{\ell}_p$ is enough larger than $\bar{\ell}_r$, then we can reverse the inequality in equation (5.2). This leads to an equilibrium city where poor bus riders occupy the center of the city, rich drivers occupy the outskirts, and there are no rich bus riders. In the context of the simple linear model we are working with here, this seems like an ad hoc fix. In the context of the monocentric city model with housing, it is more palatable, we just require that rich people consume more land than poor people.

3240

The model also fails to account for the gentrification of central cities that began to occur at the end of the 20th century. This is not really a failure of the model, with one

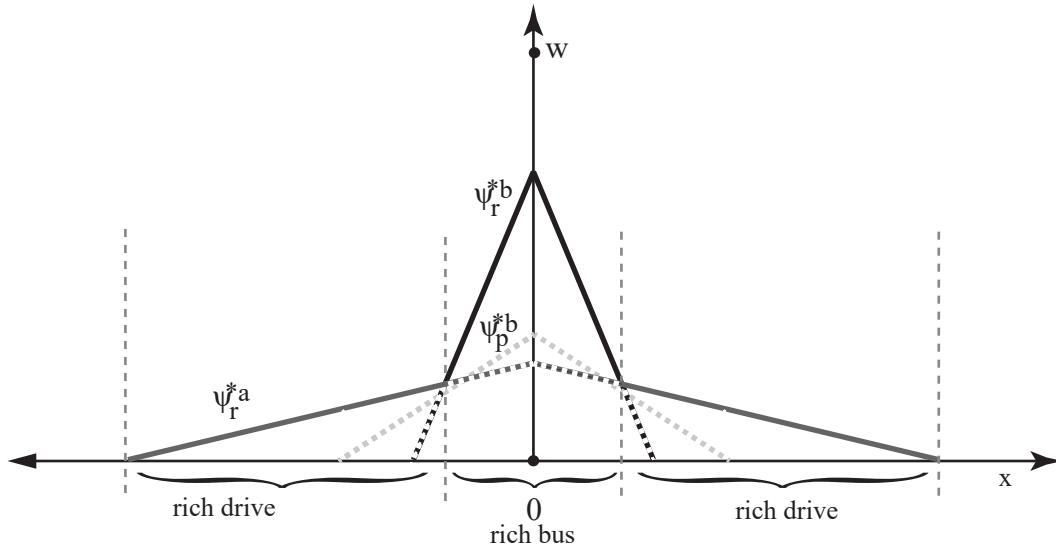
3250 comparative static exercise, we can explain only one comparative static. However, it raises the question of whether we can also explain central city gentrification in the context of this model.

As a simple way to achieve this, consider what happens as the income of the rich increases while the income of the poor stays constant. In this case, we will eventually 3255 violate condition (5.2). Once this happens, poor bus riders are no longer able to outbid rich drivers for locations not occupied by rich bus riders. In this case, the poor are better off moving out of the city to live in the countryside and getting the outside option utility level \bar{u} . Figure 5.9 illustrates this equilibrium. Once the incomes of the rich rise so high that condition (5.2) fails, then in an equilibrium city, rich bus 3260 riders live near the center, rich drivers live on the edge, and the poor are pushed out of the city altogether.

Alternatively, gentrification can also arise if the cost of cars falls far enough that the poor also want to use them. In this case, we will have rich drivers in the center and poor drivers on the edge. As with our other comparative statics exercises, this 3265 demonstrates the ability of the monocentric city model to describe the main features in the history of cities as simple, plausible comparative statics.

Summing up, once we introduce the idea of bid-rent and multiple types of households, if we allow a more realistic description of transportation costs, then the monocentric city model can predict many of the changes to cities that we observe over 3270 the 20th and early 21st century as a consequence of changes in the cost of transportation and the relative income of rich and poor. In particular, the model predicts the increasing relative concentration of central city poor, the rise of the ghetto, as a consequence of the fall in the price of cars. It predicts the subsequent gentrification

Figure 5.9: Equilibrium with poor people displaced from the city



Note: In this equilibrium, the fixed cost of cars is low enough that rich drivers or rich bus riders outbid poor bus drivers everywhere. This results in an equilibrium where the poor are excluded from the city.

of central cities as a consequence of increasing income inequality towards the end of

³²⁷⁵ the 20th century. This aligns closely with the history of cities over the 20th and early 21st centuries, particularly in the US.

With this said, we have seen evidence that ghettos were harmful, and many of the institutional barriers that helped to create them were repugnant. As we have developed it here, the monocentric city model does not reflect these problems. Because ³²⁸⁰ of the free mobility assumption, all households in the model always have the same utility level, \bar{u} , whether they drive in from the suburbs, or take the bus in the center. Similarly, the process of gentrification is not harmful. Poor bus riders excluded from the city have the same payoff as they would in the city. If we would like to think

about the welfare implications of ghettos or of gentrification, then the monocentric
3285 city model, at least as we have developed it here, is probably not up to the task, no
matter how well it does at explaining what happened.

The foundations for worrying about racism are clear. It is intrinsically harmful
to its victims, and as manifested in US cities, it probably constrained black people
3290 to live on “too little land” and to consequently pay “too high” rents to land own-
ers. On the other hand, the foundations for worrying about gentrification seem less
obvious. Particularly because, in the US case, gentrification looks (at least in part)
like a consequence of relaxing the institutional restrictions on black mobility of the
early and middle part of the 20th century. Surely, the central migration of educated
young people in the early 21st century partly reflects the relaxation of institutional
3295 restrictions that were at least partly responsible for the concentration of black people into
central city ghettos were harmful, then their relaxation, and subsequent gentrification
should be the opposite. Because of this, arguments for the harmfulness of gentrifica-
tion typically revolve around quantities that are intrinsically hard to observe and are,
3300 arguably, not capitalized into land prices. For example, the value of lost or degraded
social networks.

5.5 Conclusion

The history of cities over the past 150 or 200 years is one of growth in population
and decentralization. Cities have been getting bigger and spreading out. In the US,
3305 this partly reflected the Great Migration, which led to the creation of black ghettos

and concentration of urban poor. Over the past 20 years, this trend towards the concentration of minorities and poverty has started to reverse as more affluent people move back to city centers, even as cities continue to spread out.

The monocentric city model can explain this basic pattern as a consequence of income heterogeneity, falling prices for transportation, and increasing income inequality.
3310 In particular, the monocentric city model is able to describe and predict segregation by income.

The model does less well predicting segregation by race. This was also an important feature of the development of cities in the US, and partly reflects “white flight”
3315 along with racial policies intended to constrain the residential choices of black people. The monocentric city model also does not provide a basis for concluding that gentrification is a problem. There is no externality. History suggests that patterns of segregation that emerged during the 20th century partly reflected policies intended to confine blacks to central cities. Viewed in this light, gentrification looks like reversion
3320 to the more integrated development that would have occurred otherwise.

Problems

1. In this question we will examine the dissimilarity index of a fictional city.

Assume there are three census tracts in a city, each with population 1. Assume that, initially, the black population of the city is zero. At the end of the period,
3325 the black population increases to $\frac{1}{7}$ of the total city population (i.e., the city is around 14% black), while the non-black population does not change. Answer the following questions about the population distribution at the end of the period.

- (a) What is the total city population and black population?
- (b) Assume that the black population is equally divided among the three
3330 tracts. What is the dissimilarity index?
- (c) Assume that the entire black population is in one census tract. What is
the dissimilarity index in this case?
- (d) Given the above, how should we interpret the rapid increase in the time
series of the dissimilarity index in the left panel of 5.4?
- 3335 (e) Why is it important that the empirical analysis in Cutler et al. (1999)
focuses on the variation in the dissimilarity index across cities, instead of
over time?
2. How much lower is the black than non-black college graduation rate in 1990 in
high segregation cities compared to low segregation cities?
3. In this problem, we will examine bid-rent functions with two types of house-
holds.

3340 Assume there are two types of households, the rich and the poor, whose only difference (for now) is their wage levels, with $w_r > w_p$. Assume also that everyone takes the bus, which has cost

$$(w_r * t^b + c^b)|x| \text{ for the rich type, where } t^b, c^b > 0$$

$$(w_p * t^b + c^b)|x| \text{ for the poor type, where } t^b, c^b > 0$$

- 3345 (a) Set up the household problem from the monocentric city model for each type of household.

- (b) Assume $u(c^*) = \bar{u}$ for both types. For both types, substitute an expression for c^* in terms of \bar{u} into the constraint from the previous part. Call these functions $R_r(x)$ and $R_p(x)$, respectively.
- 3350 (c) Evaluate $R_r(0)$ and $R_p(0)$. Which is larger?
- (d) Evaluate $\frac{\partial R_r(x)}{\partial x}$ and $\frac{\partial R_p(x)}{\partial x}$. Which is steeper?
- (e) Plot $R_r(x)$ and $R_p(x)$ on one graph. Indicate the areas in which each type has the higher willingness to pay. Describe the resulting equilibrium briefly.

3355

Chapter 6

Quantitative Spatial Models and How Railroads Reorganized London

The San Francisco Bay Area is among the most productive and innovative places in
3360 the world. Between 1990 and 2005 it was responsible for nearly 15% of all patents filed in the US, and an even larger share of all venture capital funding [Carlino and Kerr, 2015]. But there are two possible explanations for San Francisco's extraordinary inventiveness. It may be that there is something about the place that is particularly conducive to invention, like the presence of excellent universities. On the other hand,
3365 it may just be that the people who move to San Francisco are especially inventive in some way that is difficult to observe. Either way, understanding the role of individual heterogeneity in inventiveness is key to understanding how the innovation process works, and why it is so concentrated in space.

More generally, a primary focus of almost any empirical investigation of the effect
3370 of almost any sort of treatment on people or households, is to understand whether the assignment of treatment is related to unobserved individual heterogeneity of the people or households. The extraordinary inventiveness of San Francisco is an example of this more general problem.

Unfortunately, the monocentric city model is not good for thinking about heterogeneity.
3375 It can describe a few types of households, e.g., rich and poor, but not much more. In particular, the monocentric city model cannot accommodate household or individual level heterogeneity. This limits its usefulness for thinking about problems like the sorting of inventive people into particular places.

We must also think about welfare differently when households are heterogeneous.
3380 Recalling Chapter 1, with just one or a few types of households, aggregate land rent measures all of the surplus in the monocentric city model. With a continuum of types of households, this is no longer true. Aggregate land rent is a lower bound on the amount of surplus created by a city.

Given the importance of individual heterogeneity, we want to figure out what a
3385 model of cities looks like when there is a continuum of different types of people. This effort will rely on tools borrowed from an old econometric technique called “discrete choice modeling”, repurposed for the study of cities. These models are often called “quantitative spatial models”, or simply QSM, and are an active areas of research in urban economics. They are, however, more complicated than the monocentric city
3390 model.

This complexity has three main advantages. First, it allows quantitative spatial models to describe actual geographies, instead of the uniform lines and planes we’ve

considered so far. Second, it allows for a more realistic description of the differences in peoples' tastes for particular locations, and their differing productivities in particular locations or occupations. Finally, quantitative spatial models can allow people to choose both their location of work and of residence. This is an important generalization of the monocentric city model, where we assume that everyone works in the center. The cost of this complexity is that the model is harder to solve, and results are less general. For example, with the monocentric city model we conclude that rent gradients must be downward sloping. With a QSM we conclude, for example, that rent declined as we traveled away from the center of London in 1920.

To allow for a continuum of household types while keeping the mathematics of the problem tractable, we need to assume that space consists of discrete locations, rather than a continuum. This raises two issues. First, how do we describe distance between a bunch of discrete places? Second, how do we think about the household optimization problem? We'll tackle them in order.

6.1 Distance in discrete space

Given three locations, $i = 1, 2, 3$, there are nine possible origin-destination pairs. If we disregard pairs where the origin and destination are the same (so no travel takes place) and assume that the cost is the same for both directions of travel, then we are down to three origin destination pairs, $ij \in \{12, 23, 31\}$. Denote the travel cost between any two locations as τ_{ij} . The convention is that the first subscript is the index for the origin location, and the second is for the destination.

The standard QSM framework assumes that transportation costs enter the problem

³⁴¹⁵ multiplicatively. This is called “iceberg transportation costs”. The idea is that if you have τ_{ij} units of value in location i and you transport it to location j , some of it melts, and you are left with 1 unit of value in j . This requires that $\tau_{ij} > 1$, unless you are going from location i to itself, i.e., not really traveling at all, and then $\tau_{ii} = 1$. A little more generally, if we send x units from location i to j , then x/τ_{ij} units arrive.

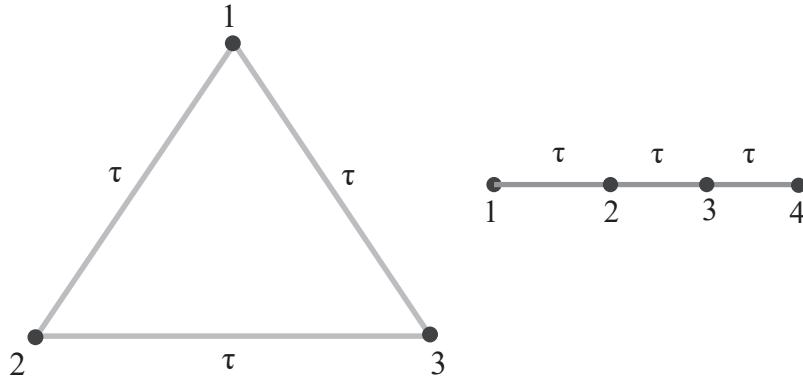
³⁴²⁰ This assumption turns out to be easy to work with because it is multiplicative, and this is why it is a common way of describing transportation costs. With that said, unless you are actually shipping icebergs, there is no reason to think that it is a particularly accurate description of how much it costs to move around. For example, the cost of commuting in a car almost surely involves both a fixed and a variable cost, as in Chapter 5. In some cases, we can also set up QSM models with other formulations of transportation costs, and in some of the examples we work in this chapter, I will still assume that transportation costs operate additively, much as they do in the monocentric city model.

³⁴²⁵ Now consider the case where our three locations are the vertices of an equilateral triangle, and that the cost to travel between any two vertices is the same, as illustrated in the left panel of figure 6.1. This means that we can write a matrix of travel costs for this set of three locations as

$$\begin{bmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{bmatrix} = \begin{bmatrix} 1 & \tau & \tau \\ \tau & 1 & \tau \\ \tau & \tau & 1 \end{bmatrix}.$$

In words, this says that the iceberg cost to travel between any pair of locations is τ , unless the pair of locations are both the same place, and then the iceberg cost is

Figure 6.1: Pairwise travel costs for two simple geographies



Note: *Left panel shows pairwise travel costs between three points located at the vertices of an equilateral triangle. Right panel shows pairwise travel costs between four points located along a line.*

³⁴³⁵ one. More simply, if we ship τ units of a good from one vertex to another, one unit arrives. If we ship one unit of the good from a place to itself, it all arrives.

For our second example, consider four equally spaced locations on a line such that the iceberg cost to travel between any two adjacent points is τ , as illustrated in the right panel of figure 6.1. In this case, if we send τ units from location 1 to 2, then exactly one unit will arrive in location 2. Similarly if we send τ units from 2 to 3. ³⁴⁴⁰ What about from 1 to 3? If we send x units from 1 to 2, then x/τ units arrive. If we then send these goods along to location 3, then they must pay the iceberg cost again, and $(x/\tau)/\tau = x/\tau^2$ units arrive. Similarly, sending goods from 1 to 4 means that x/τ^3 units arrive.

³⁴⁴⁵ For the geography illustrated in the right panel of figure 6.1, there are 16 possible origin-destination pairs. In order to describe transportation costs completely for this geography, we need an iceberg factor for each of them. That is, we need 16 τ 's. We

can write them all in a matrix as follows,

$$\begin{bmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{21} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{31} & \tau_{32} & \tau_{33} & \tau_{34} \\ \tau_{41} & \tau_{42} & \tau_{43} & \tau_{44} \end{bmatrix} = \begin{bmatrix} 1 & \tau & \tau^2 & \tau^3 \\ \tau & 1 & \tau & \tau^2 \\ \tau^2 & \tau & 1 & \tau \\ \tau^3 & \tau^2 & \tau & 1 \end{bmatrix} \quad (6.1)$$

This gives us the cost of every possible trip in the geography.

3450 We can now see how this description of travel costs will work in actual, real world geographies. Suppose you are given a map of a city with N discrete neighborhoods, census tracts, for example. One can describe travel costs in the city by calculating the iceberg travel costs between each of the possible pairs of census tracts. This results in an empirically founded $N \times N$ matrix of iceberg commute costs. In this way, we 3455 can use this method of describing travel costs to describe real cities rather than the stylized examples we've considered up until now.

3460 But there is a catch. Travel from one neighborhood to another is really travel from a point in one neighborhood, usually its center, to a point in another neighborhood. If, as will usually be the case. people are not traveling precisely between these two points, then turning a map into a set of discrete points insures that we will measure travel distances only approximately. The monocentric city model of Chapter 1 can measure travel distances precisely on a stylized geography. In models with discrete space, we can measure distance approximately in a geography based on a real place.

3465 For our final example, consider a case as much like the monocentric city of Chapter 1 as possible, a linear geography with three residence locations at $x = 1, 2, 3$ and a work location at $x = 0$. Everyone travels to work and back. Unlike the examples

above, and most QSM models, transportation costs enter the budget additively, not as multiplicative iceberg costs.

Just like the monocentric city of Chapter 1, everyone commutes from their residential location to the center, where everyone earns the same wage w . Everyone lives on a parcel of the same size, $\bar{\ell}$, at whichever of the three locations they choose. Wages are spent entirely on a composite consumption good, c , on rent for the household parcel at location i , $R_i\bar{\ell}$, and on commuting. Also like the monocentric city model, commuting costs are linear in distance from the center. Putting this all together, the budget for a household at each of the locations is

$$w = c^* + R_1\bar{\ell} + t \quad (6.2)$$

$$w = c^* + R_2\bar{\ell} + 2t$$

$$w = c^* + R_3\bar{\ell} + 3t$$

at locations 1, 2, and 3, respectively.

The (additive, not iceberg) transportation cost matrix for this city will look like,

$$\begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} & \tau_{03} \\ \tau_{10} & \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{20} & \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{30} & \tau_{31} & \tau_{32} & \tau_{33} \end{bmatrix} = \begin{bmatrix} . & t/2 & 2t/2 & 3t/2 \\ t/2 & . & . & . \\ 2t/2 & . & . & . \\ 3t/2 & . & . & . \end{bmatrix}$$

The first index on τ_{ij} is work and the second is home. Each person travels to work and then back, so the cost of each leg is half the total cost of commuting that appears in the budget. Even though there are 16 possible trips, I've only filled in the parts

of the travel cost matrix that describe trips that actually happen in the monocentric city model, those from home to work and back.

6.2 The discrete choice problem

Suppose there are two locations, $i = 1, 2$, and a single household that obtains utility

3485 V_1 and V_2 from the two locations.

This household's discrete choice problem is,

$$\max \{V_1, V_2\}.$$

This is easy. Choose your favorite. This replaces the household problem we solved with continuous space in Chapter 1. That problem requires that a household choose the best place from among a continuum of choices. Here the household just chooses
3490 between two.

If household makes a choice of housing and consumption in each location, as in the monocentric city model with housing from Chapter 3, then a household's problem is,

$$\max \{\max_{c_1, h_1} u(c_1, h_1), \max_{c_2, h_2} u(c_2, h_2)\}.$$

We imagine that households solve this problem backwards. That is, households consider the best possible choice of consumption and housing, conditional on having already chosen a location, and then ask what utility this would yield in each location. Households then choose the best of these two discrete options.
3495

The logic of this problem is the same if households choose among many discrete alternatives (the case we're really interested in) rather than two, there is just more
3500 notation. For example, we would write the problem of a household choosing between N alternatives as,

$$\max \{V_1, V_2, \dots, V_N\}.$$

Now suppose there are many households instead of just one, and index them by ν (“nu”). Formally, we want a continuum, or in the jargon, “a measure” of households.
3505 For practical purposes, this arcane term means a “length” of households, and it is a way to finesse one of the stranger properties of real numbers.¹

Denote the payoff for a household of type ν at location i by $V_i(\nu)$. Suppose that for each location, each $V_i(\nu)$ has a systematic or common component, u_i , and an idiosyncratic component that is particular to each household, $\varepsilon_i(\nu)$, with $V_1(\nu) = u_1\varepsilon_1(\nu)$ and $V_2(\nu) = u_2\varepsilon_2(\nu)$ for all households ν . Thus, each type of household, ν , is associated with its own list of location specific idiosyncratic values, one for each location. Following the literature, I'll often refer to $\varepsilon_i(\nu)$ as a “taste shock”. The notation usually used for these models does not explicitly indicate that the different draws of $\varepsilon_1(\nu)$ and $\varepsilon_2(\nu)$ are particular to household type ν , and abbreviates to $V_1(\nu) = u_1\varepsilon_1$ and $V_2(\nu) = u_2\varepsilon_2$. This is easier to write, and I'll do it from now on,

¹Consider that if I multiply every element of the set $[0, 1]$ by two, I get the set $[0, 2]$. Conversely, if I multiply every element of the set $[0, 2]$ by one half I get the set $[0, 1]$. That is, every number in the interval $[0, 1]$ can be matched to a number in the interval $[0, 2]$ and every number in the interval $[0, 2]$ can be matched to a number in the interval $[0, 1]$. In this sense, there are the same number of numbers in both intervals, even though one interval is twice as long as the other. In practice, this means that the real numbers are not good for counting things or, as in our case, enumerating a list of households. On the other hand, the lengths of the two intervals are clearly different, and so if we want to think about lengths, or quantities of households we don't have a problem.

3515 but it means that you need to remember that the ε 's are different from household to household.

Let's now go back to the two location example I started with. Suppose that the idiosyncratic part of household preferences takes one of the values one, two, or three at random. That is, $\varepsilon_i \in \{1, 2, 3\}$ with each value equally likely. In this case, each 3520 possible draw of two ε 's, $(\varepsilon_1, \varepsilon_2)$, corresponds to a type ν , and so there are nine different types of households, each making up 1/9th of the population.² Again, each type ν corresponds to a pair $(\varepsilon_1, \varepsilon_2)$, with each ε_i drawn from the set $\{1, 2, 3\}$

Next, suppose that the parts of the payoff that are common to all types are $u_1 = 1$ and $u_2 = 1.1$. Then $V_1(\nu) \in \{1 \times 1, 1 \times 2, 1 \times 3\} = \{1, 2, 3\}$ and $V_2(\nu) \in \{1.1, 2.2, 3.3\}$. 3525 Putting this all together we have that each type of household has valuations for the two locations, $(V_1(\nu), V_2(\nu))$, where $V_1(\nu)$ and $V_2(\nu)$ each take one of these three values. Because $u_1 > u_2$ location 2 is “better” than location 1 in the sense that given equal idiosyncratic values for the two locations, households always choose location 2. Notice, too, that this is a closed city model. All households have to end up somewhere 3530 in the city. There is no option to move to an outside option as there is in the open monocentric city.

This model of location choice is like the monocentric city model in an important way. In the monocentric city model, equilibrium is defined to occur when all locations give the same payoff, and so there is no incentive for a household to move. This means 3535 that, trivially, all households are choosing their favorite location. That is just the same

²Notice that there is a trick here. If the different values of ε_i were really equally likely, then there should be some randomness in the shares of households of each type that we see in any given sample. In general, it would not be exactly 1/9th for each type. It turns out that this sort of variation goes away if you make enough draws, or in our case, if you have enough households. This is why we want to work with a continuum of households rather than a finite number.

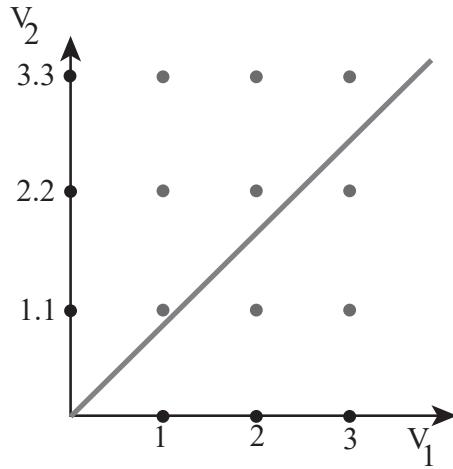
as the discrete choice rule we use here, choose your favorite.

The discrete choice model is also different from the older model in important ways. First, in the monocentric city model all households choose their favorite location, but unlike in the monocentric city model, households in the discrete choice model generally 3540 strictly prefer their chosen location to the others. In the monocentric city model all households are exactly indifferent between all locations. Second, in the monocentric city model, all households get exactly the same payoff. This is not true in the discrete choice model. In our example above, the household types that draw $\varepsilon_1 = 3$ or $\varepsilon_2 = 3$ are lucky. They have a higher payoff than the other types. The type that draws both 3545 $\varepsilon_1 = 1$ and $\varepsilon_2 = 1$ is unlucky. These households have a low payoff wherever they locate. Third, because all households in a location face the same price of real estate, the lucky households will end up paying less than the location is worth to them, while the unlucky households may be stuck paying more (because we're not letting them move out of town). This means that there may be consumer surplus in the real estate market. This means, in turn, that aggregate land rent does not measure all of the 3550 surplus as it does in the monocentric city model. We will return to this later.

To figure out where people locate, we need to figure out the share of households choosing locations 1 and 2. That is, the share s_1 with $V_1(\nu) = \max \{V_1(\nu), V_2(\nu)\}$ and the share s_2 with $V_2(\nu) = \max \{V_1(\nu), V_2(\nu)\}$. Notice that as long as housing 3555 markets clear and everyone ends up living somewhere, then we must have $s_1 + s_2 = 1$, and so if we can figure out one of the two shares, then we know everyone's choices.

To solve this problem, consider the graph shown in figure 6.2. The pair of payoffs for each of the nine types is a coordinate in $(V_1(\nu), V_2(\nu))$ space. We want the share of households for which location 2 gives a higher payoff than location 1, $V_2(\nu) > V_1(\nu)$.

Figure 6.2: A simple discrete choice Problem



Note: The figure illustrates a simple discrete choice problem. Payoffs in location 1 are on the x-axis. Payoffs in location 2 are on the y-axis. The gray line is 45 degrees. Above this line and location 2 has a higher payoff than location 1, and conversely. Thus, any household type whose payoffs for the two locations lies above the 45 degree line prefers location 2 and otherwise prefers location 1. If the shares of households with each pair of payoffs are all the same then 6/9th's of households choose location 2.

- ³⁵⁶⁰ In the figure, the gray 45 degree $V_2 = V_1$ line divides the plane in two. Below the line, $V_1(\nu) < V_2(\nu)$, and above the line $V_2(\nu) > V_1(\nu)$. To calculate the share of households that prefer location 2 to location 1, we need only calculate the share of types with a payoff pair that lies above the 45 degree line. Because each of the nine possible types of household is equally likely, this means 6 of 9, or 2/3 of households choose location 2 and 1/3 choose location 1.

We've stated this outcome in terms of a toy example where each household can have only three idiosyncratic values for each location. Nothing prevents us from allowing more types of households or from allowing the shares of households of each

type to be different, and we will need to allow this if we are to describe more realistic
 3570 geographies. This requires more general notation for thinking about the shares of different types.

To start, let $f(\varepsilon_1, \varepsilon_2) = 1/9$ denote the share of households of each type. Then we can write the share of the households with types left of the gray line in figure 6.2 as,

$$\begin{aligned} s_2 &= \sum_{\ell=1}^3 f(\varepsilon_1 = 1, \varepsilon_2 = \ell) + \sum_{\ell=2}^3 f(\varepsilon_1 = 2, \varepsilon_2 = \ell) + \sum_{\ell=3}^3 f(\varepsilon_1 = 3, \varepsilon_2 = \ell) \\ &= \sum_{k=1}^3 \left[\sum_{\ell=k}^3 f(\varepsilon_1 = k, \varepsilon_2 = \ell) \right] \end{aligned} \quad (6.3)$$

The second line is just a compact rewrite of the first. Using $f(\varepsilon_1, \varepsilon_2) = 1/9$ equation
 3575 (6.3) simplifies to,

$$\sum_{k=1}^3 \left[\sum_{\ell=k}^3 1/9 \right].$$

The indexing of ε 's with l and k is actually indexing pairs of ε 's, which means that we are really just indexing types. So, what we have done, is to write down a general way of summing over the types that choose location 2.

For our problem, with three types and two locations, all of this notation makes
 3580 our easy problem more obscure (unless you are really practiced at reading this sort of notation). But, our example involves just two locations and only three taste shocks per location. If we want to write down more complicated and realistic problems, we can use this notation without much modification.

To see this, generalize our example with two locations, ($i = 1, 2$) and three taste
 3585 shocks per location, to one with two locations and N taste shocks for each. As before,

each $V_i(\nu)$ has a systematic or common component, u_i , with $u_1 = 1$ and $u_2 = 1.1$. Each type, ν , gets a draw of one of N possible idiosyncratic taste parameters for each location i , with $\varepsilon_i \in \{1, \dots, N\}$ and $V_i(\nu) = \varepsilon_i u_i$. Let $f(\varepsilon_1, \varepsilon_2)$ be the share of households with each pair of taste parameters. If we were to generalize figure 6.2 to this problem, it would look just the same, but would involve an $N \times N$ grid of $(V_1(\nu), V_2(\nu))$ pairs rather than 3×3 .

In this case, the share of households choosing location 2 is,

$$s_2 = \sum_{k=1}^N \left[\sum_{\ell=k}^N f(\varepsilon_1 = k, \varepsilon_2 = \ell) \right]. \quad (6.4)$$

This is exactly the same calculation that we did in equation (6.3). We are calculating the share of households choosing location 2, but summing the shares of households of each type such that $V_1(\nu) > V_2(\nu)$. Here we see the value of the notation we introduced in equation (6.3). We calculate the share of households choosing location two in the more complicated case by just changing the limits of summation from 3 to N .

We can also generalize this example to allow for a continuum of possible valuations for each household for each location and a continuum of possible taste shocks for each location. In this case, the ε_i has a continuum of values in each location and we need to use integration rather than summation,

$$s_2 = \int_0^\infty \int_{\varepsilon_1}^\infty f(\varepsilon_1, \varepsilon_2) \partial \varepsilon_2 \partial \varepsilon_1. \quad (6.5)$$

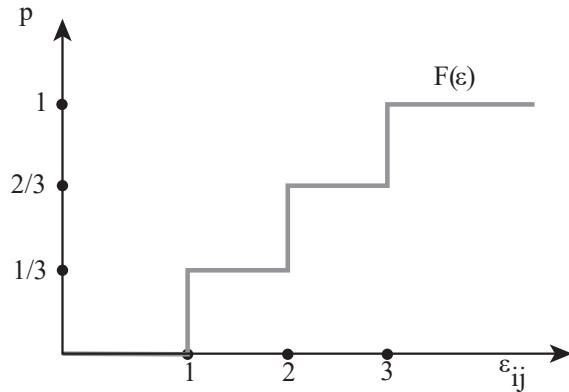
Equation (6.5) is just the same as equation (6.4), except that we've switched out integration for summation.

3605 Equations (6.4) and (6.5) let us calculate the share of households choosing location 2 when households can have N and a continuum of taste shocks for each location. This allows a rich description of household heterogeneity, but households still just choose between two locations,. This means we are still restricted to a geography that is probably too simple to be much use. What happens if we consider more locations?

3610 Start with the case where there are three locations rather than two. In this case, we need to think about generalizing the intuition we get from figure 6.2 to allow for a third payoff. We would still need the x and y -axes to describe payoffs from locations 1 and 2, but we would also require a z axis, pointing out of the page, to describe the payoff in the third location. The 45 degree line in the figure would then become 3615 a plane, separating triples of payoffs where $V_2 > V_1$ from those where the opposite condition holds. A household would choose location 2 only $V_2 > V_1$ and $V_2 > V_3$, so we would need a second 45 degree plane separating triples of payoffs where $V_2 > V_3$ from those where $V_2 < V_3$. The set of types that chose location 2 is then the set that lies above both of these planes.

3620 This is hard to visualize with three choices. Once we go to a geography with four or more locations, it's basically impossible. The formal mathematical description of the problem, the N location generalization of equations (6.4) or (6.5) is a correspondingly complicated nested sum or integral. Worse still, not only are these problems hard to write down, they are hard to evaluate. They are hard to solve analytically, that is to 3625 get an exact solution with pen and paper, and they are also hard to solve numerically using a computer.

Figure 6.3: A simple probability distribution



Note: This figure illustrates the probability distribution function given in equation (6.6). This distribution describes the distribution of draws such that the outcomes, 1, 2, and 3 all make up 1/3 of the population.

6.3 Extreme value distributions

But there is a special case. If the ε_i follow an “extreme value distribution”, then the shares of households choosing each outcome has an easy analytic solution. To understand what an extreme value distribution is, we have to first understand what a distribution is.

In the two location and three type example of the last section, ε_i takes the values 1,2,3 with equal probability. Another way to represent this is with a probability distribution function that reports the share of realizations of ε_i that are less than any given value. In the jargon, this function is a “Cumulative Distribution function

(CDF)” or “Probability Distribution Function” for ε . We can write it as

$$F(\varepsilon) = \begin{cases} 0 & \text{if } \varepsilon < 1 \\ 1/3 & \text{if } 1 \leq \varepsilon < 2 \\ 2/3 & \text{if } 2 \leq \varepsilon < 3 \\ 1 & \text{if } 3 \leq \varepsilon \end{cases} \quad (6.6)$$

$F(\varepsilon)$ satisfies $Prob(\varepsilon < \bar{\varepsilon}) = F(\bar{\varepsilon})$ or equivalently, the share of households with $\varepsilon < \bar{\varepsilon}$ is $F(\bar{\varepsilon})$. Figure 6.3 illustrates the distribution in equation (6.6). If you’ve studied statistics or econometrics, then this should be familiar.

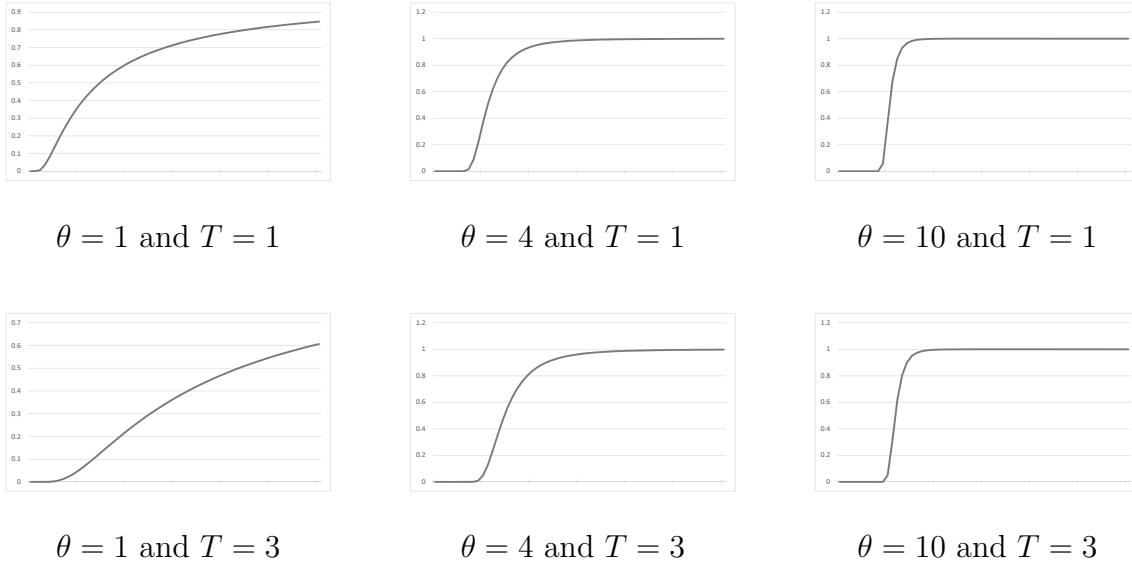
3640 We are interested in two particular distributions that have the special property that the otherwise hard integration problem required in order to figure out which location a household prefers from many alternatives is easy. These probability distributions are known as extreme value distributions, and they come in two main types, “Gumbel” and “Frechet”. For our purposes, they behave similarly and so I will just
3645 talk about the Frechet distribution.

If ε is determined by a Frechet distribution, then its Cumulative Distribution function is,

$$F(\varepsilon) = e^{-T\varepsilon^{-\theta}} \text{ for } T > 0, \theta > 1.$$

That is, $Prob(\varepsilon < \bar{\varepsilon}) \equiv F(\varepsilon) = e^{-T\bar{\varepsilon}^{-\theta}}$. This distribution is governed by two parameters, T , “level”, and θ , “dispersion”. These names are suggestive of “mean” and
3650 “variance” and are often used in the same spirit. Just as figure 6.3 illustrates the distribution described by equation (6.6), figure 6.4 illustrates the Frechet density for

Figure 6.4: Frechet probability distributions



Note: *Frechet probability distributions as the level and dispersion parameters, T and θ , vary. Notice that as θ gets larger, moving from left to right, the Frechet distribution looks more and more like a step function.*

a few values of T and θ .

Looking at figure 6.4, we see that as θ increases, F converges to a step function. This means that all households have the same ε and there is no heterogeneity. To see this, consider a probability distribution with the property that $F(\frac{1}{2}) = 0$ and $F(\frac{1}{2} + \varepsilon) = 1$. In words, this requires that the share of draws from this distribution that are less than $1/2$ is zero, and the share of draws that are less than $\frac{1}{2} + \varepsilon$ is one. It follows that all of the draws must fall in the interval $[\frac{1}{2}, \frac{1}{2} + \varepsilon]$. As we shrink ε toward zero, this is a function with a step at $1/2$. And so, as a distribution looks more like a step function, the outcomes start to concentrate closer to the location of the step. This means less heterogeneity, or dispersion, of the ε 's.

This discussion is obviously arcane, but has an important practical implication. I motivated this chapter by talking about the importance of individual heterogeneity for understanding the way people make decisions, particularly over their locations.
3665 But how *much* heterogeneity matters. Few people care much which shade of beige bathroom tile is in their home, but people may care a lot whether their home is near Yellowstone National Park, or Central Park. In the first case, it will be hard for people to find much to disagree about, but in the second, people may have different tastes. The dispersion parameter in the Frechet distribution gives us a way to tune
3670 how much preference heterogeneity there is in the population of decision makers to fit what we see in the data. The object of Quantitative Spatial Models is to construct models of cities that are able to describe real geographies and real people. To do this, we need this sort of flexibility in our model. Adjusting the dispersion parameter in a discrete choice model helps us do this.

3675 More fundamentally, in order to develop models of many locations with heterogeneous locations, we need to be able to solve the discrete choice problem when there are many choices. Simple examples, like the one we worked through in the last section, we can solve with elementary calculations. But when we consider the problem of discrete choice over many locations, the problem is hard to state in general, and
3680 harder to solve.

By assuming that household taste shocks in each location have a Frechet distribution, we can cut this knot. The following theorem shows how. The proof uses probability theory and vector calculus a bit beyond what is required for the rest of the book, and it is left for an appendix to this chapter.

Theorem. Discrete Choice Theorem: Suppose that households choose among N discrete locations, that for each location $i = 1, \dots, N$, household ν receives payoff $V_i(\nu) = \varepsilon_i u_i$, and ε_i is drawn from a Frechet distribution, $F(\varepsilon) = e^{-T\varepsilon^{-\theta}}$. Then the share of households such that

$$V_i(\nu) = \max \{V_1(\nu), V_2(\nu), \dots, V_N(\nu)\}$$

3690 is

$$s_i = \frac{u_i^\theta}{\sum_{k=1}^N u_k^\theta}. \quad (6.7)$$

Moreover, the average utility of a household making this choice is

$$E(V_i(\nu)) = \Gamma\left(\frac{\theta-1}{\theta}\right) \left(\sum_{i \in \{1,2,3\}} u_i^\theta\right)^{1/\theta}, \quad (6.8)$$

for all locations i .

Equation (6.7) is the punchline. It says that if we know the u_i 's, the parts of payoffs that are common across people, and the dispersion parameter, θ , then we can calculate the share of households making a particular choice. We'll come to the second part of the theorem, equation (6.8), a bit later.

Equation (6.7) requires two further comments. First, equation (6.7) is solving exactly the same problem as equations (6.4) and (6.5), but with a continuum of types drawn from a Frechet distribution and an arbitrary finite number of locations, N . Second, even though equation (6.7) looks complicated, once you learn to parse the

3700

notation, it is pretty straightforward to evaluate. We will work an example below.

This theorem is a big step towards a model of cities with a continuum of types of households.

6.4 A discrete linear city with heterogeneous households

3705

We have worked out how to describe distance in a discrete geography, and we have worked out a tractable way to solve discrete choice problems. I want to put these pieces together to construct a version of the monocentric city model with heterogeneous households.

3710 Consider a discrete linear city with three neighborhoods $i \in \{1, 2, 3\}$. Let x_i denote a neighborhood's distance from the CBD, with $x_1 = 1$, $x_2 = 2$, $x_3 = 3$. The cost to commute one unit distance is τ . The city is populated by a continuum of households indexed by ν . Each household chooses a neighborhood i , pays land rent R_i , and commutes to the center, at location 0, to earn wage w . Household ν 's utility 3715 in location i is $V_i(\nu) = \varepsilon_i c_i$ where c_i is consumption and ε_i is a household specific preference for location i . All ε_i 's are drawn from a Frechet distribution, $F(\varepsilon) = e^{-T\varepsilon^{-\theta}}$.

A household with a big draw for ε_i gets more utility from consumption in location i , and in this sense has an idiosyncratic taste for living in location i . Thus, this model allows for a population of households, each of which has their own particular predilection for each of the three locations. This is about as general a description of heterogeneity as you could ask for.

3720 Consider again the example that motivated this chapter. While this particular

problem is not set up to think about people sorting into neighborhoods or cities on the basis of their inventiveness, hopefully, it is clear that this is not too big a step
3725 from what we have done. Thus, we are pretty close to a description of a spatial model in which the description of individual level heterogeneity is flexible enough to form a basis for empirical research. This is really neat. The only catch is that we have to describe individual heterogeneity with an extreme value distribution like the Frechet distribution.

3730 Given their draws of $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$, each household makes the discrete choice of location,

$$\max \{V_1(\nu), V_2(\nu), V_3(\nu)\},$$

or, using the definition of $V_i(\nu)$,

$$\max \{\varepsilon_1 c_1, \varepsilon_2 c_2, \varepsilon_3 c_3\}. \quad (6.9)$$

There are going to be lucky and unlucky households. The unlucky households draw three small ε 's and are unhappy where ever they go. The lucky households draw at
3735 least one big ε .

A household budget is $w = c_i + R_i + i\tau$, so $c_i = w - R_i - i\tau$. Substituting into (6.9), households face the discrete choice,

$$\max \{\varepsilon_1[w - R_1 - 1\tau], \varepsilon_2[w - R_2 - 2\tau], \varepsilon_3[w - R_2 - 2\tau]\}. \quad (6.10)$$

This is a closed city model. Households must choose one of the three locations in

the city. To convert it to an open city model, we can specify an outside option, call it location zero, and imagine that all households get payoff $\varepsilon_0 \bar{u}$ in this location. In this case, the open city version of the household discrete choice problem is,³⁷⁴⁰

$$\max \{\varepsilon_0 \bar{u}, \varepsilon_1[w - R_1 - 1\tau], \varepsilon_2[w - R_2 - 2\tau], \varepsilon_3[w - R_3 - 3\tau]\}. \quad (6.11)$$

Because equation (6.10) is a little simpler than equation (6.11) I will concentrate on solving the closed city case. Solving the open city case is similar, but involves a little more notation and algebra.

³⁷⁴⁵ Applying the Discrete Choice Theorem to equation (6.10), we calculate the shares of households choosing each location,

$$\begin{aligned} s_1 &= \frac{c_1^\theta}{\sum_{k=1}^3 c_k^\theta} = \frac{[w - R_1 - 1\tau]^\theta}{\sum_{k=1}^3 [w - R_k - k\tau]^\theta} \\ s_2 &= \frac{c_2^\theta}{\sum_{k=1}^3 c_k^\theta} = \frac{[w - R_2 - 2\tau]^\theta}{\sum_{k=1}^3 [w - R_k - k\tau]^\theta} \\ s_3 &= \frac{c_3^\theta}{\sum_{k=1}^3 c_k^\theta} = \frac{[w - R_3 - 3\tau]^\theta}{\sum_{k=1}^3 [w - R_k - k\tau]^\theta}. \end{aligned} \quad (6.12)$$

Now what? The system of equations (6.12) is three equations in 8 unknowns $\{s_1, s_2, s_3, R_1, R_2, R_3, \theta, \tau\}$, so we can't solve them without more information.

³⁷⁵⁰ In the spirit of the continuous space monocentric city model, suppose that each location is occupied by exactly one third of the population, so that $s_i = 1/3$, and the

R_i are not observed. Then,

$$\begin{aligned}\frac{1}{3} &= \frac{[w - R_1 - 1\tau]^\theta}{\sum_{k=1}^3 [w - R_k - k\tau]^\theta} \\ \frac{1}{3} &= \frac{[w - R_2 - 2\tau]^\theta}{\sum_{k=1}^3 [w - R_k - k\tau]^\theta} \\ \frac{1}{3} &= \frac{[w - R_3 - 3\tau]^\theta}{\sum_{k=1}^3 [w - R_k - k\tau]^\theta}.\end{aligned}$$

Because the denominators are all the same, the numerators must be, too. This means that,

$$[w - R_1 - 1\tau] = [w - R_2 - 2\tau] = [w - R_3 - 3\tau].$$

In turn, this implies that $R_1 - R_2 = \tau$ and $R_2 - R_3 = \tau$. In words, the land rent gradient decreases at the same rate as commute costs increase, just like the continuous version of the model.

Suppose we also require that land rent at $x = 3$ be equal to the observed agricultural land rent, \bar{R} , and that τ is known. Then we have $(R_1, R_2, R_3) = (\bar{R} + 2\tau, \bar{R} + \tau, \bar{R})$, and we can solve for consumption at each location. For this example, consumption is the same everywhere, $c_i = w - \bar{R} - 3\tau$, just as in the monocentric city model with homogeneous households.

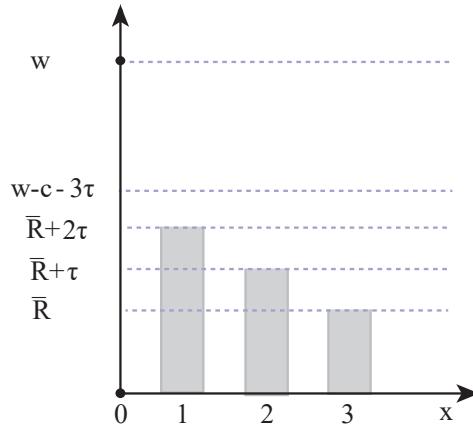
Figure 6.5 illustrates the resulting land rent gradient. It's useful to compare this with the land rent gradient for the monocentric city with homogeneous households illustrated in figure 1.8. Three differences are obvious. First, for the discrete city, we've only described the half of the city that lies to the right of the CBD. This is easy to fix, we just need to extend the analysis of the discrete city to six locations, three to

the right of zero, as now, and three to the left that are symmetric to the ones on the right. This means six locations instead of three, and correspondingly more notation and algebra, but it is otherwise an almost identical problem. Second, the discrete
3770 model consists of discrete locations and has a discrete price gradient, while in the model with identical households, everything is continuous. This is by construction. Third, and more substantively, by inspection of figure 6.5, the common consumption level is $w - \bar{R} - 3\tau$. This is greater than $w - R_i$ for all locations. In the model with homogeneous households, $c^* = w - R(0)$, for location $x = 0$. This is a consequence
3775 of the discrete geography. In our discrete geography, no one lives at the CBD. The closest residential location is at $x = 1$. This means that no one pays zero commuting costs. In the continuous model, however, there are households living right at $x = 0$, and their commute cost is exactly zero.

One of the advantages of QSM models is that they are flexible enough to describe
3780 real world cities, and much of the current research activity in urban economics revolves around more complicated versions of these models that describe actual places (we'll look at an example shortly). With this in mind, it's worth pointing out that there are lots of combinations of data that will allow us to solve for equilibrium in our example with three locations.

3785 You can see this in the example we have been working. If I observed different shares, or if the locations were not evenly spaced, I could still have solved the problem.

Figure 6.5: Land rent in (half) a discrete monocentric city



Note: *The land rent gradient for a closed discrete monocentric city must equal the agricultural land rent, \bar{R} , at the most remote urban location. When each location houses an equal share of the population, then land rent increases as distance to the center decreases in order to keep income net of rent and commuting constant.*

6.5 Welfare

In the monocentric city model with homogeneous households, each household is indifferent between all the occupied locations, and all households get the same payoff. Not so here. In the monocentric city model with heterogeneous households, a household will almost always strictly prefer the location they choose to the locations they don't, and lucky households will end up with higher payoffs than unlucky households. All of this heterogeneity means we need changes in the way we think about welfare. It's not enough just to think about aggregate land rent.

Calculating welfare when households are heterogeneous is more difficult than when they are not. To see why, recall that there is just one price for land in each location. This means that households that get particularly good draws of ε will sometimes

end up paying less to live in their preferred location than is required to make them indifferent between their top two choices. The dollar value of this difference, summed over all households, is the benefit that accrues to households because they participate in the real estate market of the city that is *not* reflected in land rent. This difference is consumer surplus.³

If we want to think about welfare in models with heterogeneous households, we need to calculate consumer surplus, in addition to aggregate land rent. But how to calculate consumer surplus? First, let's define it precisely. Consumer surplus is the benefit a household gets from living in a location, after they have paid land rent and commute costs. We can calculate this difference for an average household as,

$$E(V) = E(\max \{\varepsilon_1[w - R_1 - \tau], \varepsilon_2[w - R_2 - 2\tau], \varepsilon_3[w - R_3 - 3\tau]\}). \quad (6.13)$$

This equation describes the average over all households of the post-rent and post-commute consumption in each household's favorite location.

Now we have to figure out how to evaluate this. This turns out to be difficult in general, but when the ε 's are drawn from a Frechet distribution, there is an easy to evaluate, but complicated to write down, analytical solution. This is given in the second part of the Discrete Choice Theorem, equation (6.8).

³Consumer surplus is a conventional measure welfare and is defined as the area between the price line and the demand curve in a standard supply and demand diagram. You will find it discussed at length in any textbook on microeconomics.

Using this result we have,

$$\begin{aligned} E(V) &= E \left(\max_{i \in \{1,2,3\}} \{[w - R_i - i\tau]\varepsilon_i\} \right) \\ &= \Gamma \left(\frac{\theta - 1}{\theta} \right) \left(\sum_{i \in \{1,2,3\}} [w - R_i - i\tau]^\theta \right)^{1/\theta}. \end{aligned}$$

3815 This expression looks worse than it is. The “Gamma function”, $\Gamma \left(\frac{\theta-1}{\theta} \right)$, is a generalization of the factorial operator “!” to the real numbers; $\Gamma(n) = n!$ for counting number n . It is one of those functions that, like trig functions, you usually just have to look up. For example, if $\theta = 4$ then $\Gamma \left(\frac{\theta-1}{\theta} \right) = \Gamma \left(\frac{3}{4} \right)$, and I can look this up on the internet to find 1.22. The rest of the expression is pretty straightforward.

3820 Note that $E(V)$ is just consumers’ surplus. It does not account for land rent. Figuring out how to aggregate land rent and consumer’s surplus to think about welfare in QSM models is an open question. The most common approach is to find some reason to ignore the contribution of land rent to welfare, and then treat consumer’s surplus as the only component of welfare. A better approach, worked out in Fajgelbaum et al. 3825 [2023], is to assume that households all own a share of all the land in the city, and that they collect a corresponding share of aggregate rents. In this case, the landlords are present in the model rather than being “absentee”. The problem is that this is fantastically complicated and we still need to worry about whether we are returning rent to households in a way that reflects actual ownership patterns.

3830 This is probably a good time to point out a strange feature of discrete choice models. Equation (6.13) describes the average payoff for a household in the three location city. Let’s consider what happens if we keep everything about the city exactly the same, except that we divide each of the two locations in half. Thus,

location 1 becomes locations 1a and 1b, and so on. In this case, our household now
 3835 draws six taste shocks and must choose between six locations. Their average payoff will look like this,

$$E(V) = E(\max \{\varepsilon_{1a}[w - R_1 - \tau], \varepsilon_{1b}[w - R_1 - \tau], \varepsilon_{2a}[w - R_2 - 2\tau], \\ \varepsilon_{2b}[w - R_2 - 2\tau], \varepsilon_{3a}[w - R_3 - 3\tau], \varepsilon_{3b}[w - R_3 - 3\tau]\}),$$

where ε_{ia} and ε_{ib} are the taste parameter draws for the two halves of location i .

Compare this expression with the corresponding expression where we leave each of the three spatial units intact, equation (6.13). Which would you rather have?

3840 Recall that the way a household gets a really good outcome in this model is to draw a high ε for at least one location. When we divide all the locations in half, we are giving the households twice as many chances to do this. On average this has to make them better off. With twice as many chances, they are twice as likely to get lucky. But this means that our evaluation of welfare is sensitive the size of the spatial
 3845 units we study, even if nothing changes in the real world.

To solve this problem, we need to reduce the variation in the ε 's as we increase the number of units in order to keep the distribution of largest draws about equal. Less formally, suppose we are considering a geography of 3 units and the parameters of the Frechet distribution of taste shocks are such that one household in 10 gets a
 3850 draw above 100. If we divide each spatial unit in half, each household gets twice as many draws and is (about) twice as likely to get at least one draw above 100. Simply redrawing the spatial units that we use as the basis for analysis so that they are smaller changes the economics of the model. To compensate, we need to change

the parameters that govern the distribution of the ε 's so that the chances that a
 3855 household gets a draw greater than 100 is smaller. In this sense, the distribution of the ε 's is not really a feature of population as I have described it so far. It is really a joint feature of the population and the geography.

6.6 A discrete city with heterogeneous households and iceberg commute costs

3860 So far, I've introduced the machinery required to treat heterogeneous households and a discrete geography in a way that preserves as much of the original monocentric city model as possible. Hopefully, this makes things a little bit easier to figure out, but it is not how QSM models are most commonly used. When these models are used to describe actual geographies, as they typically are, then they usually rely on iceberg
 3865 commuting costs and they allow the choice of both residence and workplace. In the next two sections, we develop this more conventional set-up.

To start, consider a discrete linear city like the one just above, but with iceberg commuting costs. There are three neighborhoods $i \in \{1, 2, 3\}$. x_i denotes a neighborhood's distance from the CBD, with $x_1 = 1$, $x_2 = 2$, $x_3 = 3$. Commute costs are
 3870 iceberg with iceberg factor τ per unit distance. This means that after commute wages at locations 1,2,3 are w/τ , w/τ^2 , w/τ^3 . Rather than paying a fixed amount per unit distance, some of a household's work time melts away, at the rate $1/\tau$ per unit of distance.

With iceberg commute costs, a household's budget at each of the three locations

3875 becomes,

$$c_1 = w/\tau - R_1$$

$$c_2 = w/\tau^2 - R_2$$

$$c_3 = w/\tau^3 - R_3.$$

These equations are the analog of equations (6.12) for a city with iceberg commute costs.

As before, a household's utility is $V_i(\nu) = \varepsilon_i c_i$ where c_i is consumption and ε_i is a household and location specific taste shock. All ε_i are drawn from a Frechet distribution, $F(\varepsilon) = e^{-T\varepsilon^{-\theta}}$. Putting this all together,

$$V_i(\nu) = \varepsilon_i [w/\tau^i - R_i],$$

and households make the discrete choice

$$\max \{\varepsilon_1[w/\tau^1 - R_1], \varepsilon_2[w/\tau^2 - R_2], \varepsilon_3[w/\tau^3 - R_3]\}.$$

Applying the Discrete Choice Theorem we have,

$$\begin{aligned}s_1 &= \frac{[w/\tau - R_1]^\theta}{\sum_{k=1}^3 [w/\tau^k - R_k]^\theta} \\s_2 &= \frac{[w/\tau^2 - R_2]^\theta}{\sum_{k=1}^3 [w/\tau^k - R_k]^\theta} \\s_3 &= \frac{[w/\tau^3 - R_3]^\theta}{\sum_{k=1}^3 [w/\tau^k - R_k]^\theta}.\end{aligned}$$

If we have data giving the shares of households at each location, and reservation rent for location 3, then we can solve for land rent at each location more-or-less in the
3885 same way as with additive commute costs.

6.7 The big prize: choosing discrete workplace *and* residence

The defining assumption of the monocentric city model is that it is monocentric. Everyone is assumed to work in the center and only chooses where they live. While
3890 this assumption seems defensible for thinking about a 19th century mill town (where most of the economic activity in the town actually did center on single place), it is much harder to defend for a modern metropolis. For example, both midtown and Wall Street are centers of activity in Manhattan, while figure 3.4 suggests that cities with more than one employment center are common.

3895 Relaxing the monocentric city assumption in a model with homogeneous households and continuous space has attracted the attention of two generations of urban economists [Fujita and Ogawa, 1982, Lucas, 2001, Lucas and Rossi-Hansberg, 2002, De Palma et al., 2019, for example]. These efforts made modest progress, but the problem and its solution are so hard that their results were not much used.

3900 But the discrete choice problem offers another way. We can use the discrete choice framework to get away from the monocentric city assumption by letting households choose pairs of locations, one for work and one for residence, rather than just choosing where to live. In this case, at the price of a little bit more complicated notation, we can let people choose both their place of work and residence. This opens the door to

³⁹⁰⁵ models of cities with realistic geographies with many employment centers.

To see how this works, consider the simplest possible case, a city with just two locations. Households choose one location for work and one (maybe the same one) for their residence. Let $i, j = 1, 2$ index the two locations. The possible choices for a household are $(i, j) = (1, 1), (1, 2), (2, 1), (2, 2)$. Each location pays wage w_i and has residential rent R_j . Commute costs are iceberg, with $\tau_{ij} = 1$ if $i = j$ and $\tau > 1$ otherwise. This means that households have a different budget set for each of the four possible choices. These budgets are listed in the table below,

(i, j)	Budget	c_{ij}	
$(1, 1)$	$w_1 = c_{11} + R_1$	$c_{11} = w_1 - R_1$	
$(1, 2)$	$w_1/\tau = c_{12} + R_2$	$c_{12} = w_1/\tau - R_2$	(6.14)
$(2, 1)$	$w_2/\tau = c_{21} + R_1$	$c_{21} = w_2/\tau - R_1$	
$(2, 2)$	$w_2 = c_{22} + R_2$	$c_{22} = w_2 - R_2$	

Notice that households only pay commuting costs if they don't live and work in the same place.

³⁹¹⁵ Each household type ν gets a Frechet taste shock for each workplace-residence pair. This is four taste shocks, $(\varepsilon_{11}, \varepsilon_{12}, \varepsilon_{21}, \varepsilon_{22})$. A household's utility conditional on workplace-residence choice i, j is $V_i(\nu) = \varepsilon_{ij} c_{ij}$, and a household makes the discrete choice between four such payoffs,

$$\max \{V_{11}(\nu), V_{12}(\nu), V_{21}(\nu), V_{22}(\nu)\}.$$

Applying the Discrete Choice Theorem we have

$$\begin{aligned} s_{11} &= \frac{c_{11}^\theta}{\sum_{(i,j)=(1,1),(1,2),(2,1),(2,2)} c_{ij}^\theta} & (6.15) \\ s_{12} &= \frac{c_{12}^\theta}{\sum_{(i,j)=(1,1),(1,2),(2,1),(2,2)} c_{ij}^\theta} \\ s_{21} &= \frac{c_{21}^\theta}{\sum_{(i,j)=(1,1),(1,2),(2,1),(2,2)} c_{ij}^\theta} \\ s_{22} &= \frac{c_{22}^\theta}{\sum_{(i,j)=(1,1),(1,2),(2,1),(2,2)} c_{ij}^\theta} \end{aligned}$$

3920 If we substitute into equation (6.15) using the budgets from equation (6.14), then we can write the shares of households making each pairwise choice in terms of the wages and rents in each location. Once we make this substitution, we will have four equations with nine unknowns, $(s_{11}, s_{12}, s_{21}, s_{22}, R_1, R_2, w_1, w_2, \tau)$. That is, four shares, the wage and rent in each location, and iceberg commute cost. This is more or 3925 less the same problem that we solved at the end of section 6.4. Given data describing any five of these nine unknowns, then finding the others is just a (complicated) algebra problem. Box 6.7.1 gives an example.

Conceptually, there is not much difference between this problem and one with many locations and pairwise commute costs based on how long it actually takes to 3930 travel from place to place. The steps for solving this more general problem are the same as what we have just gone through, with the difference that the algebra problem you get at the end is too hard to solve by hand and needs to be solved numerically using a computer. In the next section, we'll talk about an example.

Recall that one of the more interesting things that we could do with the monocentric city model with homogeneous households, was to evaluate various comparative 3935

statics. For example, what happens to land rent as we go further from the CBD, or as commuting costs or income increase?

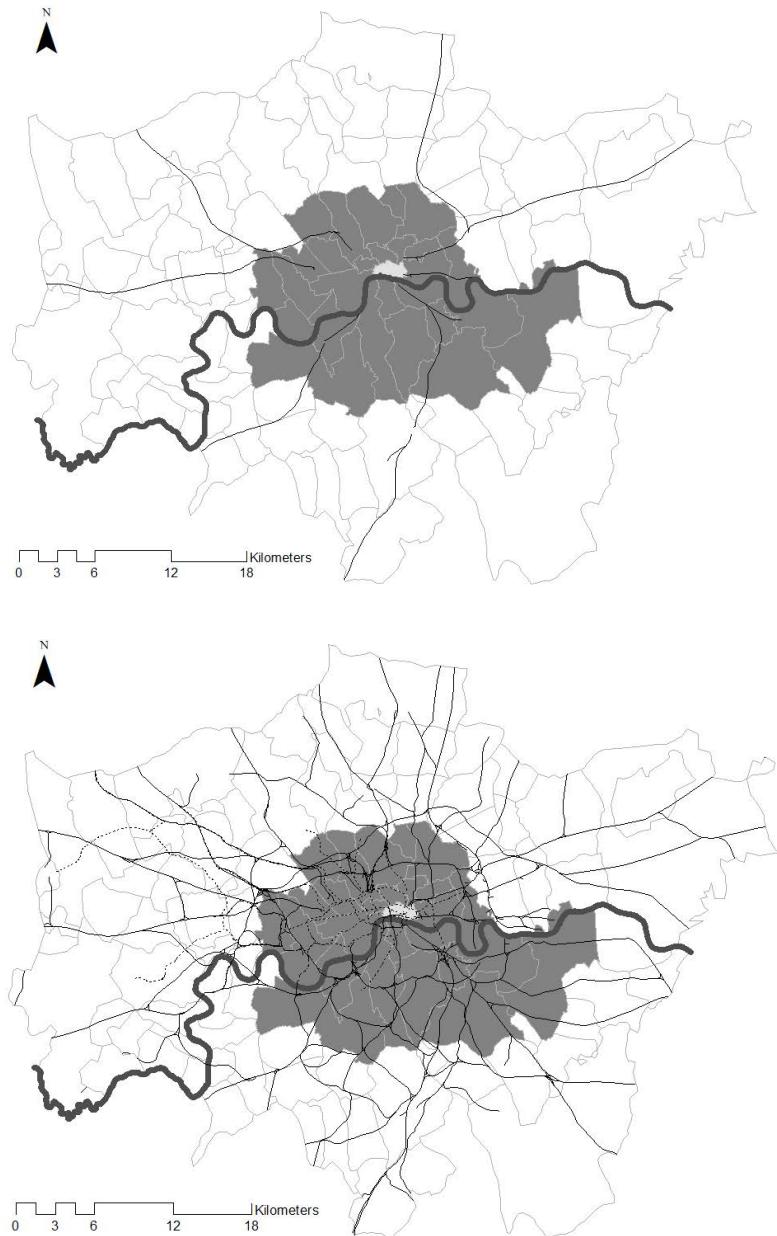
The model we've just solved doesn't have a CBD in the same way as the monocentric city model, so we can't ask about what happens with distance to the CBD, but
3940 we can ask how households' choices change as we, for example, change the iceberg commute cost, change the wage in one or both locations, or tinker with the amount of heterogeneity in the taste shocks. The difference is that the results of these calculations will not be as easy to represent as they were for the monocentric city model (where we could usually put them on a graph like 1.3) and they will not generalize
3945 beyond the particular geography we've used.

With this said, this model is pretty neat. We can use any geography for work and residence location that we like, and we no longer have to assume that everyone works at the center. We can solve the model so that it matches the available data. Once we have done this, we can evaluate comparative statics, usually called "counterfactuals"
3950 to ask how the city changes as we change commuting costs, wages, taste dispersion, or any other parameter of the model.

6.8 Railroads, subways, and the economic geography of London, 1866-1921

London was one of the first cities in the world to build a subway system. Construction
3955 began in the 1860s, and a substantial network was in place by the 1920s. Even before the advent of the subway, starting in the early 1800s, the city was served by a fairly extensive rail network, which allowed the movement of both people and goods in and

Figure 6.6: Railroads and subways in Greater London in 1841 and 1921



Note: Figure shows the modern extent of Greater London. The medium gray region in the center is London County, municipal London. The small light gray region at the very center is the City of London, the CBD and home to much of the modern financial sector. Light gray lines indicate Borough boundaries. The top panel shows the extent of the railroad in 1841 as thin black lines. The bottom panel shows the extent of the railroad in 1921 as thin black lines, and the subway as thin black dashed lines. Figure based on figures 1b and d of Hebllich et al. [2020], ©Oxford University Press.

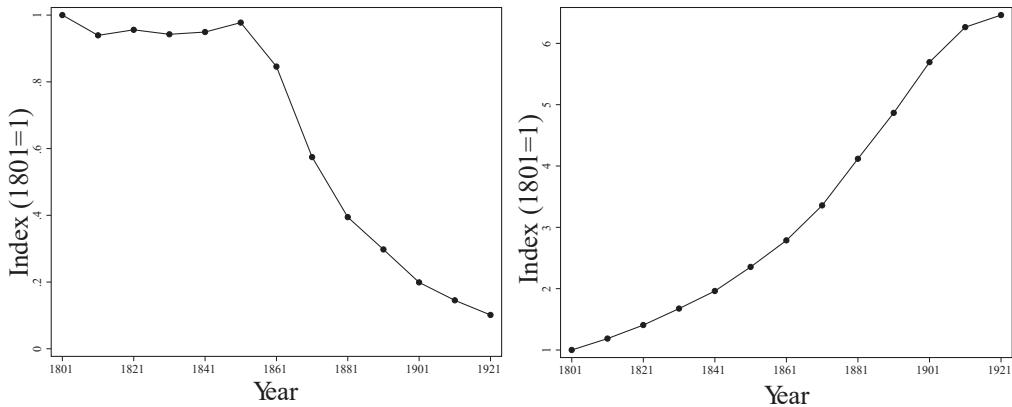
out of the city. In an early application of quantitative spatial models, Heblich et al. [2020] construct a model that allows them to estimate how the railroad and subway network constructed during the 19th century changed the economic geography of London.

The top panel of figure 6.6 illustrates the extent of the rail network in 1841, before the construction of the subway, in the area that makes up the modern extent of Greater London. The heavy black line running left to right through the center of the image is the River Thames, which approximately divides the city. The light gray lines are the boundaries of the various Burroughs that make up the Greater London. The medium gray center of the picture is London County, municipal London. The light gray area at the center of the picture, just visible if you look closely, is the City of London, the CBD for the city and location of the modern financial district. The thin black lines show the pre-1841 railroad lines. Depending on how you count, there are six or seven rail lines in service before 1841. Notice that all of the rail lines travel approximately radially from the City of London at the center, but that none of them quite reach it.

The bottom panel of figure 6.6 is similar to the top, but shows the extent of the railroad and subway network as of 1921. The railroad is still thin black lines, and by 1921, the railroad network serving London was extensive. Much of the modern subway network was also in operation, and subway lines are indicated by thin dashed lines. The subway serves mostly municipal London, and in particular, serves the City of London.

In light of what we have learned so far, it is natural to expect that this change in transportation infrastructure, and hence transportation costs, had important implica-

Figure 6.7: Change in the population of Greater London and City of London

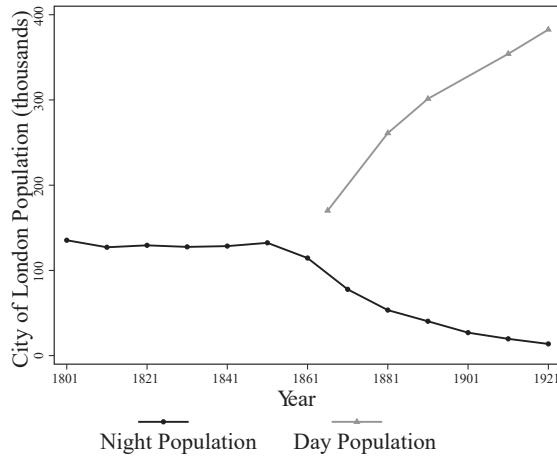


Note: *Left panel shows the residential population of the City of London, London's CBD, relative to its residential population in 1801. By 1921, the City of London housed fewer than a fifth as many people as it did in 1801. The right panel shows the residential population of Greater London over the same time period. The population of greater London increased by about a factor of six between 1801 and 1921. Figure based on figure 2 of Hebllich et al. [2020], ©Oxford University Press.*

tions for how the city was organized. Figure 6.7 suggests that this was the case. The left panel shows the residential population of the City of London, the very central part of the city, between 1801 and 1921, relative to its level in 1801. We see that the residential population of the City of London was about constant at its 1801 level until 1861, two years before the subway first began operation. Between 1861 and 1921, the residential population of the City of London fell to about an eighth of its 1801 value. From these data alone, it looks like the effects of the subway on the center of London were catastrophic.

The right panel of figure 6.7 shows the population of Greater London over the same time period, also relative its level in 1801. Here we see just the opposite. The

Figure 6.8: City of London population and employment



Note: Residential population of the City of London in thousands as the black line, and daytime population, i.e., employment as the light gray line. Even as residential population fell in the city of London, employment increased even more rapidly. Figure based on figure 3 of Hebllich et al. [2020], ©Oxford University Press.

population of Greater London grew steadily between 1801 and 1921, and the city in 1921 was nearly seven times as large as it was in 1801. Unlike the population decline in the center however, the growth of population in Greater London does not show a dramatic trend break around the opening of the subway in 1861.
3995

In all, this looks a lot like what we saw in US cities as result of the construction of radial interstate highways (in Chapter 2), but is even more dramatic. US cities in the second half of the 20th century experienced a 17% decline in population, the City of London saw its population decline by about factor of 8.

4000 The data that Hebllich et al. [2020] use records both residential population and daytime population. Daytime population is the count of people in every building during the day, workers and residents, and together with residential (or nighttime)

population, it gives us a pretty good measure of employment. Figure 6.8 reports on daytime and residential population in the City of London between 1801 and 1921.

4005 The black line reports residential population in thousands (this is exactly the same plot as in left panel of figure 6.7, but with the y -axis rescaled.) The light gray line shows daytime population, from 1861 (when the data is first available) until 1921. We see that daytime population increased even more rapidly than residential population decreased. Between 1861 and 1921, the residential population of the City of London decreased from about 120,000 to about 20,000. Over the same period,

4010 daytime population increased from about 170,000 to nearly 400,000. This suggests that the City of London, the CBD, transformed from a place where the number of residents about matched the number of jobs and not many people commuted, to a place where jobs outnumbered residents by more than a factor of ten. This could

4015 only occur if at least nine of ten workers commuted into the City to work.

This transformation appears to begin around the time the first subway line was opened in 1863, and so it is natural to suspect that the subway was partly responsible. However, much else was happening in London during this time. Not only was there a massive expansion of the railroad network, but this was the time when the industrial revolution really took root, and so many other things changed during this time. In particular factories got bigger, and the city built more buildings. While it is natural to suspect that the subway is responsible for some of the transformation of the City of London, it is also natural to suspect that it is not responsible for all of it.

4020 One of the main goals of Heblich et al. [2020] is to figure out how much of the transformation of the city of London was due to the railroads and subway. To do this, they build a quantitative spatial model describing 19th century London, and

then consider what would happen if they removed the subway, practically, if they changed the iceberg commuting cost matrix in a way that reflects how it would have changed if the subway had not been built.

4030 To get from the example with two locations that we worked in section 6.7 to the model in Heblich et al. [2020], we need four main changes. First, instead of two possible locations for work and residence, each household chooses between each of the 55 boroughs of Greater London. These boroughs are illustrated in figure 6.7. In addition to these 55 boroughs, households are also allowed a 56th choice, out of town, much as in the open city version of discrete linear city described in equation 4035 (6.11). As before, household taste shocks are drawn from a Frechet distribution, though households draw 56^2 shocks, instead of four.

4040 Second, the cost to commute between workplace and residence is given by a transportation cost matrix, like equation (6.1), but with 56 rows and columns instead of four. Each of the elements of this matrix is filled out with an estimate of the pairwise cost of travel constructed using GIS software and estimates of the cost to travel between each pair of Boroughs using the best available means of travel from walking, train, subway and trolley. Loosely, the pairwise travel cost is what you would get if tried to calculate pairwise travel time from the center of Borough to the center of 4045 another if you could travel back in time use a modern route finding app to estimate your travel time.

4050 Third, households in Heblich et al. [2020] have a different utility function than we have used in our example in this chapter. In our examples, households have tried to maximize their consumption after paying for commuting and rent. This leads to household payoffs of the form, $V_{ij} = \varepsilon_{ij} \left(\frac{w_i}{\tau_{ij}} - R_j \right)$. Households in Heblich et al. [2020]

have payoffs $V_{ij} = \varepsilon_{ij} \frac{w_i}{\tau_{ij} R_j^\beta}$. That is, wages divided τ and R . Looking back at Chapter 3, we see that this is the indirect utility function that results when households are given the choice over both their location and the amount of housing to consume, and their preferences over housing and consumption take the Cobb-Douglas form given 4055 in equation (3.11). The two main changes to equation (3.11) are the addition taste shocks and that wages are discounted to account for commuting.

Finally, Heblich et al. [2020] allow wages and rents to vary with the number of people choosing to live and work in a location. All else equal, if more people want to work in a location, wages fall, and if more people want to live in a location, rents 4060 rise. This requires changing our problem so that wages, instead of being a constant, are a function of the share of households working in the location, and similarly for rents. This leads to a system of equations much like the one described in equation (6.15), but whose difficulty is beyond the scope of this book. If you are interested in seeing how this works, look at Heblich et al. [2020] or Thisse et al. [2024].

4065 While the model that Heblich et al. [2020] actually use is much more complicated than the one that we develop in section 6.7, at the heart of both models is a discrete choice problem where households choose a pair of locations, one for work and one for residence, accounting for the cost of commuting, for wages, and for the price of housing. Because the model in Heblich et al. [2020] is more complicated, it explains 4070 more things, for example, the gap between the rent of commercial and residential land and the amount of building floor space in each Borough. In spite of this, the solution method has much in common with the solution method we used for the toy example worked out in Box 6.7.1. The discrete choice model, together with other equations of the model give rise to a system of equations. After filling in enough

4075 of the variables in this system of equation with observed quantities, it is possible to solve for the remaining variables.

Once the model is solved once for actual data, it is possible to evaluate counterfactual behavior in which transportation infrastructure is different. Doing this requires the evaluation of counterfactual matrices of iceberg commuting costs that
4080 reflect transportation costs under imagined configurations of infrastructure.

This is just what they do. In particular, taking the railroad and subway network in 1921 as a baseline, they evaluate the actual transportation cost matrix, and consequent number of people commuting in or out of the City of London, just the calculation we performed in box 6.7.1. They then repeat this exact exercise twenty
4085 years earlier in 1901. Compared with 1921, in 1901, many things about London have changed, the transportation network is less developed, there are fewer buildings and people, and wages and rents are different. Thus, the changes in commuting reflect all of these changes. They then repeat this calculation for census years back to 1831. The black line in both panels of figure 6.9 plots their results. We see that the number
4090 of people predicted to commute between the City of London per day fell from about 360,000 in 1921 to about 30,000 in 1831.

Notice that these are not actual measured commuters, but model predicted commuters, just as the calculation we did in box 6.7.1 gave us a prediction for commuters from other data. Given this, it is natural to be a little suspicious, and to wonder
4095 whether the model predicted values look anything like the actual values. To check this, Hebllich et al. [2020] checks these predicted commute flows against data measuring actual flows and finds close agreement during the second half of their sample when the commute flow data is available. Therefore, while the commute flows for the

early part of the 19th century are imputed from other data, they are imputed using
4100 a process that works pretty well during the later part of the 19th century.

This tells us how much commuting to the City of London increased during the period when the railroad and subway were constructed. It does not tell us how important the subway and railroad were to this increase. The commute levels described by the black line in both panels of figure 6.9 also reflect changes in things other than the
4105 rail and subway networks. What we would like to do is to evaluate the comparative static, here a counterfactual, where we hold everything else constant and ask what would have happened if we dug up the railroad and the subway, or just the subway. This is exactly what they model lets Hebllich et al. [2020] do.

To perform this counterfactual exercise for 1891, we would like to consider the
4110 transportation cost matrix that we already constructed for this year. This transportation cost matrix will reflect iceberg commute costs calculated on the basis of railroads and subways actually present in 1891. To understand the effect of reducing railroad and subway infrastructure from its 1921 to 1901 levels, holding everything else constant at its 1921 levels, we would like to solve for the share of people commuting
4115 to or from the City of London in 1921, if travel costs were at counterfactual 1901 levels. This is similar, conceptually, to the counterfactual calculation we performed in box 6.7.1. Once this is done for 1901, we can repeat the exercise for all of the census years back to 1831.

The light gray line in the left panel of figure 6.9 reports these results. We see
4120 that if the 1921 railroad and subway network were reduced to its 1831 level, holding everything else constant at 1921 levels, then commuting would fall from its 1921 level of about 360,000 to about 100,000 in 1831. That is, most of the increase in commuting

to the City of London during the later part of the 19th century can be explained by improvements to the railroad and subway network. With that said, even without
4125 the new railroads and subways, commuting to the City of London would still have increased by a factor of three relative its level in 1831. All of the other things that changed between 1831 and 1921 are responsible for this difference.

We would also like to know whether the railroad or the subway was more important. To answer this question, we perform exactly the same counterfactual exercise
4130 we've just performed, but ask what would happen if we remove just the subway network, and leave the railroad intact at its 1921 levels. Thus, for 1901, we would like to calculate the travel cost matrix that results if we leave the railroads as their 1921 levels, but restrict the subway network to its 1901 levels. With this done, holding everything else constant at its 1921 level, and solve for the number of people commuting
4135 to the City of London. Having done this for 1901, repeat it for all of the census years back to 1831.

The light gray line in the right panel of figure 6.9 reports this result. Commuting the City of London falls only modestly as we reduce the extent of the subway level to zero in 1861. Of the about 360,000 commuters in 1921, about 300,000 of them would
4140 still have solved their discrete choice problem with a residence outside and a job inside the city of London, even if the subway had not been constructed. Comparing the light gray lines in the two panels of figure 6.9 we conclude that the railroad and subway network played an important role in reshaping the economic geography of London during the later half of the 19th century. Without the improvements in transportation infrastructure, it seems unlikely that people would have made choices that led to the
4145 rise of the City of London as a center of employment with few residents. More than

that, the model also suggests that the railroads did most of the work. Subways were a bit player.

Heblich et al. [2020] conduct a similar pair of exercises to examine the effect of
4150 railroads and subways and of subways alone on the total population of greater London.

They find much more modest effects. In particular, in 1830 the population of Greater
London was about 2 million and this increased to about 7.5 million by 1920. If we
increase travel costs to the their levels in 1830, before the subway had been constructed
and before much of the railroad network was built, and hold everything else constant
4155 at 1920 levels, the model predicts a population of about 6.4 million for greater London.
That is, of the 5.5 million increase in the population of Greater London between 1830
and 1920, 4.4 million would have occurred even without the construction of railroads
and subways. The effect of removing only the subways is even smaller.

Taken together, these results indicate the railroads and subway together had a
4160 dramatic effect on the economic geography of London. By allowing the separation
of workplace and residence, they allowed a massive decentralization of residence, and
a correspondingly important concentration of employment in the City of London.
This occurred, in spite of the fact that transportation infrastructure had only a small
4165 effect on the overall population of the city. Taken together, this suggests that the
main effect of the railroads and subways was to reshuffle the locations of work and
residence in a way that permitted much greater concentration of employment. Of
railroads and subways, railroads did most of the work. The effects of the subways
were qualitatively similar to those of the railroads, but much smaller.

It is worth comparing the model based prediction if Heblich et al. [2020] with
4170 the purely empirical results about highways and subways in Chapter 2. Here we

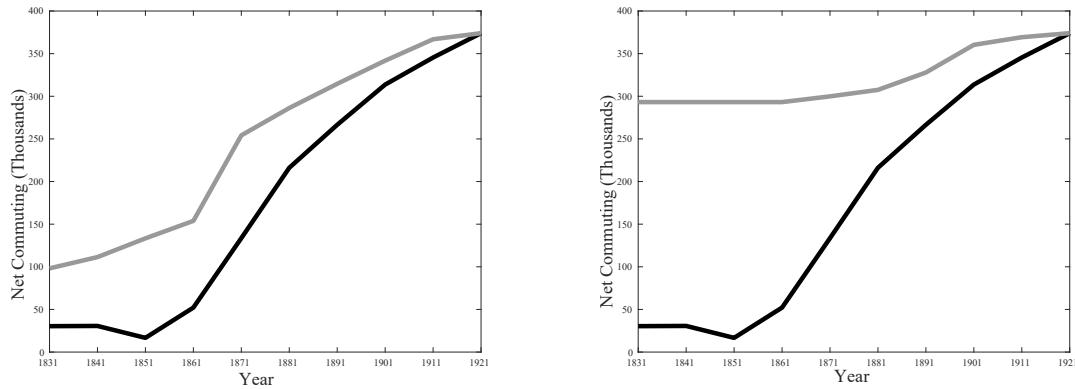
found that radial highways lead to a dramatic decentralization of population but that they have only small effects on the level of population. We found similar effects for subways. Subways clearly lead to a decentralization of lights at night, and seem not have any effect on population in a city. It is not clear how closely we should expect
4175 the effects of railroads and the subway in London to match the effects of highways in a sample of US cities, nor the effects of subways in a world sample of subway cities. With that said, the model predictions for London seem to line up qualitatively with results from studies of modern infrastructure. It is also not clear how close an agreement we should expect, nor how much disagreement is required before we should
4180 start to question the predictions of the model.

6.9 Conclusion

The monocentric city model makes a large number of qualitatively correct predictions about cities on the basis of a relatively small set of assumptions. More specifically, the monocentric city model assumes that households solve their maximization problem,
4185 that no one wants to move, and that commuting is costly. Using these assumptions, the model predicts downward sloping gradients for rent, population density, building height and housing consumption. Beyond these basic results, the monocentric city model also makes predictions for how the organization of a city should change as wages and transportation costs change. All of these predictions are broadly consistent with
4190 observation.

However, the monocentric city model rests on two assumptions that are obviously false. First, that everyone is the same, and second, that everyone works in the center.

Figure 6.9: Actual and counterfactual commuting with removal of all London rail and just subway, by decade



Note: *The dark gray line in both panels indicates total commuting flows between the City of London and Greater London, by decade. The rapid increase tracks the rapid increase in commuting to the City of London over the course of the 19th century. The light gray line in the left panel indicates counterfactual model predictions of the commuting that would have occurred if the rail network were reduced to its extent in that decade, but buildings, wages and rents remained at their 1921 levels. That the light gray line approximately tracks the dark gray line indicates that most of the increase in commuting can be explained in the model by the decreases in commuting costs that resulted from the expansion of the rail and subway networks. The right panel is similar, but considers what would happen if only the subway network were removed. That counterfactual commute flows change little from their 1921 level suggests that, conditional on the presence of the rail network, the subway is not very important for determining the level of commute flows into the City of London. Figures based on figure 9 of Heblitch et al. [2020], ©Oxford University Press.*

We should be suspicious of a model whose foundations are so obviously at variance with observation.

⁴¹⁹⁵ The assumption that households are all the same also limits the usefulness of the monocentric city model as a foundation for empirical investigations. A central question for many empirical investigations is to understand the relationship between

individual level heterogeneity and whether or not the individual receives whatever treatment is the subject of inquiry. The monocentric city model is basically useless
4200 for this sort of inquiry. It assumes away individual heterogeneity.

Quantitative spatial models offer a solution to these problems. At heart, these models are based on a discrete choice problem describing how an arbitrarily heterogeneous population chooses among a discrete set of alternatives. By casting these discrete alternatives as locations, we can incorporate the discrete choice machinery
4205 into the monocentric city model. This yields a monocentric city model with heterogeneous households, thereby relaxing one of the two onerous foundational assumptions of the monocentric city model.

By using a clever trick, we also relax the assumption that everyone work at the center. If the discrete alternatives that households face are *pairs* of locations, one for
4210 work and one for residence, then framing the household problem as a discrete choice problem means that we can let households choose both work and residence relaxing the requirement for central work in the monocentric city model. This trick has given rise to quantitative spatial models, and we went through an early example studying transit infrastructure in London.

4215 Quantitative spatial models have the advantage of being able to describe realistic, discrete geographies and allowing arbitrary configurations of work and residence. They also allow for a rich description of individual heterogeneity, and so in principle, can form a basis for thinking about the problem that I described to motivate this chapter, the sorting of inventive people into San Francisco.

4220 Quantitative spatial models are one of the most active areas of current research in urban economics. While we have learned much about them, a number of questions

remain to be resolved. In particular, the foundations for making welfare statements in these models are quite different than in the monocentric city model. In the monocentric city model, aggregate land rent is a complete measure of the surplus created by a city. In a quantitative spatial model, we must also account for consumer surplus.
4225 This means that we could find ourselves asked to choose between a city with a lot of land rent and not much consumer surplus, and the opposite case. The literature is just beginning to tackle this problem.

Beyond this, the complexity of quantitative spatial models makes it hard to tell
4230 whether comparative static results, i.e., counterfactuals, result from some unimportant simplifying assumption, or if they are more general. More intuitively, figures 6.9 and 3.2 both present similar comparative static results. Figure 6.9 describes changes in commuting patterns as transportation costs fall, and figure 3.2 describes what happens to land rent as we move further from the city center. In the case of figure 3.2
4235 it is transparent that the result follows from the shape of the indifference curve. In figure 6.9, it is less clear what mechanism or mechanisms, exactly, are at work.

Box 6.7.1: Discrete choice of workplace and residence: A numerical example

Consider the example described by equations (6.14) and (6.15). Suppose that we observe that $w_1 = w_2 = 3$, that $R_1 = R_2 = 1$, and that $\tau = 2$. Assume that we also know that $\theta = 2$. We can use this information to solve for the shares of the population making each pairwise choice of workplace and residence.

Using equation (6.14) and the given data, we have that $c_{11} = c_{22} = 2$ and that $c_{12} = c_{21} = \frac{1}{2}$. Substituting these values into equation (6.15) we have that

$$\begin{aligned}s_{11} &= s_{22} = \frac{8}{17} \\ s_{12} &= s_{21} = \frac{1}{17}.\end{aligned}$$

Thus, we can use data on wages, rent, iceberg commute costs and taste dispersion to solve for the share of households choosing each possible pair of workplace and residence.

Now suppose that we would like to evaluate a policy that improves transportation infrastructure and reduces the iceberg commute cost to $\tau = \frac{3}{2}$. In this case, we can go through the same logic to evaluate household choices in this hypothetical, or counterfactual case.

In this counterfactual case, we have that $c_{11} = c_{22} = 2$ and that $c_{12} = c_{21} = 1$. Substituting these values into equation (6.15) we have that

$$\begin{aligned}s_{11} &= s_{22} = \frac{4}{10} \\ s_{12} &= s_{21} = \frac{1}{10}.\end{aligned}$$

Thus, the reduction in transportation costs reduces the share of people who live where they work (i.e., choose $i = j$) from $\frac{16}{17}$ to $\frac{8}{10}$ and increases the share who commute from $\frac{1}{17}$ to $\frac{2}{10}$.

Problems

1. In this problem, we will work through an example of the discrete choice model with heterogeneous households. Consider a discrete linear city with three neighborhoods $i \in \{1, 2, 3\}$. Let x_i denote a neighborhood's distance from the CBD, with $x_1 = 1, x_2 = 2, x_3 = 3$. The cost to commute one unit distance is τ . The city is populated by households indexed by j . Each household chooses a neighborhood i , pays land rent R_i , and commutes to the center, at location 0, to earn wage w . A household's utility is $V_{ij} = A_i \cdot c_i z_{ij}$ where $A_i = i$ is the amenity value in location i , c_i is consumption and z_{ij} is the household and location specific valuation. All z_{ij} are drawn from a Frechet distribution, $F(z) = e^{-Tz^{-\epsilon}}$.

4240

- (a) Let consumption be $c_i = w - R_i - i\tau$. Set up the household's problem.
- (b) Using the big theorem from the lecture, solve for the share of household s_i in each location.
- 4245 (c) Let the share of households in each location $s_1 = s_2 = s_3 = \frac{1}{3}$, wage $w = 5$ and the price of agricultural land $\bar{R} = 1$. Assume that the land rent at $x = 3$ is equal to \bar{R} . Solve for R_1, R_2 and R_3 in terms of τ .
- (d) Solve for consumption in terms of τ .
- (e) Plot land rent and commuting costs as a function of i . How does this compare to the monocentric city model with a continuum of locations?
- 4250 (f) Do all households at location i have the same utility? What does this suggest about the usefulness of R to measure welfare?

6.10 Appendix: Proof of the discrete choice theorem

4260 **Theorem. Discrete Choice Theorem:** Suppose that households choose among N discrete locations, that for each location $i = 1, \dots, N$, household ν receives payoff $V_i(\nu) = \varepsilon_i u_i$, and ε_i is drawn from a Frechet distribution, $F(\varepsilon) = e^{-T\varepsilon^{-\theta}}$. Then the share of households such that

$$V_i(\nu) = \max \{V_1(\nu), V_2(\nu), \dots, V_N(\nu)\}$$

is

$$s_i = \frac{u_i^\theta}{\sum_{k=1}^N u_k^\theta}. \quad (6.7)$$

4265 Moreover, the average utility of a household making this choice is

$$E(V_i(\nu)) = \Gamma\left(\frac{\theta-1}{\theta}\right) \left(\sum_{i \in \{1,2,3\}} u_i^\theta\right)^{1/\theta}, \quad (6.8)$$

for all locations i .

Proof: To begin, solve $V_i(\nu) = \varepsilon_i u_i$ to get $\varepsilon_i = V_i/u_i$. Thus, if $F(\varepsilon) = e^{-T\varepsilon^{-\theta}}$, we can substitute to get a distribution for V_i , $G_i(V) = e^{-Tu_i^\theta V^{-\theta}}$. To make things easier to write, let $T_i = Tu_i^\theta$, and we have $G_i(V) = e^{-T_i V^{-\theta}}$.

4270 To make things a little easier, we'll restrict attention to the choice of the best of two alternatives rather than the best of N .

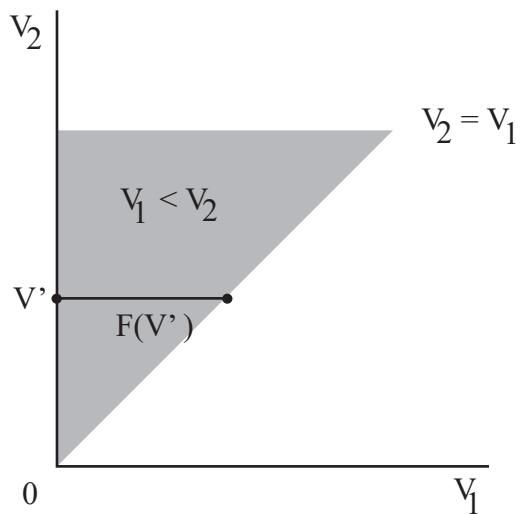
Consider two Frechet distributions,

$$F_1(V) = e^{-T_1 V^{-\epsilon}}$$

$$F_2(V) = e^{-T_2 V^{-\epsilon}}.$$

Suppose the two distributions are independent and that we take a draw from each, V_1 and V_2 . Agents choose their favorite alternative from these two possibilities. In 4275 this case, in a large sample, the share of agents choosing V_2 is $Pr(V_1 < V_2)$.

Figure 6.10: $Pr(V_1 < V_2)$



To evaluate this probability, we integrate the joint density of V_1 and V_2 over the gray region in figure 6.10. This is the region where $V_1 < V_2$. With V_1 and V_2

independent, this integral is,

$$\begin{aligned}
 Pr(V_1 < V_2) &= \int_0^\infty F_1(V)dF_2(V)dV \\
 &= \int_0^\infty e^{-T_1 V^{-\epsilon}} [T_2 \epsilon V^{-\epsilon-1} e^{-T_2 V^{-\epsilon}}] dV \\
 &= \int_0^\infty T_2 \epsilon V^{-\epsilon-1} e^{-(T_1+T_2)V^{-\epsilon}} dV \\
 &= \frac{T_2}{T_1 + T_2} \int_0^\infty (T_1 + T_2) \epsilon V^{-\epsilon-1} e^{-(T_1+T_2)V^{-\epsilon}} dV \\
 &= \frac{T_2}{T_1 + T_2} [e^{-(T_1+T_2)V^{-\epsilon}}]_0^\infty \\
 &= \frac{T_2}{T_1 + T_2} [1 - 0] \\
 &= \frac{T_2}{T_1 + T_2}
 \end{aligned}$$

Substituting into this expression from the definition of T_i establishes the first part of
the theorem.

4280

To prove the second part of the theorem, consider V_i distributed Frechet. Then,

$$\begin{aligned}
 F_i(V) &= e^{-T_i V^{-\theta}}, \\
 E(V) &= \Gamma(1 - \frac{1}{\theta}) T_i^{1/\theta}.
 \end{aligned}$$

Again restricting attention to the case of choosing the best of two alternatives, we would like to evaluate the expected value of the best of these two draws. Define

$$V^* = \max\{V_1, V_2\},$$

for V_i independent Frechet draws as above.

4285 We can calculate the distribution of V^* , $F(V)$, as follows:

$$\begin{aligned}
 F(V) &\equiv Pr(V^* < V) \\
 &= Pr(V_1 < V \cap V_2 < V) \\
 &= Pr(V_1 < V)Pr(V_2 < V) \\
 &= F_1(V)F_2(V) \\
 &= e^{-T_1 V^{-\theta}} e^{-T_2 V^{-\theta}} \\
 &= e^{-(T_1 + T_2)V^{-\theta}}
 \end{aligned}$$

Therefore F is also Frechet. Hence

$$E(V^*) = \Gamma(1 - \frac{1}{\theta})(T_1 + T_2)^{1/\theta}$$

Substituting from the definition of T_i gives the result.

Chapter 7

The Roback Model and the Value of Amenities

7.1 What is the value of amenities?

To think about the optimal provision of location specific amenities, from parks to police protection, to environmental regulation, to building codes, we need to balance the benefits of a marginal unit against its cost. To do this, we must first measure the value of amenities to households and how they affect the productivity of firms.
4295 Fire protection makes a location a better place to live, but it can also make it a more productive place to work. If we want to think about the value of fire protection, then we care about how it affects both households and firms.

You might conjecture, and hope, that after spending so much time thinking about equilibrium in the monocentric city model, that we would not need to introduce a new model to solve this problem, that we could use the monocentric city model to

think how the value of amenities is related to wages and rents. This turns out to be correct. But, the Roback model provides a more realistic, more defensible description of firms and allows firms and households to compete for land, and the monocentric city does not. As such, the Roback model is probably a better foundation for empirical analysis.

The Roback model [Roback, 1982], sometimes also called the Rosen-Roback model, is a cousin of the monocentric city model that is particularly useful for thinking about the value of amenities. It is the second workhorse model of urban economics. By dropping commuting, the Roback model gives an easy way to estimate the values that households and firms assign to amenities from data describing wages and rents in different locations.

A little more concretely, if we want to know how to use cross-location differences in wages and rents to think about the value of changes to a location's attractiveness or productivity, the Roback model is the answer. For example, we expect that climate change will affect the attractiveness and productivity of cities differently. Can we, somehow, infer the costs of climate change from cross-city differences in rent, wages, and climate? This is exactly the sort of question that the Roback model answers.

7.2 Amenities and productivity in the monocentric city model.

The foundations of the Roback model are basically the same as those of the monocentric city model, everyone optimizes and no one wants to move. But the math is quite different, and a little more complicated. Given this, it's helpful to start thinking

about the problem in the context of the more familiar model.

4325 Consider the standard monocentric city model, with two modifications. First, each city has an amenity value, A_c , such that $u(A_c c) = \bar{u}$, just as in Chapter 1. Second, each city has a “productivity”, A_y , and this location specific productivity affects wages according to $w = w(A_y)$. We want more A_y to be good, so assume that wages are increasing with A_y .

4330 Amenities and productivities are not choice variables for the household, so the household’s problem is largely unchanged from the standard monocentric city model,

$$\begin{aligned} & \max_{c,x} u(A_c c) \\ \text{s.t. } & w(A_y) = c + R(x)\bar{l} + 2t|x|. \end{aligned}$$

In spatial equilibrium, no one wants to move, which means that consumption must be the same everywhere, or

$$u(A_c c^*) = \bar{u}.$$

Solving for c^* , we have

$$c^* = \frac{u^{-1}(\bar{u})}{A_c}. \quad (7.1)$$

4335 With free mobility, c^* must be constant. If it goes up, people move into the city, driving rents up, if it goes down, people move out, driving rents down.

Putting this together with the household budget, we have,

$$w(A_y) = c^* + R(x)\bar{\ell} + 2t|x|, \quad (7.2)$$

for all locations x .

What happens if productivity A_y increases? By assumption, an increase in A_y
 4340 causes an increase in $w(A_y)$. Looking at the equilibrium budget, equation (7.2), an increase in the left hand side must be offset by an increase on the right. But c^* is fixed by the free mobility condition, so the rent payment must go up, dollar for dollar. Therefore, when productivity increases, wages *and* rents both change by the same amount.

4345 What happens if A_c increases? Looking at equation (7.1), with \bar{u} fixed, free mobility requires that c^* go down. If c^* goes down in the equilibrium budget, then either rents must go up, or wages go down to preserve the equality. But wages are fixed at $w = w(A_y)$, so rents must go up. Therefore, when an amenity changes, *only* rents change.

4350 This suggests the basic logic of the Roback model. Suppose we change some attribute of a city, e.g., climate, public transit, or pollution, and we see wages and rents change. Then the change must have affected productivity. If we change an attribute of a city and *only* rents change, then the changes must have been an amenity.

If we are careful about our accounting, we can back out the value of the change
 4355 in amenity to households from how much it affects rents and wages. To do this accounting exercise, make the simplifying assumption that there is just one amenity, something like “days of sunshine” or “absence of crime”, that affects both households

and firms positively. Denote this amenity by A , and assume $A = A_c = A_y$.

Suppose that A increases from a low initial value, A_0 , to a higher value, A_1 . Let
 4360 $w_0 = w(A_0)$, $w_1 = w(A_1)$, $c_0^* = u^{-1}(\bar{u})/A_0$ and $c_1^* = u^{-1}(\bar{u})/A_1$. Let $R_1(x)$ and $R_0(x)$ be the corresponding equilibrium rent gradients. Finally, let Δx indicate the change in variable x when we change from A_0 to A_1 , for example, $\Delta A = A_1 - A_0$.

Using this notation, we can write the household budget constraint for the two cases as,

$$w_1 = c_1^* + R_1(x)\bar{\ell} + 2t|x|$$

$$w_0 = c_0^* + R_0(x)\bar{\ell} + 2t|x|.$$

4365 If I fix location, x , and subtract the second equation from the first, I get

$$\Delta w = \Delta c^* + \Delta R_0(x)\bar{\ell}.$$

Solving this expression for Δc^* and dividing both sides by ΔA gives

$$\frac{\Delta c^*}{\Delta A} = \frac{\Delta w}{\Delta A} - \frac{\Delta R_0(x)\bar{\ell}}{\Delta A}.$$

The left hand side describes a change in consumption (possibly a negative amount) per unit of amenity. The right hand side describes the corresponding change in wages and housing expenditure required to keep the household indifferent. This change in income is the change in wage per unit of A minus the change in rent per unit of A .
 4370

But the change in consumption that a household is willing to accept per unit of A , is a price. Actually, it is minus one times a price. A positive price indicates that

a household is willing to accept a decline in consumption in exchange for a unit of a good, and conversely.

4375 Thus we have,

$$\begin{aligned} p_A &\equiv -\frac{\Delta c^*}{\Delta A} \\ &= \frac{\Delta R_0(x)}{\Delta A} \bar{\ell} - \frac{\Delta w}{\Delta A}. \end{aligned} \quad (7.3)$$

This is a “Baby Roback Theorem”. It shows that we can calculate something that looks like the price of an amenity from rents and wages by using the basic building blocks of spatial equilibrium.

In the rest of the section, we develop the actual Roback Theorem. It is better than the baby version because it allows for a much richer description of firms, and allows households to choose both their level of consumption and their level of housing. However, as we will see, it actually looks quite similar to equation (7.3).

7.3 Roback Model

The details of the Roback model are different from those of the monocentric city model of Chapter 1 in four main ways. First, commuting within a city is free and is dropped from the model. Second, households choose their level of housing and other consumption, as in the monocentric city model with housing from Chapter 3. Third, firms consume land, and the Roback model describes firms more carefully than the monocentric city model. Fourth, firms and households compete for land in the city, while in the monocentric city model, firms are all at a point in the center and so,

4390

implicitly, do not consume land.

Reading Roback [1982] requires a familiarity with multivariate calculus beyond the scope of this book. In order to understand this model, we're going to work through an example based on Cobb-Douglas production and utility functions. This is one of
 4395 the main teaching examples in an intermediate microeconomics course.

The particular example we'll use assumes a city level amenity, A , that increases both the utility of residents and productivity of firms. This is something like “days of sunshine”. One can also imagine “amenities” that have opposite effects on firms than households. For example, air quality regulation is good for households but usually
 4400 bad for firms. To accommodate this in our framework, one could replace A with $1/A$ in either the production or utility function given below.

Households in the Roback model are similar to those in the monocentric city model with housing, with two changes. First, there is no commuting. Second, utility depends on a local amenity. To describe them, let c be consumption and ℓ_c be consumption of
 4405 residential land. Wages are w , rent is r . The Roback model is an open city model and reservation utility is \bar{u} . Each city, or location, has an amenity level A . Households have Cobb-Douglas utility functions over housing and residential land, and amenities enter multiplicatively. That is, $u(c, \ell_c) = Ac^\alpha \ell_c^{1-\alpha}$.

Households divide their wage between consumption and residential land to make
 4410 themselves as well off as possible. In math, households solve

$$\max_{c, \ell_c} U(c, \ell_c, A) = Ac^\alpha \ell_c^{1-\alpha} \quad (7.4)$$

$$\text{s.t. } w = c + r\ell_c \quad (7.5)$$

We also need to assume that A takes a value in some known interval to rule out the cases where A is zero or infinitely large.

Equation (7.4) is the problem that we solved in box 3.2.1. Using equations (3.3) and (3.4), we have household demand for housing,

$$\ell_c(w, r) = (1 - \alpha)w/r,$$

4415 and household demand for consumption,

$$c(w, r) = \alpha w.$$

If we substitute both demand functions back into the utility function, we get the indirect utility function. That is, utility as a function of prices and income rather than as a function of land and consumption,

$$\begin{aligned} V(w, r, A) &\equiv U(c(w, r), \ell_c(w, r), A) \\ &= A(\alpha w)^\alpha ((1 - \alpha)w/r)^{1-\alpha} \\ &= \alpha^\alpha (1 - \alpha)^{1-\alpha} A \frac{w}{r^{1-\alpha}}. \end{aligned} \tag{7.6}$$

It is conventional to write an indirect utility as $V(\cdot)$ to distinguish it from a regular

4420 utility function $U(\cdot)$.

Solving for the indirect utility function when households have Cobb-Douglas preferences, what we just did, is a standard exercise in a microeconomics class. We didn't change anything about the standard problem except the variable names.

In a spatial equilibrium no one wants to move, so we must have utility equal to

⁴⁴²⁵ the reservation utility for all households,

$$\bar{u} = U(c, \ell_c, A).$$

But the indirect utility function is still a utility function, so it must also be true that

$$\bar{u} = V(w, r, A). \quad (7.7)$$

Substituting our expression for the indirect utility function from equation (7.6) into the right hand side of equation (7.7), we have

$$\bar{u} = \alpha^\alpha (1 - \alpha)^{1-\alpha} A \frac{w}{r^{1-\alpha}}. \quad (7.8)$$

Rearranging gives

$$r = \left[\frac{\alpha^\alpha (1 - \alpha)^{1-\alpha} A}{\bar{u}} \right]^{1/(1-\alpha)} w^{1/(1-\alpha)}. \quad (7.9)$$

⁴⁴³⁰ Equation (7.9) is a complicated expression, but an important one.

Any pair r and w that satisfies equation (7.9) has the property that if the household chooses the utility maximizing bundle, taking rent and wages as given, they can just achieve the reservation utility level. Thus, this equation describes an indifference curve in (r, w) space. If a wage-rent pair is on this curve, an optimizing household gets exactly reservation utility level \bar{u} , but otherwise, an optimizing household gets a different payoff.

Using this last expression, we can also evaluate the rate at which r changes with

w in order to keep utility constant,

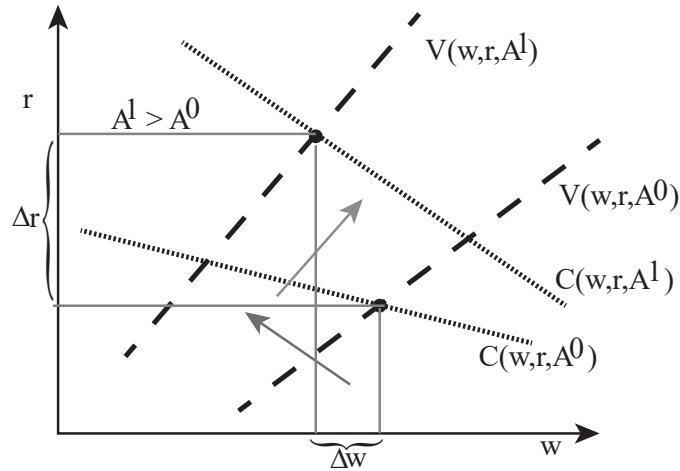
$$\frac{dr}{dw} = \left[\frac{\alpha^\alpha (1-\alpha)^{1-\alpha} A}{\bar{u}} \right]^{1/(1-\alpha)} \frac{1}{1-\alpha} w^{\frac{\alpha}{1-\alpha}}. \quad (7.10)$$

Because equation (7.9) describes an indifference curve, equation (7.10) describes the slope of this indifference curve.

From equations (7.9) and (7.10) we can make three statements about the behavior of indifference curves in (w, r) space. First, the slope of the indifference curve is positive. Second, as the amenity increases indifference curves shift up. Third, as the amenity increases, the slope of the indifference curve increases.

The two dashed lines in figure 7.1 illustrate. Both dashed lines are indifference curves in (w, r) space, with $A_1 > A_0$. Importantly, both curves describe pairs of wages and rent at which the same reservation utility can be achieved, but one curve describes the wages and rent that just permit this level of utility when amenities are A_0 , and the other when they are A_1 . Both curves are positively sloped, so higher wages are offset by higher rents to keep utility constant. The curve for A_1 lies above the curve for A_0 . As amenities improve, we need some combination of higher rent and lower wages to keep utility constant. Finally, as amenities increase, the indifference curve gets steeper. The intuition behind this is in three steps. First, as amenities increase, rent must increase in order to keep utility constant. Second, as rents increase, the household substitutes away from housing and toward consumption. Third, this means that a change in rent is operating on a smaller amount of housing. Therefore, to offset a one dollar increase in wages, we need a larger increase in rent. Equivalently, indifference curves in (r, W) space get steeper as amenities increase.

Figure 7.1: Indifference and isocost curves in wage-rent space #1



Note: The dashed lines illustrate two indifference curves in (w, r) space and how they change as the amenity A increases from A_1 to A_0 . Higher wages must be offset by higher rents to stay on an isoquant. People accept lower wages and higher rents as amenities increase. The dotted lines illustrate two isocost curves in (w, r) space and how they change as the amenity increases productivity. As productivity increases, wages and rents must increase to keep costs constant. In this example, the increase in amenities from A_0 to A_1 leads equilibrium rents to increase and equilibrium wages to decrease.

We now need to know how firms respond to rents, wages and amenities. Let N
⁴⁴⁶⁰ be the amount of labor employed by the firm and ℓ_p be the amount of land used in production. Firms make the production good Y from labor and land according to the production function,

$$Y = AN^{1-\beta}\ell_p^\beta, \quad (7.11)$$

where A is the same amenity as in the household problem. Notice that this production function is constant returns to scale. Doubling inputs exactly doubles outputs (see

4465 box 3.3.1 for a formal definition).

The firm chooses inputs to produce output Y as cheaply as possible. Hence, the firm solves,

$$\begin{aligned} \min_{N, \ell_p} C(w.r, Y) &= wN + r\ell_p \\ \text{s.t. } AN^{1-\beta}\ell_p^\beta &= Y. \end{aligned}$$

Because we restrict attention to a constant returns to scale production function, if we can find the way to produce *one* unit as cheaply as possible, we can just multiply by Y to get the cost for Y units. This means we can solve for the “unit cost function” without loss of generality. This problem is the same as the general cost minimization problem, but with $Y = 1$,

$$\begin{aligned} \min_{N, \ell_p} C(w.r, 1) &= wN + r\ell_p \\ \text{s.t. } AN^{1-\beta}\ell_p^\beta &= 1. \end{aligned}$$

To solve this minimization problem for the unit cost function, first solve the constraint for ℓ_p ,

$$\ell_p = A^{-1/\beta} N^{\beta-1/\beta}. \quad (7.12)$$

4475 Next substitute back into the firm’s problem. This gives an unconstrained minimiza-

tion problem in one variable,

$$\min_N wN + rA^{-1/\beta} N^{\frac{\beta-1}{\beta}}. \quad (7.13)$$

This is a calculus problem. To solve it, we evaluate the derivative with respect to N and set it equal to zero. This gives the first order condition,

$$0 = w + rA^{-1/\beta} \frac{\beta - 1}{\beta} N^{-1/\beta}.$$

Solving this equation for employment gives an expression for the cost minimizing level

4480 of employment to produce one unit of output,

$$N^*(w, r) = \left[\frac{w}{r} \frac{\beta}{1-\beta} \right]^{-\beta} A^{-1}. \quad (7.14)$$

If we substitute the expression back into the conditional factor demand, equation (7.12), we have

$$\ell_p^* = A^{-1/\beta} \left[\left[\frac{w}{r} \frac{\beta}{1-\beta} \right]^{-\beta} A^{-1} \right]^{-(1-\beta)/\beta}$$

Simplifying, this leads to,

$$\ell_p^*(w, r) = \frac{1}{A} \left[\frac{w\beta}{r(1-\beta)} \right]^{1-\beta}. \quad (7.15)$$

In the jargon, equations (7.14) and (7.15) are “conditional factor demands”. That

4485 is, the amount of labor and land a cost minimizing firm demands conditional on producing Y units of output (in our case $Y = 1$).

The total cost to produce one unit of output in the cost minimizing way is

$$C(w, r, 1) = wN^*(w, r) + r\ell_p^*(w, r)$$

Substituting from equations (7.15) and (7.14), and doing a lot of algebra, we get

$$C(w, r, 1) = \frac{1}{\beta^\beta(1-\beta)^{(1-\beta)}} \frac{w^{1-\beta}r^\beta}{A} \quad (7.16)$$

This is the cost minimizing way to make one unit of output at given input costs.

4490 $C(w, r, 1)$ is the unit cost function.

Next, we assume there is free entry of firms. An implication of free entry is that profits are driven to zero: if profits are positive, firms enter, if they are negative, firms exit. Recalling that with constant returns to scale, every unit costs the same to make, if unit cost is equal to price, then profits must be zero. To make things easy, 4495 we assume that the firm sells its output at a price of one. Altogether, this means that the formal statement of the zero profit condition is,

$$1 = C(w, r, 1) \quad (7.17)$$

This equation describes the set of wages, rents and (implicitly) amenities that let the firm produce an average unit of output at the market price.

Substituting from equation (7.16) into equation (7.17), we have

$$1 = \frac{1}{\beta^\beta(1-\beta)^{(1-\beta)}} \frac{w^{1-\beta}r^\beta}{A}.$$

4500 Solving this expression for rent gives,

$$r = \left(\frac{1}{\beta^\beta (1 - \beta)^{(1-\beta)} A} \frac{1}{w} \right)^{\frac{-1}{\beta}} w^{\frac{\beta-1}{\beta}}.$$

This function describes all of the pairs of (w, r) such that one unit dollar of output can be produced from one dollar of inputs. That is, it is an isocost curve.

We would like to know what this curve looks like. To start, differentiate it with respect to w , to learn its slope,

$$\frac{dr}{dw} = \left(\frac{1}{\beta^\beta (1 - \beta)^{(1-\beta)} A} \frac{1}{w} \right)^{\frac{-1}{\beta}} \frac{\beta-1}{\beta} w^{\frac{-1}{\beta}}. \quad (7.18)$$

4505 We can now establish some comparative statics. Because $0 < \beta < 1$, the right hand side of equation (7.18) is negative. That is, the firm's isocost curve has a negative slope in (r, w) space. Second, $\frac{dr}{dw}$ decreases (becomes more negative) as A increases (note the negative exponent on $1/A$ in equation (7.18)).

The dashed lines in figure 7.1 describe isocost lines for the firm for two values of 4510 the amenity, with $A_1 > A_0$. As A increases, productivity increases. To keep costs constant, we require a combination of higher wages and higher rent, so the isocost curve must shift up. As A increases, the isocost line also gets steeper. That is, more negatively sloped.

We've now worked out how isocost and indifference curves move around in (r, w) space as amenities change. We would like to understand how *equilibrium* wages and rents change as amenities change.

An equilibrium in the land and labor markets must satisfy four conditions; households optimize and no household wants to move, firms optimize and no firm wants to

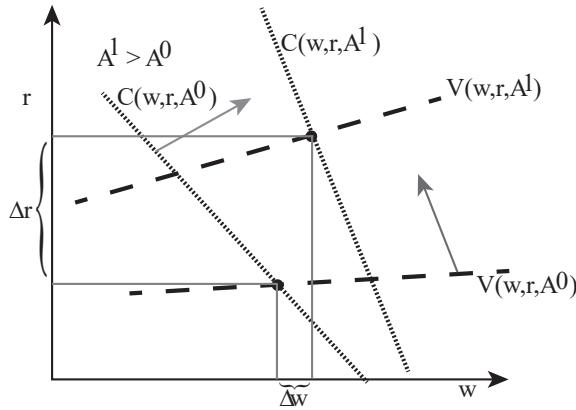
enter or exit. These conditions are stated formally in equations (7.4), (7.8), (7.13) and
4520 (7.17). To satisfy these conditions a pair of wages and rent, (w, r) , must lie on both the household indifference curve and the firm isocost curves. We see in figure 7.1 that this occurs where an isocost and indifference curve intersect. Each such intersection point is an equilibrium.

This means that we can also ask how equilibrium wages and rents change when
4525 amenities change. We know how isocost and indifference curves change, so we can work out how their intersection moves. This is what we have been working towards. It is illustrated in figure 7.1. As drawn, as amenities increase from A_0 to A_1 , equilibrium wages decrease and equilibrium rents increase.

This seems odd. As amenities increase, the city is better for households and more
4530 productive for firms. How can it be that wages and rents move against each other? What is happening is that the improvement in amenities is relatively more important. Because amenities are better, rents must be higher to keep utility constant. But as rents increase, firms substitute away from land. Because land and labor are complementary in the production process, the decreased use of land in production
4535 drives down the productivity of labor, and hence the wage. In this case, this indirect effect on wages is larger than the direct positive effect of increased productivity on wages, and the net effect is negative.

Things do not have to work out this way, and figure 7.2 illustrates another possibility. This figure is just like figure 7.1, but the slopes of the indifference and isocost
4540 curves are a little different. Here, as amenities increase, both wages and rents also increase.

Figure 7.2: Indifference and isocost curves in wage-rent space #2



Note: The dashed lines illustrate two indifference curves in (w, r) space and how they change as the amenity A increases from A_1 to A_0 . Higher wages must be offset by higher rents to stay on an indifference curve. People accept lower wages and higher rents as amenities increase. The dotted lines illustrate two isocost curves in (w, r) space and how they change as the amenity increases productivity. As productivity increases, wages and rents must increase to keep costs constant. In this example, the increase in amenities from A_0 to A_1 leads equilibrium rents and wages to increase.

7.4 Roback Theorem

All of this hardware sets us up to state the Roback Theorem (actually there are two).

These theorems will give us a way calculate the value of place specific characteristics that affect utility and productivity. This is useful because many place specific characteristics are things that do not explicitly transact in the market, like “the absence of crime”, “reliable bus service”, or “days of sunshine”, and so they are difficult to value because they don’t have prices.

Before we state these results, we need two tricks. First, we can define a price for

an amenity directly from the indirect utility function V as a ratio of the derivative

of the indirect utility with respect to the amenity to that with respect to the wage. That is,

$$p_A \equiv \frac{\frac{\partial V}{\partial A}}{\frac{\partial V}{\partial w}}.$$

This fraction is the ratio of the sensitivity of utility to the amenity to the sensitivity of utility to income. To see why it makes sense to think of it as a price, write each element of this expression just in terms of units. For example, $p = \$/A$. That is, the units of the price of amenities are dollars per unit of amenity. Rewriting the whole expression this way, we have

$$\frac{\$}{A} = \frac{\text{utils}}{\frac{A}{\text{utils}}}.$$

Simplifying the fraction on the right, shows that the units of the two expressions are the same. Thus, the ratio of partial derivatives, in fact, has the right units to be a price.

We also need another fact about logarithms. If x is a function of A , then recalling the derivative of a logarithm, we have

$$\frac{1}{x} \frac{dx}{dA} = \frac{d}{dA} \ln x(A). \quad (7.19)$$

This is really just an application of the chain rule.

We can now state the first part of Roback's main theorem:

$$p_A = \ell_c \frac{dr}{dA} - \frac{dw}{dA} \quad (7.20)$$

⁴⁵⁶⁵ Do not let the technical statement of this result distract you from how simple and intuitive it is. The Roback theorem says that the value of a marginal increase in the amenity A to a household is the sum of the change in wage and the change in total change housing expenditure required to obtain it.

⁴⁵⁷⁰ Notice that theorem is “marginal” in the sense that we are considering small enough changes in A the household does not reoptimize its choice of ℓ_c . For a sufficiently large change in A , we also expect ℓ_c to change, and no term involving $\frac{\partial \ell_c}{\partial A}$ appears in the theorem.

⁴⁵⁷⁵ The Roback theorem is important because it provides an elegant intuition for thinking about the value of amenities. At least as important, it also leads to a simple, practical way to calculate the value of these amenities. To see how this works, consider a set of cities $i = 1, \dots, K$, and suppose that you have data describing rents, r_i , wages, w_i , and amenities, A_i , for this set of cities. If we estimate the regressions,

$$r_i = B_0 + B_1 A_i + \epsilon_i$$

$$w_i = C_0 + C_1 A_i + \mu_i,$$

then $\frac{dr}{dA} = B_1$ and $\frac{dw}{dA} = C_1$. We can substitute these estimates directly into in the Roback theorem to get,

$$p_A = \ell_c B_1 - C_1. \quad (7.21)$$

⁴⁵⁸⁰ With B_1 and C_1 known constants, the only unknown on the right hand side is ℓ_c .

Recalling that ℓ_c is household land consumption. If we replace this variable with an estimate of average housing consumption in the city, then we can evaluate the

right hand side of equation (7.21) to calculate the value of a marginal unit of the amenity A , whether it is climate, crime, or good schools, from easily available data.
 4585 This is an important and widely used tool for this purpose. If you want to know whether you should direct public dollars to crime reduction or better schools, this is a pretty reasonable place to start.

For the purpose of the discussion above, choosing units for ℓ_c is a little fussy. To see this note that the rent in city i , r_i , only makes sense if we carefully define the object being rented so that it is the same everywhere. Does r_i describe the rent for a square foot of living space, a square foot of living space with in unit laundry? Or does r_i describe rent for an “average” apartment, where the average apartment needs to be the same across all cities? The important thing is that, whatever ever units we use to define r_i , we use the same units for ℓ_c .
 4590

4595 Dividing both sides of equation (7.20) by w , we have

$$\frac{p_A}{w} = \frac{\ell_c r}{w} \frac{1}{r} \frac{dr}{dA} - \frac{1}{w} \frac{dw}{dA}.$$

Using the fact about logarithms from equation (7.19), we get a widely used alternate statement of the Roback theorem,

$$\frac{p_A}{w} = \frac{\ell_c r}{w} \frac{d \ln r}{dA} - \frac{d \ln w}{dA}.$$

The ratio second $\frac{p_A}{w}$ is the value of amenities in a city as a share of the wage. This ratio is often used as an index for “quality of life” in a city. It reports the importance 4600 of amenities in producing utility relative to the importance of income.

We can use the same data and regressions that we used to evaluate p_A to evaluate

the quality of life index. Start with the same city level data and regressions that we used to estimate $\frac{dr}{dA} = B_1$ and $\frac{dw}{dA} = C_1$ above. Use these estimates, together with the fact about logarithms in equation (7.19) to calculate,

$$\begin{aligned}\tilde{B}_1 &= \frac{d \ln r}{dA} \\ &= \frac{1}{r} \frac{dr}{dA} \\ &= \frac{1}{B_0 + B_1 A_i} B_1.\end{aligned}$$

4605 Similarly,

$$\begin{aligned}\tilde{C}_1 &= \frac{d \ln w}{dA} \\ &= \frac{1}{C_0 + C_1 A_i} C_1.\end{aligned}$$

Substituting into the expression for the Quality of Life index, we get,

$$\frac{p_A}{w} = \frac{\ell_c r}{w} \tilde{B}_1 - \tilde{C}_1.$$

The only unknown left on the right hand side is $\frac{\ell_c r}{w}$. This expression is total expenditure on housing divided by total income, the share of household expenditure on housing. This is easy to observe, and in fact, figure 2.1 reports these estimates 4610 for a subset of US MSAs from Davis and Ortalo-Magné [2011]. On the basis of these results, it is common to use 0.25 as the housing share of expenditure for the purposes of calculating quality of life index values.

We have so far presented Roback's results that relate the value of amenities to

households to wages and rents. There is a parallel result that relates the value of
 4615 *amenities* to *firms* to wages and rents. This second part of the main Roback Theorem applies about the same logic to firms as the results above apply to households, to arrive at estimates of the effect of amenities on costs. Recalling that N is the population of the city, labor, and Y is total output, this result is,

$$\frac{dC}{dA} = \frac{\ell_p}{Y} \frac{dr}{dA} - \frac{N}{Y} \frac{dw}{dA}.$$

This result, too, has an easy, common sense interpretation. The left hand side is
 4620 the change in the cost to produce a single unit of Y that results from a small change in the amenity A . The first term on the right hand side is land per unit of output times the change in rent. That is, change in expenditure on land. The second term on the right is change in expenditure on labor. Thus, the value to the firm of a change in the amenity, really the change in unit cost that results from a change in
 4625 the amenity, is equal to the change in expenditure on inputs that results from the change in amenities.

Like the result for p_A , this theorem lets us calculate the marginal effect of amenities on the unit cost of output from things we can observe, city (or location) specific wages and rents. Also like the result for p_A , this result is marginal. We're considering small
 4630 enough changes in A that we don't need to worry about N , Y , or ℓ_p adjusting.

This result is used less widely than the result for households. However, if you want check a statement like “better public transit improves productivity”, this might be a good place to start.

7.5 Comments

4635 Roback [1982] generalizes her model to allow households to consume housing instead of land. This requires the addition of a construction sector where free entry drives profits to zero and construction is constant returns to scale. Conceptually, this generalization is similar to the case we have treated, but results in a system of three equations, one for firms, one for households, and one for the construction sector.

4640 This more complicated formulation is often used as the basis for empirical work.

While we have worked with particular functional forms, Roback [1982] works with arbitrary constant returns to scale production functions. This allows her to demonstrate that the Roback Theorem, equations (7.20), (7.7), and (7.13) hold for any economy where the production function is constant returns to scale and the utility function has convex indifference curves. It is useful here to compare the Baby Roback theorem, equation (7.3) with the grown up version, equation (7.20). Looking carefully, we see that the two equations are identical, except that one describes discrete changes in variables, “ Δ ”, where the other describes infinitesimal changes. This is a striking result because the models on which the Baby Roback and actual Roback theorem’s are based are quite different. This suggest that the Roback theorem may actually be more general than has so far been considered.

4645

7.6 Application #1: Valuing climate

There have been several papers that use the Roback framework, or something similar to it, to try to value climate. For example, Albouy et al. [2016] or Sinha et al. [2021].

4655 The basic quality of life index from Roback is,

$$\frac{p_A}{w} = \frac{\ell_c r}{w} \frac{d \ln r}{dA} - \frac{d \ln w}{dA}.$$

If we estimate the regressions,

$$r_i = B_0 + B_1 A_i + \epsilon_i$$

$$w_i = C_0 + C_1 A_i + \mu_i,$$

for a set of cities $i = 1, \dots, K$, then we can use the estimated coefficients to calculate $\frac{d \ln r}{dA}$ and $\frac{d \ln w}{dA}$ as in the discussion of the last section. This allows us to calculate the Roback quality of life index, for every city i , and also allows us to calculate the 4660 quality of life index in counterfactual cases where amenities are different.

To use this framework to think about the value of climate change, Albouy et al. [2016] makes several changes to this set-up. Of these, some are technical and I'll skip them. Three of them are more substantive and deserve comment,

The share of expenditure on housing is typically around 25%. However, for the 4665 purpose of calculating quality of life indices, it probably makes sense to think of the “income share of housing” as the “income share of housing and all other goods not traded across cities”. This is to reflect the fact that local services, like plumbers and restaurants also tend to be more expensive in places where rents are higher. This increases the weight placed on the $\frac{d \ln r}{dA}$ term in the index to about 33%.

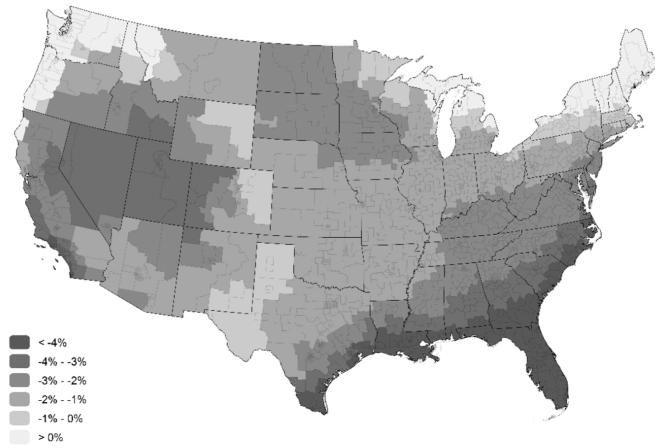
4670 Most income is taxed, and tax rates vary across cities. Given this, it probably makes sense to base the analysis on the after tax wage. Practically, this means scaling down the $\frac{d \ln w}{dA}$ term by $1 - \tau$, where τ is the relevant tax rate on income.

In addition to this, “climate” is complicated, and if we want to describe it, we’ll need a vector, e.g. spring, summer, fall, winter temperature and rainfall. Up until 4675 now, we’ve been thinking of A as a scalar. we’ll need to evaluate each of these climate variables, and then consider the total value of a hypothetical change in all of them.

Albouy et al. [2016] conduct this exercise. They use wage, and real estate data from the 2000 US census, and a change in climate consistent with ‘business as usual’ in 2100, that is, no effort at carbon reduction. This leads them to produce the map 4680 shown in figure 7.3. In the map, darker colors indicate places where p_a/w decreases by more in response to a change in climate. This seems pretty sensible. The biggest declines in the quality of life are in hot places, and some places that are too cold get better. Note that there is some ambiguity in the interpretation of the index. If 4685 climate change leads to wage increases in some places then the index will decrease, just as it would if amenities decreased.

Notice that when we apply the Roback model to a set of cities it does not “conserve people”. That is, there is nothing in the model to enforce the requirement that as we change amenities, the population of the whole system of cities, in this case the 4690 whole US, stays constant. Sinha et al. [2021] develops an extension of the model that remedies this. It is based on a discrete choice model that looks a lot like those we studied in Chapter 6.

Figure 7.3: Estimated changes in quality of life index from climate change



Note: Darker colors indicate places where the decrease in the quality of life index from predicted climate change by 2100 is larger. Figure reproduced from Alouy et al. [2016], ©University of Chicago Press, Association of Environmental and Resource Economists.

7.7 Application #2: Public finance in spatial equilibrium

Thinking about income taxes in the context of the Roback model raises a number
4695 interesting issues.

People require higher wages to live in worse places. With an income tax, these higher wages are taxed. On the other hand, if a household accepts lower wages to live in a nicer place, they do not pay taxes on the amenities they consume. That is, the benefit a household gets from amenities is not taxed, but the benefit from wages is.

4700 This creates an implicit subsidy for amenities and an incentive to move to nicer, lower wage places in order to cut your tax bill. In a spatial equilibrium, this means that an income tax tends to shift people away from productive places with bad weather.

Increasing marginal tax rates make the problem worse. The incentive to move away from a really productive place is stronger as the marginal tax rate increases.

- 4705 To make matters worse, places in the US that pay high taxes, on average, receive less money back from the federal government than they pay in, and conversely. This means that high productivity places are taxed partly to subsidize low productivity places. If the economy is described by a spatial equilibrium, then a place with good amenities and low productivity should have high rents and low wages in equilibrium.
- 4710 Because the US tax system redistributes income from more to less productive places, our high amenity, low productivity places also get a handout from the productive unpleasant places.

For the final complication, the “Home Mortgage Tax deduction” allows US households to pay mortgage interest with before tax dollars. This benefit is more important
4715 for people with high wages, living in expensive real estate markets, and conversely, it is least valuable to people with lower marginal tax rates living in cheap real estate markets. Note that the Home Mortgage Tax Deduction partly offsets the effects of the income tax. Where the income tax encourages people to move to low productivity places, the mortgage deduction encourages people to move to high productivity places.
4720

In Chapter 1 we considered how property taxes are capitalized into real estate prices. The Roback model, together with the discussion above, suggests that income taxes can also be capitalized into local real estate prices and can even affect local wages. The discussion above also suggests that but it will be hard to figure out how
4725 these effects will net out. The tax system is complicated, and in theory, it could affect local wages and real estate prices in different ways.

Albouy [2009] considers this problem and finds that the tax system has important implications for the economic geography of the US. To accomplish his analysis, he constructs a generalization of the Roback model describing all of the cities in the
4730 US.

Relative to the version of the Roback model that we have considered, the model in Albouy [2009] has the following additional features. First, it allows for three types of amenity per location, two that affect firms, and one that affects households. Second, he allows for capital in the production process. Third, he fixes the total number of
4735 workers in the economy. Fourth, he explicitly models the government as collecting taxes and redistributing them across cities. Fifth, he develops a detailed model of the tax system that allows for; progressive income tax rates, cross-state variation in state income taxes, and the home mortgage tax deduction. Finally, he describes the extent to which the federal government transfers tax revenue from one city to another.

4740 While the resulting model is more complicated than the version of the Roback model we consider, the same basic mechanisms are still at work. In a spatial equilibrium, households in places with better weather must suffer a combination of higher rents and lower wages to make them indifferent to a place with worse weather. Similarly, firms in less productive places must benefit from a combination of lower wages
4745 and lower rents to stop them from moving to more productive places. However, in Albouy's model the capitalization of amenities and productivity into wages and rents is complicated and distorted by the tax system, and the tax system is described in detail.

Table 7.1 presents his main results. To start, Albouy ranks cities by their net
4750 receipts from the federal government as a share of city income. To do this he first

calculates the sum of all federal income taxes paid by city residents. From this he subtracts all federal expenditures and transfers in the city, e.g., social security and highway construction. Finally, he divides this difference by the sum of all wages in the city. By this measure, San Francisco was the most disadvantaged city in the country
4755 (in 2000). As a share of the wage of an average resident, the federal government collected almost 4.8% more than it sent back. New York City was second most disadvantaged at 4.3%. At the other end of the table, as a share of the income of a typical resident, Killeen Texas received almost 6% more from the federal government than they paid in taxes.

4760 In light of our analysis of the Roback model, we should expect these tax differentials to operate on wages and rents through a number of channels. First, they act as an amenity (or disamenity). People in Killeen get free money, while people in New York must work a little harder. All else equal, this should depress rents in New York and raise them in Killeen. In addition, the income tax provides an incentive to move away from high wage New York to someplace where wages are lower and amenities are higher. Third, the home mortgage tax deduction partly opposes the income tax.
4765 It helps to lower the after tax price of real estate in productive, high wage places.

Albouy's model provides a logically coherent framework in which to make an educated guess about how all of these effects will net out, taking account of the
4770 measured productivity and amenities of each city. Relative to a counterfactual case where the tax system sent each city exactly as many dollars as it paid in, wages in San Francisco are 1.5% higher but real estate prices and employment levels are 48% and 28% lower. The effects in Killeen Texas are about opposite. Relative to what would happen with spatially neutral tax system, wages in Killeen are about 1.8%

Table 7.1: Tax differentials and their effects on prices and employment in US cities in 2000

	Tax Payment Rank	Total Tax Differential	Wage	Land Rent	Employment
San Francisco	1	.048	.015	-.480	-.288
New York, NY	2	.043	.013	-.434	-.261
Detroit	3	.036	.013	-.355	-.213
Norfolk, VA	188	-.030	-.009	.300	.180
Tucson, AZ	195	-.032	-.010	.322	.193
Killeen, TX	241	-.058	-.018	.583	.349

Note: *Table reproduced from Albouy [2009].*

⁴⁷⁷⁵ lower, but real estate prices and the employment level are 58% and 35% higher.

People in San Francisco must be compensated for their exported tax dollars with higher wages and lower rents. In spite of the higher wages and lower rents, labor must be relatively scarce before it can be productive enough to justify the higher wages, and so employment falls. Something about opposite happens in Killeen.

⁴⁷⁸⁰ 7.8 Conclusion

After the monocentric city model, the Roback model is one of the most used in the field. It gives us a tool to think about why people will pay higher rents to live in more polluted places, something that is difficult to think about in the monocentric city model. The big achievement of this model is to give us a way to use easily observable information on wages, rents and amenities to formulate a notion of the value of the amenity. This is really useful, both practically and conceptually.

Except for the focus on particular functional forms, the development of the model

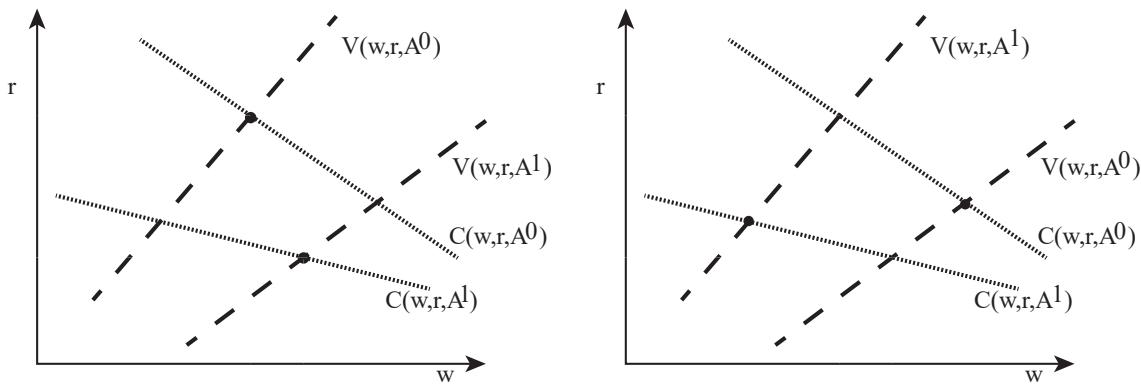
we have done here is general. Indeed, most applications of the Roback model rely on functional forms similar to those we have used here.

Problems

4790

1. Consider the following two figures showing indifference curves and iso-cost curves for two levels of an amenity, where $A_1 > A_0$.

Figure 7.4:



4795

- (a) For the economy in the left panel, does the amenity increase or decrease productivity? Does it increase or decrease utility and welfare? Explain briefly.
- (b) For the economy in the right panel, does the amenity increase or decrease productivity? Does it increase or decrease utility and welfare? Explain briefly.
2. This problem asks you to let $U(c, l_c, A) = \bar{u}$. Assume the household problem is

4800 given by

$$\max_{c, l_c} U(c, l_c, A) = Ac^{2/3}l_c^{1/3} \text{ such that } w = c + rl_c$$

- (a) Solve the constraint for c . Plug your expression for c into the utility function.
- (b) Solve the maximum problem for l_c .
- (c) Find the indirect utility function $V(w, r, A)$ by substituting demand for housing and consumption into $U(c, l_c, A)$.
- (d) Define an indifference curve by $V(w, r, A) = \bar{u}$. Solve for r in terms of A, \bar{u} , and w .
- (e) Evaluate $\frac{\partial r}{\partial w}$. What is the sign of this derivative?

Is A an amenity or dis-amenity from the perspective of the consumer?

4810 Explain briefly.

3. This problem asks you to calculate the importance of amenity A in real terms.

4815 Assume you have data on rents, wages, and amenity A for a cross-section of cities. That is, your data is $\{r_i, w_i, A_i\}$ for a set of cities $i = 1, \dots, J$. You may also assume that housing expenditure is one-third of the city wage. Describe the regressions you would run, and any subsequent analysis you would do, to determine the importance of amenity A in real terms (that is, as a share of the city wage).

Chapter 8

Agglomeration Economies, or Why

Are There Cities Anyway?

4820

8.1 Why are there cities?

The top right panel of figure 4.1 shows the rise of the urban share of population in the US over the 19th and 20th century, from about 5% in 1800 to more than 80% by early in the 21st century. The bottom right panel of this figure shows the simultaneous
4825 rise of income per capita. This is a period when per capita mean income in the US increases by about a factor of 13. Looking at these two figures, it is natural to wonder whether there is some relationship between the increasing population share of cities and increasing income.

Buttressing this conjecture, one of the main things people do in cities is work, and
4830 the nature of industrial production after the beginning of the industrial revolution suggests that packing people together for work is important. Thus, casual empiricism

Figure 8.1: Factory floors in the 19th century US and 21st century China



(see figure 8.1) also suggests that crowding into cities makes people more productive.

Why else would we have factories? And it is hard to imagine organizing this sort of employment density but not for cities.

4835 The monocentric city model assumes that, for some reason, people earn a high enough wage in the CBD that many of them want to pay rent and commute in order to be near it. We have so far not considered why this might be. We now consider this question. Why do people want to crowd together so badly that they put up with higher rents and commuting?

4840 An easy, if not very helpful answer is, “because people like to be near each other.” This raises two further questions. How much do people like to be near each other? Why do people like to be near each other? How does this desire for density affect how cities are organized? We have a an extensive empirical literature estimating how much people benefit from being near each other. We also have some evidence for why 4845 people benefit from being near each other.

8.2 Returns to scale and a planner's problem

At the heart of our explanation for why we pack ourselves into cities is the hypothesis that working at high densities, or in big groups, makes us more productive. There is some handy mathematical jargon for talking about this idea, “returns to scale,” and 4850 to investigate the relationship between the size of an enterprise and its productivity we will talk about returns to scale in production.

Let production in a city be Y , labor be N , and the production technology f , with

$$Y = f(N). \quad (8.1)$$

To keep things simple, suppose each person provides one unit of labor. This means that labor and population coincide and we can use the same symbol for both.

4855 Elaborating on the definition of constant returns to scale from box 3.3.1, say that f is

(IRS) Increasing Returns to Scale if $\alpha f(N) < f(\alpha N)$ for any $\alpha > 1$

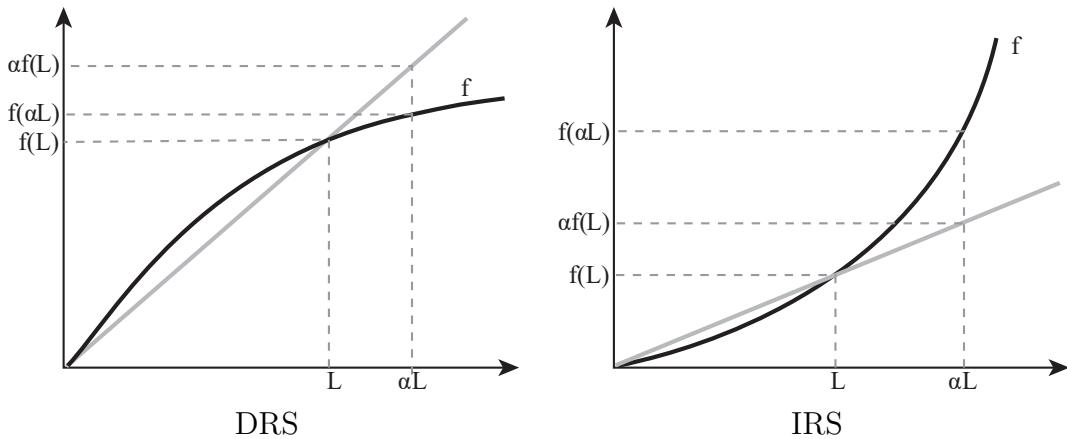
(CRS) Constant Returns to Scale if $\alpha f(N) = f(\alpha N)$ for any $\alpha > 1$, and,

(DRS) Decreasing Returns to Scale if $\alpha f(N) > f(\alpha N)$ for any $\alpha > 1$.

4860 In words, production is increasing returns if doubling inputs more than doubles outputs; decreasing returns if doubling inputs less than doubles output, and constant returns if doubling inputs exactly doubles output. Figure 8.2 illustrates the three cases.

By inspection of figure 8.2, we see a second important implication of returns to 4865 scale. Looking at the decreasing returns to scale function in the left panel, we see

Figure 8.2: Illustration of increasing and decreasing returns to scale production functions



Note: *The heavy black line in the left panel illustrates a decreasing returns to scale production function. In the right panel, the heavy black line illustrates an increasing returns to scale function. In both figures, the heavy gray line is constant returns to scale.*

that as x increases, the function gets flatter, or more formally, f' decreases. This means that each successive unit of x results in a smaller increase in output than the one before. With decreasing returns to scale, the marginal product of inputs is decreasing. Looking at the right panel of figure 8.2, we see that the opposite is true
4870 for increasing returns to scale. If production is increasing returns to scale, then the marginal product of inputs is increasing. With constant returns to scale, the heavy gray line in both panels, the marginal product of inputs is constant. In fact, decreasing (increasing) marginal product is equivalent to decreasing (increasing) returns to scale when we only have a single input to the production process.

4875 One last feature of returns to scale will be useful to us later. A decreasing returns to scale production process not only has decreasing marginal productivity, but the average productivity of inputs is decreasing in scale as well. Since each successive

unit of input makes less output than the one before, this drags the average down. Conversely, the average product of inputs is increasing in scale for an increasing
4880 returns to scale production process.

Given this language for thinking about productivity and scale, we can now consider how city size and returns to scale interact to affect the level of output. Suppose we have two cities, A and B . In each city, labor is converted into output as in equation (8.1). We would like to divide 1 unit of population between the two cities to maximize
4885 aggregate output.

Note that this analysis is a “planner’s problem”. We are asking what would happen if a single benevolent agent were in charge of everyone’s location choices. In general, the solution to a planner’s problem need not be a spatial equilibrium. That is, a configuration where no optimizing household wants to move.

4890 Because $N_A + N_B = 1$, we can always write the populations of cities A and B as,

$$\begin{aligned}N_A &= \frac{1}{2} - \varepsilon \\N_B &= \frac{1}{2} + \varepsilon,\end{aligned}$$

for some $-\frac{1}{2} \leq \varepsilon \leq \frac{1}{2}$.

Suppose that both cities have a DRS technology. If $\varepsilon > 0$ then by decreasing ε slightly, we shift population from the larger to the smaller city. Because the marginal product of labor is decreasing with DRS, when we shift a marginal unit of labor from
4895 the larger to the smaller city, we are shifting it from lower to higher productivity employment. It follows that decreasing ε increases total output. A symmetric argument applies for $\varepsilon < 0$. Therefore, aggregate output is maximized when $N_A = N_B = \frac{1}{2}$.

With DRS, there is no incentive to gather together, to agglomerate. Rather production should be as dispersed as geography allows. In this case, this means being evenly
 4900 split between the two locations. The left panel of figure 8.3 illustrates this case.

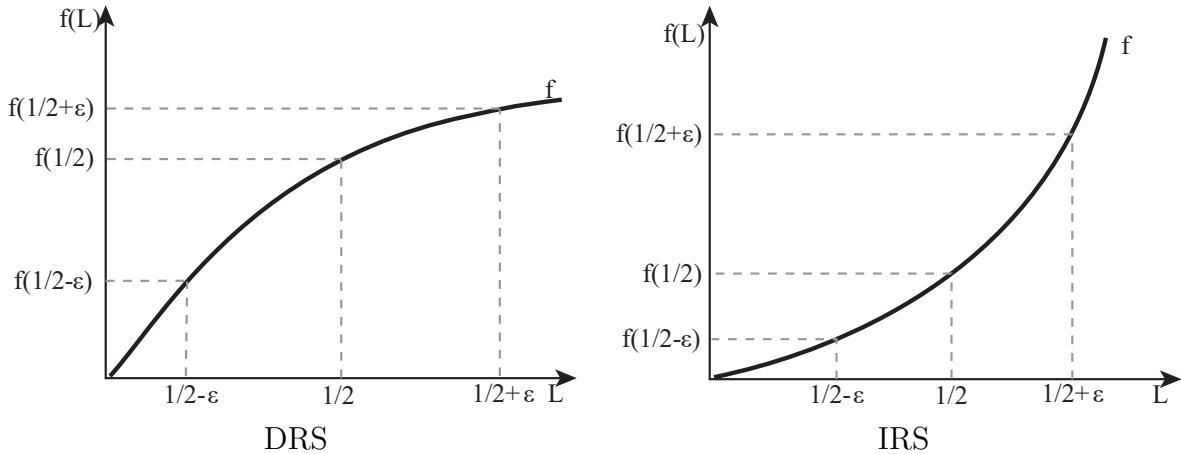
Now suppose that both cities have an IRS technology. If $\varepsilon > 0$, so city A is smaller than city B , then by decreasing ε slightly, we shift population from the larger to the smaller city. Because the marginal product of labor is higher in the larger city than the smaller one, this means we are shifting employment from more productive to less productive employment. It follows that increasing ε increases total output. The more unequal are city sizes, the more we exploit IRS in the larger city.
 4905 Interestingly, A symmetric argument applies when $\varepsilon < 0$ and city B is smaller than city A . With IRS, there is an incentive for people to gather together, to “agglomerate”. If we want to maximize output, we want everyone working in the same place, either,
 4910 $N_A = 1, N_B = 0$ or $N_A = 0, N_B = 1$. The right panel of figure 8.3 illustrates.

8.3 Returns to scale and equilibrium

With a competitive labor market workers are paid their marginal product. To understand the implications of this condition for wages under different sorts of returns to scale, we need notation to describe the actions of a “small” worker.

4915 For this purpose, index a set of small workers with $i = 1, \dots, N$ and let y_i be output for a particular small worker. Let n_i be the labor supply for small worker i . Let $N = \sum_j n_j$ be total labor supplied by all workers, and finally, let $Y = \sum_j y_j$ be total output for all workers.

Figure 8.3: Increasing and decreasing returns and aggregate output



Note: Left panel illustrates how aggregate outcome changes as population is divided more unequally between two cities with DRS technology. Right panel is the same, but when the two cities have IRS technology.

Suppose that our technology is,

$$y_i = AN^\sigma n_i,$$

for $i = 1, \dots, N$. That is, each worker's output depends on their own labor, n_i , on the total amount of employment in the city or factory where they work, N , and on a firm or city specific productivity parameter, A . A perfectly competitive labor market should operate by paying each worker their marginal product, that is $w = \frac{dy_i}{dn_i}$.

Using the definition of Y and y_i , we have

$$Y_i = \sum_{j=1}^N AN^\sigma n_i$$

Recalling that $N = \sum_{j=1}^N n_1$ and doing a little algebra, we can state the technology

for transforming labor into output as,

$$Y_i = AN^{1+\sigma}. \quad (8.2)$$

It is easy to check that this technology is IRS, CRS or DRS as σ is positive, zero, or negative. Urban economists often refer to σ as the “agglomeration effect,” or for reasons that will become clear below, as the “wage elasticity of city size.”

4930 Now consider the problem of a single worker choosing where to locate. We suppose that all firms and workers are “small enough” that their choice of n_i does not have a perceptible effect on AN^σ . This is the substance of the assumption that labor markets are competitive; all workers are too small to recognize their effects on other workers. Keeping this assumption in mind, calculate the marginal product of labor 4935 for a single small worker,

$$\begin{aligned} \frac{d}{dn_i} y_i &= \frac{d}{dn_i} AN^\sigma n_i \\ &= \left[\frac{d}{dn_i} (AN^\sigma) \right] n_i + (AN^\sigma) \left[\frac{d}{dn_i} n_i \right] \\ &= \sigma(AN^{\sigma-1})n_i + (AN^\sigma). \end{aligned}$$

As n_i gets small, the first term disappears, and we are left with

$$\frac{d}{dn_i} y_i = AN^\sigma. \quad (8.3)$$

That is, when workers are sufficiently small, they ignore the effect that their location choice has on aggregate output.

This gives us a way to think about how labor markets can work with IRS or
4940 DRS. Equation (8.3) is an expression for the marginal product of labor, even if it
is a bit myopic. A competitive labor market operates by setting the wage equal to
the marginal product of labor, as calculated in equation (8.3). The trick is to notice
that with IRS or DRS we need fudge a little bit on the calculation of the marginal
product, though in a way that is consistent with the behavior of workers who are
4945 small compared to the whole economy.

If you look carefully at equation (8.3), you will see that it is the average product
of labor, not the marginal product (divide the expression for total output in equation
(8.2) by N , and you get equation (8.3)). That wages are the average product of labor
creates a public goods problem. With increasing returns to scale, if a single worker
4950 decides to work in a location, then they increase the productivity of all the other
workers on average. This means that it is in the interest of incumbent workers to
pool their pennies and offer a new worker a bonus to come and work in their location.
But when we think about a market or competitive equilibrium, we rule out this sort of
coordinated activity by incumbents. This means that market equilibrium should not
4955 maximize output when we have increasing returns to scale. When there are increasing
aggregate returns, the individual incentive to crowd into a city is less than the social
incentive. Another way of saying this, and one that is common in the literature, is
that there is an “agglomeration externality” or just an “agglomeration effect”.

A symmetric opposite logic applies with DRS production. In this case, each
4960 marginal worker drives down the average productivity of all workers. This leads
to an equilibrium with too much concentration of workers. This is a just negative
agglomeration externality, but is usually called a “congestion externality”.

Our example focuses on how concentrations of people affect labor productivity. Similar logic allows us to ask how concentration of people affect firm productivity, or
 4965 other measures of economic output. It is also natural to think about agglomeration economies in consumption. There are clearly increasing returns to scale in the provision of museums, sports teams and the latest, coolest stuff, while decreasing returns to scale probably operate for things like traffic congestion, pollution, crime and park space.

4970 We can now think about what an equilibrium looks like with increasing returns to scale. To start, look at the right panel of figure 8.2. Suppose that workers are distributed across the two locations, just as indicated by the dashed lines. That is, with $1/2 - \varepsilon$ in one location and $1/2 + \varepsilon$ in the other. What should happen? Consider the problem of a worker in the smaller location, the one with only $1/2 - \varepsilon$. Recalling
 4975 that workers are paid the wage given by equation (8.3) in whatever location they choose, any worker in the smaller location who moves to the larger location should see a wage increase. Then what? With IRS, the movement of a worker from the smaller to the larger locations increases the gap in labor productivity, and so increases the incentive for the next marginal worker to move. This is going to continue until all
 4980 of the workers are in one place or the other. That is, when we start thinking about cities where the production technology is IRS, all of the workers should end up in one location or the other, though we don't know which location will win.¹

Notice that when we consider the case of IRS, we have two equilibrium outcomes. All the workers end up in city A or all the workers end up in city B. Multiple equilibria

¹Notice that there is also the special case where there are exactly half of the workers in each location. In this fragile situation, no worker wants to move because the payoffs are exactly the same in both locations.

4985 of this sort are common in models with IRS and they create three problems. First, they are an indication that the theory is incomplete. The theory does not provide a basis for selecting between multiple equilibria. Second, the model gives us no guidance about how, or whether economies can switch between different equilibria. Third, if we think that multiple equilibria are a feature of real world cities, then we may not
4990 know which equilibrium we are observing in measurements of cities.

Though there is pretty good evidence for some sort of increasing returns to scale in cities, we don't see the complete agglomeration of people into a few places the way that our little model suggests. We observe some large cities, so there must be some IRS. But, IRS must attenuate, somehow, with city size. The monocentric city
4995 model itself suggests a mechanism for this. Even if labor is more productive as cities get larger, commuting uses up more resources as cities get larger, so the costs of running a city increase with city size, too. The tradeoff between increasing costs and productivity as cities grow will help determine city sizes when we discuss systems of cities.

5000 Returns to scale and agglomeration economies give rise to all sorts of interesting incentive problems, *in theory*, and so they have attracted a lot of interest from economists. However, New York City MSA is about 20m people, while many MSAs of less than 100k also exist. The fact that we typically observe cities with wildly different sizes in the same country, and even quite near each other suggests either
5005 that increasing returns to scale is either not very important, or that it is about offset by the increasing costs of commuting and operating a bigger city. We will consider estimates of the magnitude of returns to scale in section 8.6.

Table 8.1: Possible outcomes when three firms choose between two locations at random

outcome #	firms			locations	
	1	2	3	# A	#B
1	A	B	B	1	2
2	A	A	B	2	1
3	A	B	A	2	1
4	A	A	A	3	0
5	B	B	B	0	3
6	B	A	B	1	2
7	B	B	A	1	2
8	B	A	A	2	1

Note: *This table describes all of the possible arrangements of firms. Each row describes one possible outcome. For example, in the first row, firm 1 chooses location A and firms 2 and 3 choose B. Two firms choose B and one chooses A. If firms choose randomly each outcome/row is equally likely.*

8.4 How to measure agglomerations?

One of the implications of agglomeration economies seems to be that production activity will concentrate in space. This is widely observed. Technology firms are concentrated in Silicon Valley. Automobile firms are concentrated in Detroit. On first look, this seems like strong evidence for increasing returns to scale.

But, could it be random? If firms were distributed randomly, what would it look like? Can we tell whether the distribution of firms across space is the result of random choices of location, or do firms choose to be near each other to exploit increasing returns to scale?

To think about these questions more carefully, consider an economy with three firms, $i, j = 1, 2, 3$, each choosing between two locations, A and B , at random. Half

Table 8.2: Distributions of firms and pairwise distances and their frequencies

s_n	#A	#B	$d_{ij} = 0$	$d_{ij} = 1$
1/8	3	0	6	0
3/8	2	1	4	2
3/8	1	2	4	2
1/8	0	3	6	0
Mean			$4\frac{1}{2}$	$1\frac{1}{2}$

Note: *Firms are completely concentrated in one location one time in four and at least two thirds of firms are always in one location.*

the time they choose A , half the time they choose B . What should we expect to

5020 observe?

Table 8.1 lists the eight possible outcomes. The top row of the table describes the outcome where firm 1 chooses location A and firms 2 and 3 choose location B , one firm in A and two in B . The successive rows of the table describe the other seven possible things that can happen. If the firms are choosing their locations at random, 5025 then each possibility should be equally likely and should occur in about one eighth of all trials.

Table 8.1 describes the way firms arrange themselves across two possible locations. To develop a test that we can use to test whether real life firms are choosing locations at random, we need to think about the distribution of pairwise distances between 5030 firms.

Table 8.2 describes these distances. With three firms, $i = 1, 2, 3$, there are six possible pairs of firms, $(1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3)$, although three of these pairs are trivial, a firm paired with itself. For each outcome listed in table 8.1, we can evaluate the distance, d_{ij} , between each of these six possible pairs of firms. The

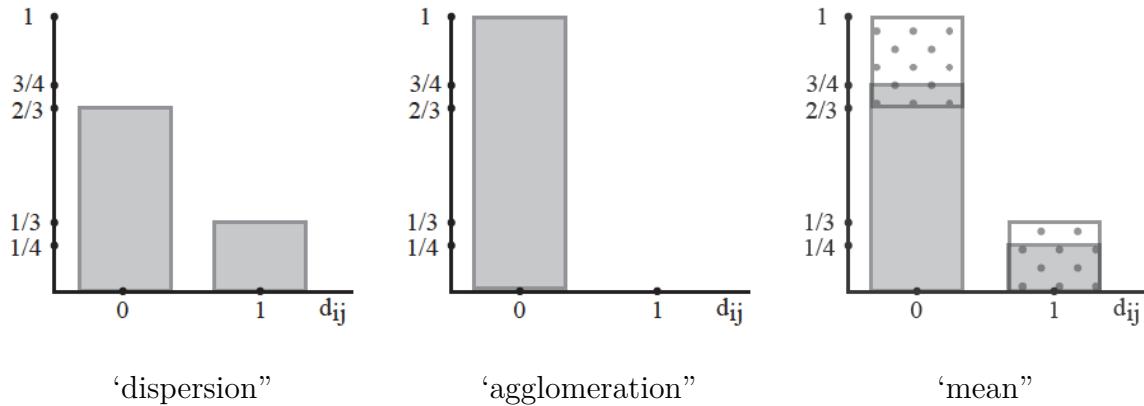
5035 distance between a firm and itself must be zero, so $d_{11} = d_{22} = d_{33} = 0$. To evaluate the others, suppose that if two firms are in the same location, either both in A or both in B , then $d_{ij} = 0$, and if they are in different locations, one in A and one in B , then $d_{ij} = 1$.

5040 Looking at all of outcomes listed in table 8.1, there are three outcomes (#2, #3 and #8) with two firms in A and one in B . The second row of table 8.2 describes these outcomes. With two firms in A and one in B , of the six possible pairwise distance between firms, two are equal to one, and the remaining four are zero, just as listed in the last two columns of the second row of table 8.2. The remainder of table 8.2 performs similar calculations for the remaining outcomes in table 8.1.

5045 Tables 8.1 and 8.2 describe what happens when three firms choose between two locations at random. Now suppose that we observe three actual firms choose between two locations, and we would like to know whether they have chosen at random, or if they have chosen with an eye to being near each other, as we would expect if IRS was important. In particular, when should we conclude that the concentration of firms is 5050 not consistent with random choices across symmetric locations?

To think about this, plot the histograms of pairwise distance. Looking at table 8.2, we see that there are only two cases. In one quarter of the outcomes, there are 6 pairwise distances equal to zero, and none equal to one. The middle panel of figure 8.4 plots this distribution of pairwise distances as a histogram. This is the case where 5055 firms agglomerate in the same location. In the other case described in table 8.2, four of the six pairwise distances are zero, and two are one. The left panel of figure 8.4 plots this distribution of pairwise distance as a histogram. This is the case where firms are as dispersed as this example allows.

Figure 8.4: Histograms of pairwise distances between three firms choosing between two locations at random



Note: *In the last panel, dotted area indicates the range of outcomes we can expect to see. If, after many trials, we see something at the edge of this range, then firms are probably not choosing locations at random.*

In the right panel of figure 8.4, we plot the average of the two panels to the left. If
 5060 we observe many trials of three firms choosing between these two locations at random, then for an average trial, we should see three quarters of pairwise distances between firms equal to zero, and one quarter equal to one. The dotted bands in the right panel indicate the range of outcomes we can possibly see.

If, after many trials, we see something near the edges of these bands, then firms
 5065 are probably not choosing at random. Either the firms have a systematic preference for one location or the other (natural advantage) or the firms coordinate to be near each other.

Duranton and Overman [2005] do exactly the exercise that leads to figure 8.4, but instead of using three pretend firms choosing between two pretend locations, they
 5070 look at the location choices of all firms in the UK. The basis for their study is the

Figure 8.5: Maps showing the locations of UK establishments in 1996.



Note: *The left panel illustrates the locations of all UK establishments in SIC classification 2441, “Basic Pharmaceuticals”. The right panel illustrates the locations of all UK establishments in SIC classification 2932, “Agricultural and Forestry Machinery”. Each dot shows the location of an establishment. Pharmaceutical establishments are concentrated in London. Agricultural and Forest Machinery Establishments are more dispersed. Figures reproduced from Duranton and Overman [2005], ©Oxford University Press.*

1996 UK census of establishments. In these data, “establishments” are either “firms” or, for firms with more than one facility, “plants”, so the data really describe plants rather than firms.

The UK census of establishments reports two main pieces of information about each establishment, its post code and its Standard Industrial Classification Code (SIC). Postal codes in the UK are a little bit smaller than a US zipcode (8-10,000

people), so knowing an establishment's post code lets Duranton and Overman locate the establishment on a map pretty precisely. Standard Industrial Classification codes describe the industry of the establishment in a detailed way. For example, SIC 2441
5080 is "basic pharmaceuticals" and 2932 is "other agricultural and forestry machinery".

Figure 8.5 illustrates the locations of establishments for these two SICs.

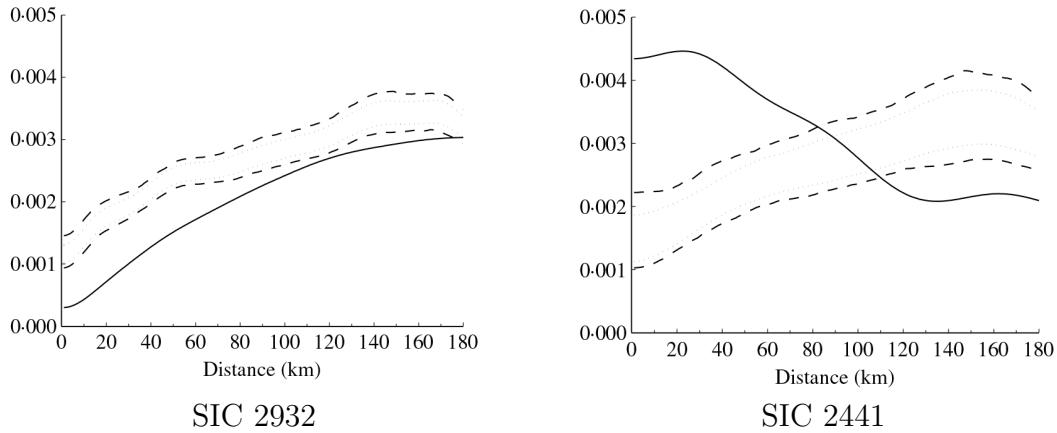
The data underlying the two maps in figure 8.5 is just the same as the data behind the example described by tables 8.1, 8.2 and figure 8.4. That is, the data describe the set of firms, really establishments, in an SIC, and a list of possible locations, all of
5085 the post codes containing one of the establishments in the SIC. There are just many more firms and locations than in our example.

Using these data, for each SIC, we can go through a process like the one that we went through with our toy example in tables 8.1, 8.2 and figure 8.4. That is, we can enumerate all of the possible distributions of establishments across locations, as in
5090 table 8.1. Using this list of outcomes, we can calculate the distributions of pairwise distances between firms for each outcome, as in table 8.2. Finally, in figure 8.6 we can plot a histogram of mean pairwise distances along with variation around these means, as in the right panel of figure 8.4.

These figures are constructed in much the same way as in our example. The solid
5095 line shows the realized histogram of pairwise distances. The dashed lines are like the "polka-dot envelope" in the figure based on our simple example. They show the range of outcomes we expect if firms chose randomly over the universe of all establishment locations (for all SICs). Pharmaceuticals are "too close" for randomness. Other Ag. etc., are a little too spread out.

5100 Repeating this exercise for each of the 234 industries described by the UK census

Figure 8.6: Histograms showing likely ranges for histograms of establishment pairwise distances and the observed histogram for two industries.



Note: Real world versions of figure 8.4 based on data for all UK establishments in SIC 2441, “Basic Pharmaceuticals” on the right and SIC 2932, “Agricultural and Forestry Machinery” on the left. The x-axis in both figures is pairwise establishment distance. The y-axis is the share of establishment pairs with pairwise distance x . The dashed lines give 10% and 90% bounds on the shares of establishment distances equal to x when establishments choose their locations at random. The solid lines describe the histogram of pairwise establishment distances for the actual locations of establishments. That the solid line does not lie between the two dashed lines means that the observed pattern of firm locations is unlikely to have come about as a consequence of random choices of location. Pharmaceutical firms are too close to each other for random choice, and Agricultural and Forest Machinery establishments are not close enough. Figures reproduced from Duranton and Overman [2005], ©Oxford University Press.

of establishments, 177 have distance profiles that don't seem random. Industries that are “too close” are much more common than “too dispersed”.

Ellison and Glaeser [1997] conduct a qualitatively similar (but much more complicated) exercise using US data, and reach a similar conclusion. Interestingly, they also find that employment is more concentrated than firms. That is, bigger firms are

more concentrated than smaller ones.²

8.5 How to measure agglomeration economies?

Recalling equation (8.3), when we are considering small workers, and workers are paid the marginal product of their labor, we have,

$$w = \frac{\partial}{\partial n_i} y_i = (AN^\sigma).$$

5110 If we take logs, this gives,

$$\begin{aligned} \ln(w_i) &= \ln(AN^\sigma) \\ &= \ln(A) + \sigma \ln(N). \end{aligned}$$

This means that we can learn about the extent of aggregate returns to scale by looking at the relationship between log wage and log total employment. This sort of log-linear relationship means that σ is an elasticity. That is, the wage elasticity of city size.

5115 Using data describing lots of workers, i , distributed across many cities, j . we can estimate this equation with the regression,

$$\ln(w_{ij}) = B + \sigma \ln(N_j) + \varepsilon_{ij}. \quad (8.4)$$

This is straightforward, and the resulting estimate of the wage elasticity of city size,

²We don't really have a theory for why, but this raises an important conceptual question: are bigger firms in cities because cities make them more productive (and therefore bigger), or do bigger firms locate in cities?

σ , gives us an answer to one of the two questions that led off this chapter, “How much do people want to be near each other?” The magnitude of σ tells us how much benefit people derive from being in the same city as other people.

5120 We can do this same exercise with other measures of worker output, e.g., patents per tech worker. We can also do something similar with firms, e.g. output per unit of input. The idea is similar, but the details are different.

Variations of this regression have been estimated in hundreds (if not thousands) of academic research papers. In spite of this, it remains an area of active research. There 5125 are three reasons this regression and its cousins have attracted so much attention.

First, anything to do with productivity and productivity growth sheds light on the remarkable increase in human wealth and population since the industrial revolution. This is surely one of the most important things to happen in the world in the last several thousand years, so understanding it better, rightly, attracts a lot of attention.

5130 Second, understanding why people are drawn together in cities is central to understanding why we have cities, so understanding the foundations of agglomeration economies is central to the study of cities.

Finally, a number of econometric problems arise when we try to estimate σ using equation (8.4). First, we expect that people will migrate to more productive places, 5135 for example, those close to a good harbor or a coal mine. In this case, we will estimate a large value for σ in equation (8.4), but this does not really reflect increasing returns to scale. Second, it may be that more productive people migrate to larger cities. In this case, estimates of equation (8.4) will reflect sorting of people rather than returns to scale. Finally, the same two effects can operate on firms. A lot of the research 5140 effort devoted to estimates of equation (8.4) is devoted to sorting out these problems.

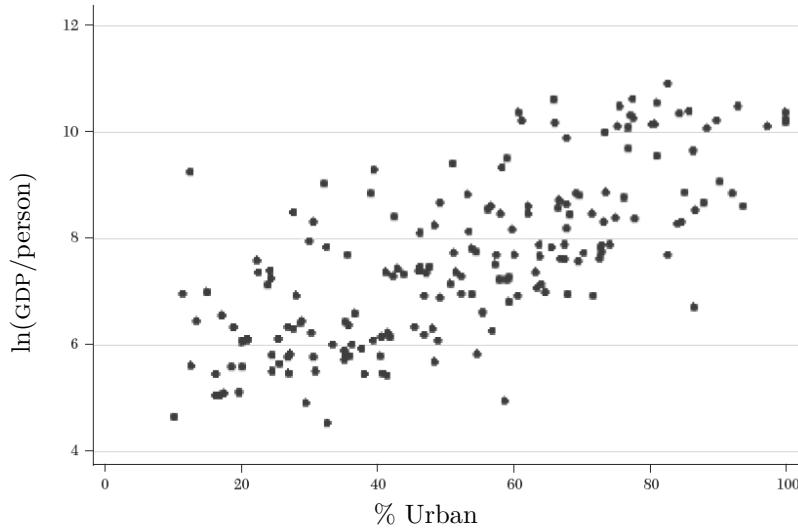
Notice that why agglomeration effects arise is important. If larger cities are more productive because they are located at better harbors, then there is no externality. Adding workers to such cities does not increase the productivity of incumbent residents. On the other hand, if there really are scale economies, then there is an
5145 externality and adding workers does increase the productivity of incumbent workers. This makes it less likely that a decentralized (spatial) equilibrium will have good welfare properties.

There are also a lot of details and practical problems to address. Does returns to scale vary with employment or population? Does returns to scale vary with employment in your own industry? What about in industries that make your inputs and buy your output, or all industry? Does returns to scale vary with the total size of the city, or with the density of employment near a particular firm? Does increasing returns depend equally on all workers, or just those with particular attributes, like the college educated? Are different production activities subject to different degrees of returns to scale, e.g., new versus mature industries? Finally, do agglomeration economies affect different types of workers, e.g. college educated and not, differently?
5155

The answer to these questions, pretty broadly, seems to be “yes”. The framework we used to develop some intuition about agglomeration economies, and which gives rise to the wage regression of equation (8.4), is far too simple. There is not one
5160 agglomeration economy, there are many, and they probably don’t operate at the same strength or through the same mechanism in all cities or for all people.

We have a pretty good idea of the magnitude of σ for an average worker in an average city, and we know a little bit about how σ differs from industry to industry, worker type to worker type, and mechanism to mechanism. We now turn to a short

Figure 8.7: Plot of log country mean per capita GDP against percent urban



Note: Units of observation are countries. Data are from the World Development Indicators database. GDP per capita is measured in 2000 US dollars. Figure and note reproduced from Glaeser and Gottlieb [2009] figure 10, ©American Economic Association.

⁵¹⁶⁵ survey of these results.

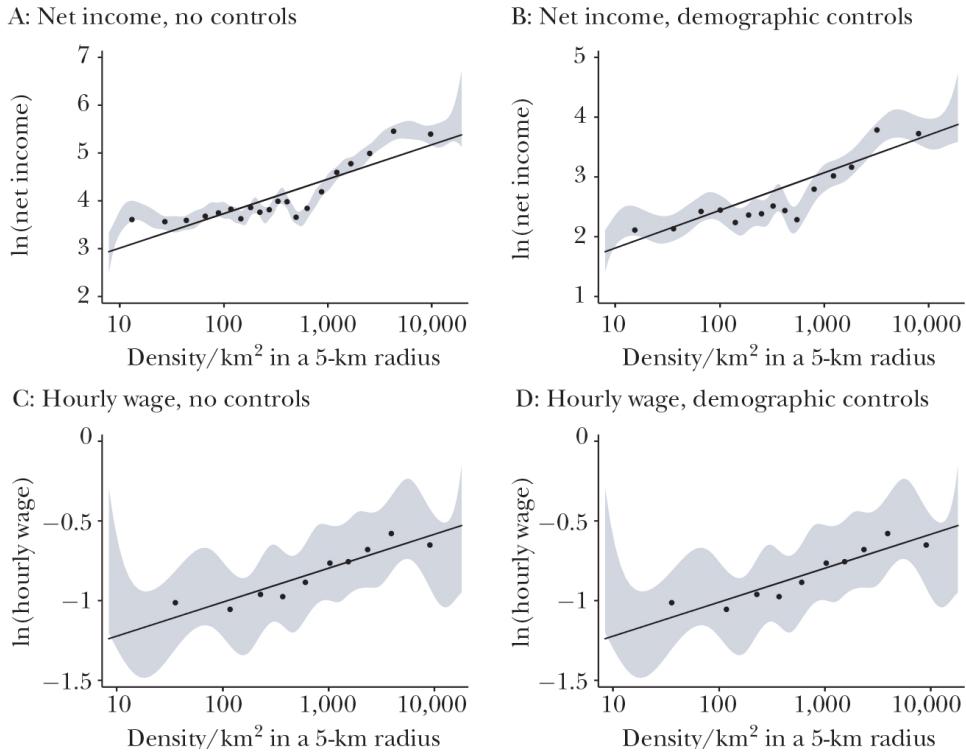
8.6 Measuring agglomeration economies: Empirical results

Figure 8.7 shows the relationship between urbanization and the level of per capita GDP in *country* level data. The x -axis reports the percentage of urban population.

⁵¹⁷⁰ The y -axis is log of per capita GDP. Countries with more people in cities are richer. Eyeballing this picture, going from 20% to 100% urban increases log per capita GDP from about 6 to about 10 on average.

To understand how big this is, let y_{60} and y_{100} be per capita income in a typical

Figure 8.8: Wages and household income as a function of nearby population density



Note: In the developing world, Africa in particular, an increase in density increases wages and household income. Household income increases faster with density. Figure reproduced from Henderson and Turner [2020], ©American Economic Association.

country with 60% and 100% urban share. From the previous paragraph, we have

5175 $\ln(y_{100}) \approx 10$ and $\ln(y_{60}) \approx 6$. It follows immediately that $y_{100} = e^{10}$ and $y_{60} = e^6$. Taking ratios, we have that $y_{100}/y_{60} = e^{10}/e^6 = e^4 \approx 54$. That is, on average, per capita income in a country that is about 100% urban is about 54 times that of a country that is 40% urban. Some caution is required in interpreting this result, however. It may be that countries become rich because they are urban, but it may 5180 also be that they become urban because they are rich.

Figure 8.8 reports estimates of agglomeration economies using survey data representing 40 developing world countries. In all four panels the x -axis reports the number of people in a disk with a 5km radius centered on the survey respondent. The x -axis is in log scale with each unit of displacement a factor of 10. The black dots in this figure are a histogram showing the y -axis value in a bin centered on the dot, and the shaded area describes the range of variation around these bin means.

In the bottom two panels, the y -axis is the log of the hourly wage. The bottom left panel is a plot of log wage against log population density, exactly the regression described by equation (8.4), but with nearby population density in place of city size. The black line plots this regression line and has slope $\sigma = 0.12$. Because both axes are in logarithms, this means that the wage elasticity of population density is 12%. Because the effect of density could reflect the sorting of people into more productive places, or of more productive people into larger cities, the same care is required in the interpretation of this result as for figure 8.7.

The bottom right panel of figure 8.8 is exactly the same as the bottom left, but addresses the possibility that more productive people migrate to cities by controlling for a short list of worker demographic characteristics. Loosely, this figure shows how wages change for an average person, relative to a person with the same demographics, as density increases. This decreases the wage elasticity of density to almost exactly 5%, the typical value for these sorts of wage regressions. Comparing the two bottom panels, the decrease in the estimate of σ when we add demographic controls suggests that of the 12% wage elasticity of density in the raw data, 7% is due to the fact that workers in cities are more productive people, and the remaining 5% could be caused by density. In fact, this is a bit too favorable to agglomeration economies. Our short

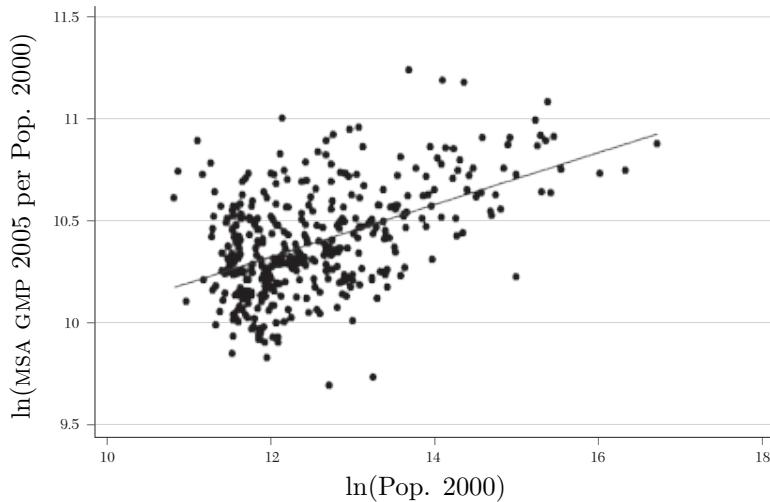
5205 list of controls does not include “ambition” or “dilligence”, and so the possibility remains that people with more of these traits sort into larger cities, and sorting on these unobserved traits accounts for some or all of the 5% elasticity we see in the bottom right panel of figure 8.8. In fact, there is an econometric tool for controlling for these sorts of unobserved time invariant individual traits, we’ll discuss it below,
5210 and they are generally not that important once we control for a short list of easily observed characteristics, like age and education.

The top two panels of figure 8.8 are just like the bottom two, but show the log of household income on the y -axis instead of wage. In the top left panel, the household income elasticity of density is about 31%. This is much larger than the corresponding
5215 wage elasticities, 12% and 5%, in the bottom panels. In addition, the relationship in the top left panel is unchanged when we add household demographic controls in the top right. This means that the relationship between density and household income is harder to explain away as resulting from the sorting of more productive households and workers into denser places. To my knowledge, this surprisingly steep
5220 relationship between household income and density has not been observed elsewhere in the literature, but seems consistent with the large gaps between rural and urban income that we see in the top panel of figure 4.8 . My best guess is that this reflects an increase in the participation of women in the labor force in cities, but this is just a guess.

5225 Rosenthal and Strange [2004] survey the literature that estimates σ . They find that whether N is employment or population, most studies that estimate equation (8.4) find that $\sigma \in [0.03, 0.08]$, with most estimates around 0.04.

To understand the importance of the wage elasticity of city size, consider the

Figure 8.9: Plot of log mean per capita MSA income against the log of city population.



Note: The log of per capita GDP increases with log of city population for US cities around 2005. Units of observation are Metropolitan Statistical Areas under the 2006 definitions. Population is from the Census. Gross Metropolitan Product is from the Bureau of Economic Analysis. Figure reproduced from Glaeser and Gottlieb [2009], ©American Economic Association.

implications of moving from a city of $N_0 = 10,000$ to one of $N_1 = 1,280,000 = 2^7 \times N_0$.

⁵²³⁰ Let w_0 and w_1 denote the corresponding wages. If $\sigma = 0.04$, then using equation (8.4) we have

$$\begin{aligned}\ln(w_0) &= \ln(AN_0^{0.04}) \\ \ln(w_1) &= \ln(A(2^7 N_0)^{0.04}).\end{aligned}$$

Using the rules of logarithms and doing a little algebra, this implies that $w_1/w_0 = 2^{7 \times 0.04} = 1.21$. That is, the central estimate of the strength of agglomeration economies implies that moving from a city of 10,000 to a city of 1,280,000 increases wages by a

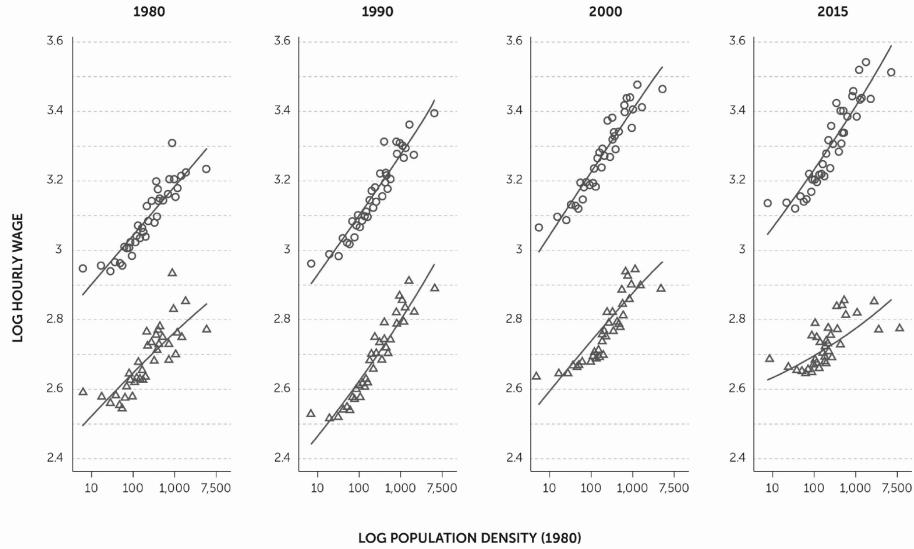
5235 little over 20% on average.

The estimation framework we've discussed so far, equation (8.4), does not allow for workers to work in different sectors. There are not service workers and tech workers, just workers. This is clearly not the case. For example, Henderson et al. [1995] estimate agglomeration effects by industry. They find that wage growth in an
5240 industry responds to (1) own industry employment, and to a measure of how diverse is employment in the city across sectors, and (2) that different industries respond differently.

Not only do agglomeration effects differ by industry, they also appear to differ by person and even over time. Figure 8.10 reports on the urban wage premium in the US
5245 for people with a college education and without. The x -axis in all panels is the log of population density in a neighborhood of each person. The y -axis is the log of the wage. The left-most panel is based on data from 1980 and the year advances to 2015 as we move through successive panels to the right. Each panel reports two plots. The upper plot describes the relationship between the log of wages and the log of density
5250 for college educated workers and the lower plot for workers who did not finish college.

Unsurprisingly, the wages for college educated workers are above those for the less educated. Looking a little more carefully, we also see that the gap between the two groups also grew during this period. However, the main feature of this series of plots is that the wage elasticity of density for college educated workers stays constant or
5255 increases over time, while the corresponding quantity for less educated workers gets smaller. In 1980, moving to a larger city resulted in about the same wage gains for college educated US workers as for less educated workers. By 2015, this was no longer true. In 2015, educated workers saw much larger increases in wages from moving to

Figure 8.10: Plots of log wage against log density for college graduates and non-college graduates by year



Note: *This remarkable plot shows the urban wage premium for college graduates (circles) and non-college graduates (triangles) over time in the US. The density premium for the wages of low skilled workers decreased dramatically relative to high skilled workers between 2000 and 2015. Figure reproduced from Autor [2020], courtesy of Aspen Strategy Group.*

a city than did the less educated.

The discussion above is sloppy about distinguishing between whether productivity varies with city size, as in figure 8.9 or population density near a worker, as in figures 8.8 and 8.10. A little more formally, whether σ is the wage elasticity of city size or nearby population density.

If we would like to tidy this up, we are led to ask questions about the scale at which agglomeration economies operate. Do increasing returns to scale operate at the scale of the whole city, or do they just depend on nearby density? If the latter, how near must people be?

We have some basis for answering these questions. Arzaghi and Henderson [2008] and Rosenthal and Strange [2003] both estimate how rapidly the effect of density falls off as it is further away. They find that (at least sometimes) agglomeration effects operate over very short spatial scales, about a mile in Rosenthal and Strange and a few hundred yards in Arzaghi and Henderson. Rosenthal and Strange [2004] surveys the literature on the topic. They find that when N is a measure of employment or population density near the worker (typically in the same county, or within a disk of radius 5 or 10 miles), then estimates of the wage elasticity regression,

$$\ln(w_{ij}) = B + \sigma \ln(N_j) + \varepsilon_{ij},$$

usually give estimates for σ that are between 0.04 and 0.05.

Does this mean that there are separate agglomeration effects for city size and density? Actually, no. A little bit of math shows that we can unify the two effects. The population of a city is the product of its area and population density. Letting a_j denote the area of city j and D_j its density, we can write

$$\begin{aligned} \ln(w_{ij}) &= B + \sigma \ln(N_j) + \varepsilon_{ij} \\ &= B + \sigma \ln(a_j D_j) + \varepsilon_{ij} \\ &= B + \sigma \ln(D_j) + \sigma \ln(a_j) + \varepsilon_{ij} \end{aligned} \tag{8.5}$$

So, if we regress log wages on city area and city population density, if the agglomeration effect is really from city size, we should find the coefficients on density and area are the same.

Using French data from 1976-1996, Combes et al. [2008] conduct this regression

5285 (more-or-less) and find that,

$$\ln(w_{ij}) = B + 0.037 \ln(D_j) + 0.011 \ln(a_j),$$

and the estimates are precise enough to reject the hypothesis that the coefficient on density and area are the same. This result suggests that the effect of population density is more important for worker productivity than is the total size of the city. If we double a city's population by doubling its area, density constant, wages increase by about 1%. If we double a city's population by doubling its density, area constant, wages increase by about 4%.³

Summing up, agglomeration economies look like they are about as complicated as can be. Their effect varies by industry, person and time, they vary with density and city size, and they are probably not quite the same from country to country.

5295 8.7 Mechanisms

Because productivity is so central to economic development and to how cities are organized, we would like to understand why cities contribute more to productivity as they are bigger. There are several candidate explanations. The first three are; labor market thickness, input market thickness, and knowledge spillovers.

5300 With more people, comes more workers, and so labor markets are bigger, “thicker”, in big cities. If you are an employer looking for someone with a particular skill, an expert in making molds for police badges or in urban economics, you have a better

³There is still a slight cheat here. Density at the individual level, as in figures 8.8 and 8.10 and the city average density, as in equation (8.5), are not quite the same thing.

chance of finding someone with exactly the right skills for your job in a bigger city. The better matching of skills to tasks, along with the reduction in search costs, should
5305 make people in bigger cities more productive.

Like labor markets, input markets are also more competitive and bigger in big cities. Thus, just as bigger cities make it easier to find the right worker, they also make it easier to find exactly the right input.

Finally, if you are trying to learn how to make a new product, or refine your
5310 process for making an old one, this will be easier if there are lots of people with specialized knowledge around for you to talk with, hire, or observe. This process is usually called “knowledge spillovers”.

One can imagine more reasons why city size might boost output, but these three mechanisms attract a lot of attention because they are consistent with the sort of
5315 increasing returns to scale aggregate technology described above: As the city gets bigger, labor market thickness, input market thickness, and knowledge spillovers should all increase productivity (if they are present).

Two other mechanisms are probably also important. First, a place might have some natural advantage, like a good natural harbor, that leads people to locate there.
5320 Second, a city might be productive just because it is a destination for particularly productive people or firms.

Figuring out which of these mechanisms are at work is important. Recall that when there is an externality, workers won’t capture the full benefit of their decision to urbanize, and so we should worry that big cities will be under-provided in equilibrium.
5325 If we think that the increasing relationship between city size and productivity occurs because of knowledge spillovers, or market thickness, then we need to worry about

this, and think about whether cities are too small in equilibrium. On the other hand, if productivity is higher in bigger cities because the people in these cities are more ambitious or skillful, or because the places they are located are different, we
5330 have one less reason to worry that spatial equilibrium will fail to lead to a desirable arrangement of people and economic activity across locations.

8.8 (Some of) what we know about mechanisms

Input sharing

One likely implication of thicker input markets is that firms will buy more of their
5335 inputs rather than make them in house. Holmes [1999] looks for evidence of this effect. To understand what he does, suppose a firm sells a widget for price p_{out} . To make this widget, the firm uses labor, capital, and p_{in} worth of intermediate inputs purchased from other firms. If firms in denser cities are more productive because they can specialize into smaller parts of the production process, then we should see that $\frac{p_{in}}{p_{out}}$
5340 increases with city size and density. Holmes does exactly this calculation and finds that this happens for most industries. Using US data, for a typical plant, increasing own industry employment in nearby counties from less than 500 to somewhere between 10,000 to 25,000, increases $\frac{p_{in}}{p_{out}}$ by about 0.03. Because labor is about half of all firm expenditure, this is a 6% increase in the share of purchased inputs.

5345 This seems like pretty strong evidence that firms are reorganizing in exactly the way we would expect if agglomeration economies arise because input markets get thicker with city size. That is, as the city gets bigger, there are more firms making widgets, and they demand more widget inputs. In response, we see the rise of a

network of increasingly specialized widget input suppliers who begin to take over
 5350 steps of the manufacturing process that widget manufacturers in smaller cities do in house.

This is suggestive, but stops short of providing direct evidence that widget manufacturers are more productive in larger cities because of thicker input markets. With that said, this result does let us conclude that the effect of input markets is at most
 5355 a small part of total agglomeration economies.

To see this, suppose that the 3% increase in purchased inputs results in a 3% decrease in the cost to produce a widget, as if all of the incremental purchased inputs were free. To accomplish this, we need to increase the number nearby own industry employees from 500 to at least 10,000. That is, density needs to increase by a factor
 5360 of 20.

Now let's see what a 5% wage elasticity of density implies about the change in wages when we increase the size of the city by this same factor of 20. Using equation (8.4), this means that with a size of 500, we have

$$\ln w_0 = A + .05 \ln(500)$$

and

$$\ln w_1 = A + .05 \ln(10000),$$

5365 Subtracting the first from the second and using rules of logarithms, we have

$$\ln(w_1/w_0) = 0.05 \ln(20) \approx 0.15.$$

Exponentiating, we have

$$w_1/w_o \approx \exp(0.15) \approx 1.16.$$

That is, if the wage elasticity of city size takes the 5% value that we've seen estimated above, then increasing city size by a factor of 20, as in Holmes's example, then we ought to see an increase in productivity of about 16%. But, our calculation suggests that productivity will increase by at most 3% because of thicker input markets.
5370 Therefore, while it is clear that thicker input markets affect firm behavior in bigger cities, with the estimates that we have in hand, it seems unlikely that they are the main reason for increasing returns to scale in cities.

Knowledge spillovers

Economists have been interested in knowledge spillovers at least since Marshall picturesquely described them in 1890,
5375

The mysteries of the trade become no mysteries; but are as it were in the air, and children learn many of them unconsciously. [Marshall, 2013]

This quote motivates many of the many papers trying to estimate agglomeration economies. This has always puzzled me. Taken out of context this way, this quote could as easily appear in Harry Potter as in a foundational text for the whole discipline of economics. In fact, Marshall had something much more specific and less magical in mind, the constant communication and observation that occurs among a group of people working nearby on more or less the same thing.
5380

With that said, while something like this is almost surely going on, the exact
5385

mechanism remains murky. Does it matter which industries you are near? Do you learn by talking to people at lunch? Hiring knowledgeable co-workers? There has been too much work on these questions to cover in detail, but the data make a case for some sort of knowledge spillovers.

5390 Moretti [2021] makes the case as well as we can hope for. He begins with data that reports the identity of inventors who file patents, their residential address, and the sector in which they work (e.g. Computer Science, Biology and Chemistry, Semiconductors). He is able to follow inventors over time, so he can see how their productivity varies as their environment changes from year to year.

5395 Moretti would like to ask what happens to the patent output, measured in patents per year, of an inventor when other inventors move nearby or when he or she moves to a city with more (or fewer) other inventors. To describe what he does, first define a cluster of inventors as the count of inventors in a city-sector-year. The size of the cluster plays the role of city size or density in our previous discussion of agglomeration economies and measures the scale of local inventive activity.

5400 Let i , j , and t index inventors, clusters, and years. Let y_{ijt} be patents filed, the outcome of interest. Next, let δ_i and γ_j be inventor and city fixed effects. This is conventional, sloppy, shorthand notation. δ_i really describes a list of variables, one for each inventor, that takes the value one when the inventor is inventor i , and zero for all the others. Similarly, γ_j is shorthand for a list of indicator variables that take the value one if the inventor in question is in city i and zero otherwise. Finally, let x_{jt} be the size of cluster j in year t .

Moretti runs (more-or-less) the following regression,

$$y_{ijt} = \delta_i + \gamma_j + \sigma x_{jt} + \varepsilon_{ijt} \quad (8.6)$$

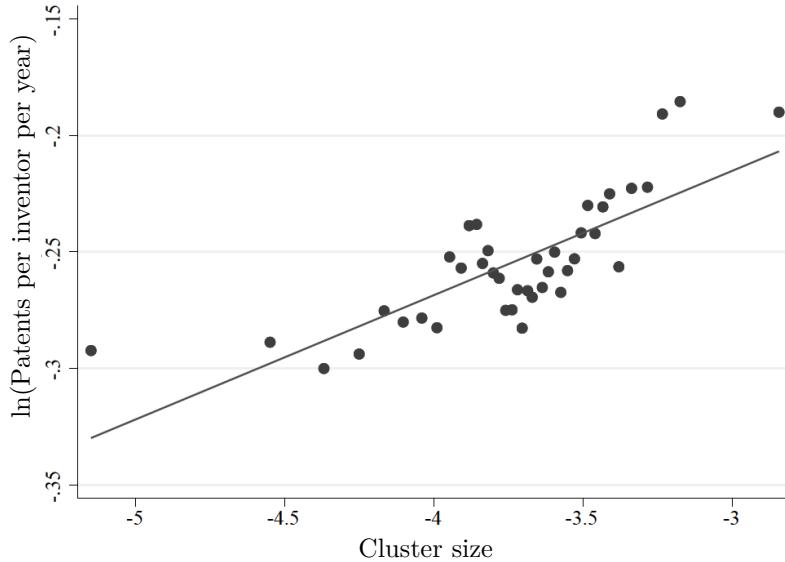
To understand this regression, take first differences. That is,

$$\begin{aligned} y_{ijt+1} &= \delta_i + \gamma_j + \sigma x_{jt+1} + \varepsilon_{ijt+1} \\ -y_{ijt} &= -\delta_i - \gamma_j - \sigma x_{jt} - \varepsilon_{ijt} \\ \Delta y_{ijt+1} &= \sigma \Delta x_{jt+1} + \Delta \varepsilon_{ijt+1}. \end{aligned} \quad (8.7)$$

Comparing equations (8.6) and (8.7), leads to two conclusions. First, while the second equation estimates fewer parameters than the first, the parameter σ in both equations is identical. We can estimate the patenting elasticity of cluster size in levels with equation (8.6), or in first differences with equation (8.7). Either way, we recover the same parameter. Second, when we take first differences, all time invariant individual and city characteristics cancel out. They don't change, and so they don't show up when we difference. In the bottom right panel of figure 8.8 we estimated the wage elasticity of city size, while controlling for observed individual demographics. If we have panel data, then we can estimate equation (8.7) and control for *all* time invariant individual characteristics, whether we observe them directly or not. The same comment applies to time invariant location characteristics. Thus, estimating a regression like equation (8.7) is an easy way to control for unobserved individual and location specific characteristics.

For an average inventor, patenting increases with the number of other inventors in that the same sector-city-year. Interestingly, in Moretti's preferred estimate, σ is

Figure 8.11: Plot of log of patents per inventor per year against log of cluster size in the late 20th century US.



Note: *y-axis shows city mean of log patents per inventor, conditional on year, sector and city. x-axis shows log of cluster size. Slope of the solid line 0.053, so that the cluster size elasticity of patents per inventor of 5.3%. Figure reproduced from Moretti [2021], ©American Economic Association.*

⁵⁴²⁵ about 0.07, just a little bigger than we usually find for wages. This looks like pretty strong evidence for some sort of “knowledge spillover”. Being around more inventors somehow makes inventors more productive, even if we don’t know quite how it works.

Sorting and learning

Consider again the difference between the bottom left and bottom right panels of ⁵⁴³⁰ figure 8.8. The bottom left plots log wages against log of nearby population density, and finds that the wage elasticity of density is about 12%. In the bottom right panel, we control for a short list of demographic characteristics, and find that the wage

elasticity of density drops to about 5%. This suggests that some of the relationship between wages and city size reflects the fact that people in denser places (or bigger cities) are different from those in less dense places.

This invites the question of whether city size would affect wages at all if we could control for all of the ways that people are different, e.g., ambition and diligence, not just the few ways that we can observe. The logic that equations (8.6) and (8.7) describe offers us a way to do just this. If we have panel data describing a person's wages and nearby density, then we can estimate a version of equation (8.7) to recover the wage elasticity of local density conditional on all unobserved time invariant attributes of people and the places that they live. The trick is to get this sort of individual level panel data.

This is exactly what is done in papers by Combes et al. [2008] and de la Roca and Puga [2017]. The first using data describing a panel of French workers, and the second using an almost identical panel describing Spanish workers.⁴ In a regression like the one reported in the bottom left panel of figure 8.8, they find a wage elasticity of density of about 5% (instead of the 12% in the figure). They then estimate the wage elasticity of density controlling for unobservable individual characteristics using a technique similar to the first differencing described in equation (8.7). They find a wage elasticity of density of about 3.7%. de la Roca and Puga [2017] also present two estimates the wage elasticity of density. The first, controls for a short list of demographic characteristics (like the bottom right panel of figure 8.8) and is about 4.5%. The second controls for all time invariant for unobservable as in equation (8.7) and is about 2.3%. That is, both papers find that much of the wage elasticity of

⁴In fact, Glaeser and Maré first looked at this question using US data in 2001 [Glaeser and Maré, 2001].

density is due to differences in unobservable time invariant attributes of workers.

Up to now, we have considered the possibility that workers may be more productive in cities because more productive people sort into cities. That is, workers in big cities are different from workers in small cities because they were different when they arrived in the city. de la Roca and Puga [2017] check whether something else entirely could be going on. It could be that people are the same when they arrive in cities but that they become different from each other as they spend more time in cities of different sizes. In this case, people in big cities are different from people in smaller cities, not because they sorted into locations on the basis of attributes related to productivity, but because people in bigger cities learn and acquire skills more quickly than in smaller cities.⁵⁴⁶⁰⁵⁴⁶⁵

Figure 8.12 summarizes de la Roca and Puga's main result. For reference, Madrid is Spain's largest metropolitan area, Sevilla is fourth, and Santiago is much smaller. The x -axis in the figure is years since the start of a person's career. The y -axis is their wage, really the average wage of someone with the same experience, relative to the wage of someone with the same experience who starts their career in Santiago and stays there. The figure describes the earnings path for four different people. The lower solid line describes the wage premium relative to Santiago of someone who moves to Seville at the start of their career and stays there for 10 years. The upper solid line describes the corresponding path for someone who moves to Madrid. The two dashed lines describe the wage premium for someone who starts their career in Sevilla or Madrid, spends five years there, and then moves to Santiago.⁵⁴⁷⁰⁵⁴⁷⁵

Three features of this graph are noteworthy. First, for a new worker, one with zero years of experience, there is a small premium for working in a bigger city, 3 or

5480 4% for Sevilla and almost 10% for Madrid. This probably reflects a pure size effect
and not sorting. Second, the wage premium relative to Santiago grows over time, and
grows faster in larger cities than smaller. Thus, after 10 years, the wage premium for
someone who has spent their whole career in Madrid is almost 35% and in Sevilla,
about 15%. Third, if someone moves from Madrid or Sevilla to Santiago after five
5485 years of experience, then their wage premium immediately drops by about the same
amount as the zero year worker gains, and then stays about constant. Duranton and
Puga [2023] replicate this result using a (small) sample of US workers, so this effect
is probably not a special feature of the Spanish economy.

It is hard to understand these patterns in the data as evidence of anything but the
5490 more rapid acquisition of skills in larger cities. Moreover, it turns out that seven years
of experience gives just about the same wage premium as we had previously attributed
to sorting. This suggests that indeed, people in big cities are more productive than
people in small cities because they are different from them. However, they are different
because they became different after they arrived in the city, not because they were
5495 different before they arrived.

This is an important result for at least two reasons. First, prior to this result, my
best guess was that a substantial fraction of the wage elasticity of city size reflected
the fact that more productive people sorted into bigger cities. This would mean that
much of the increase in wages is not actually due to city size at all, but rather to
5500 the characteristics of the residents of bigger cities. de la Roca and Puga require that
we revise this conclusion. People in bigger cities are indeed more productive than
those in smaller cities, but this is *probably* because bigger cities make them more

productive.⁵ This conclusion means that the city size externality is larger. People do not have a large enough incentive to move to cities because they do not account for
5505 the fact that their presence speeds the acquisition of skills by other city residents.

To understand the second problem, consider what would happen if everyone in Spain spent the first 5 years of their career in Madrid, before returning to their original hometown. In this case, Madrid would contribute to the average productivity of *everyone* in Spain. In this case, comparing the productivity of workers in Madrid to
5510 those elsewhere will underestimate the value of time in Madrid because the productivity of everyone in the comparison cities will reflect the skills acquired in Madrid. In this case, we will almost surely underestimate the benefits of time spent in big cities. Not everyone in Spain spends the first five years of their career in Madrid, but some people do, so it could be that something like this is going on. It is not yet clear the extent to
5515 which this concern should cause us to inflate existing estimates of the wage elasticity of city size.

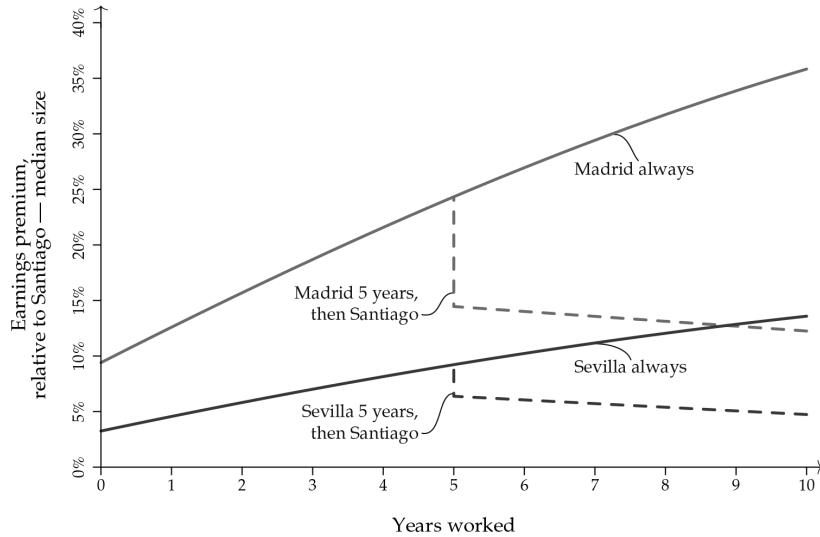
8.9 Conclusion

We now have the basis for answering the question that started this chapter, “why are there cities anyway?” People are willing to put up with commuting and higher rents
5520 in order to be close to one another. They want to be close to each other because it makes them more productive, and this productivity premium is increasing in the size of cities throughout the range observed city sizes.

Beyond this, things get complicated. The effects of agglomeration economies are

⁵It could still be the case that quick learners sort into bigger cities. In this case, we can explain the de la Roca and Puga as a consequence of sorting, too.

Figure 8.12: Path of earnings premium by experience of city size in Spanish data 2004-9.



Note: *x-axis is years of employment. y-axis is earnings premium as a percent relative to Santiago. The upper solid line describes the growth in earnings of a worker in Madrid as their time in Madrid increases. The lower solid line describes the same path for a worker in Sevilla. The upper dashed line describes the earnings premium for a worker who spends five years in Madrid before moving to Santiago. The lower dashed line is similar, but for a worker who moves from Sevilla to Santiago. That the worker Santiago with experience in Madrid earns more than the worker from Sevilla tells us that some of the wage increase in Madrid and Sevilla must be due to the accumulation of human capital by the worker.*

Figure reproduced from de la Roca and Puga [2017], ©Oxford University Press.

complicated and it probably makes sense to think about many agglomeration effects, not just one. In particular, agglomeration effects are probably: larger for density than city size; about the same for the developing and the developed world; larger for people who are more educated or have more experience; larger for knowledge intensive industries; about the same for levels and growth rates of productivity; contribute to human capital accumulation. With that said, if you were going to guess that the

5530 elasticity of whatever you output measure is to city size improved is about 5%, you would likely be within a factor of two of our best estimate, and probably much closer.

The mechanisms behind this have been the subject of speculation at least since Alfred Marshal in 1890. We have pretty good evidence for each of the following mechanisms for agglomeration economies: knowledge spillovers; labor market thickness; input market thickness; and sorting of high productivity people into larger cities
5535 (or the more rapid acquisition of skills in larger cities).

We do not have a good sense for the extent to which these mechanisms interact. For example, it is not clear if the estimated effects from each of the mechanism sums to more or less than the total agglomeration effect, or whether the effect of city size operates through different mechanisms than the effect of density. Given that empirical research is constrained by the fact that the samples of cities available for study number in the hundreds, it is not clear that it will be possible to make a lot of progress on these issues.
5540

Problems

5545 1. Let $Y = f(x) = x^\alpha$ describe the production process, where f is the production technology.

(a) Verify that f is increasing returns to scale if $\alpha > 1$.

(b) Verify that f is constant returns to scale if $\alpha = 1$.

(c) Verify that f is decreasing returns to scale if $\alpha < 1$.

5550 2. Suppose $f(n_i) = n_i^\alpha$ is decreasing returns to scale in individual labor n_i , and

output is given by

$$y_i(n_i) = AN^\sigma f(n_i), \text{ for } N = \sum_i n_i$$

Verify that

$$\frac{\partial y_i}{\partial n_i} \approx AN^\sigma f''(n_i)$$

as n_i gets small.

- 5555 3. Consider an economy with two firms (call them Firm 1 and Firm 2) choosing between three locations, A , B and C .

- 5560 (a) Create a table with all of the possible combinations of firm/location choice.
- (b) Assuming the firms are choosing location randomly, create a new table with the share of outcomes where one firm is in A and the other is in B , where both firms are in A ,etc., for each location pair you listed above.
- (c) Define pairwise distance, d_{ij} , to be 1 if the firms are in different locations, and 0 if the firms are in the same location. Add a column for pairwise distance to the table from the previous step.
- 5565 (d) Assuming that the firms choose location at random, plot three histograms about pairwise distances: First, if the firms are not in the same location, what are the relative frequencies of d_{ij} being 0 versus 1? Second, if the firms are in the same location, what are the relative frequencies of d_{ij} being 0 versus 1? Third, compute a weighted sum of these two histograms, weighing each by the relative frequency with which it occurs in your table, to create a “mean” histogram of pairwise distance.

- 5570 (e) Assume that you have data on 100 industries, each of which has two firms with three possible choices of location. You observe that in 50 industries, the two firms are in the same location, and in the other 50 industries, the firms are in different locations. Do you think the observed location choices are consistent with firms randomly choosing locations? Explain briefly.
- 5575 4. Suppose you observe only the part of the Basic Pharmaceuticals graphs from Duranton and Overman [2005] (right panel, figure 8.5) for pairwise distances between 88 and 92km. Should you conclude that pharmaceuticals are more agglomerated than would occur by chance? Explain briefly.
- 5580 5. Zipf's law (which we will encounter again later) tells us that the n^{th} largest city in a country is $\frac{1}{n}$ times as large as the largest city.
- (a) How many times would the 32nd largest city need to double to be the same size as the largest city?
- (b) Assume our current best estimate of agglomeration economies is $\sigma = 0.04$. How much more productive would we expect a unit of labor to be in the largest city than in the 32nd largest city?
5585

Chapter 9

Systems of Cities

Up until now, our focus has been on understanding the forces that affect the organization of a single city. Yet even a quick look at an image like figure 3.4 makes clear
5590 that this is only a small part of the larger geography of an economy. We now turn to thinking about this larger geography, and the forces that organize the geography of a region, or in the language we will use, a system of cities.

Our first task is simply descriptive. What does the size distribution of cities look like? How does it change over time? What are patterns of sectoral specialization?
5595 Do cities specialize in different activities? Does this change over time? This done, we ask whether we can explain these patterns as a consequence of spatial equilibrium.

9.1 Some basic facts about systems of cities #1

Consider a set of cities indexed by i and a set of industries indexed by j . Let s_{ij} be the share of industry j employment in city i . Define an index of specialization for

5600 city i as,

$$ZI_i \equiv \max_j(s_{ij}).$$

That is, the specialization of a city is the share of employment in its largest sector.

This seems pretty intuitive, but is subject to two problems. First, consider two cities, A and B. City A has employment equally divided between two sectors, and so $ZI_A = 0.50$. The second city has 51% employment in one sector and the remaining 49% evenly divided between 49 other sectors, and so $ZI_A = 0.51$. Thus, $ZI_B > ZI_A$. This is not obviously the commonsense way to rank the specialization of these two cities. There is no solution to this problem. Fundamentally, any index number, like ZI_i , is condensing many numbers into one, and so some information is lost, and one can usually construct a problem as a consequence.

5610 Second, it is also useful to consider relative specialization. Providence, Rhode Island MSA makes a large share of all submarines produced in the US. In spite of this, this sector accounts for a small share of total metropolitan employment. To measure this sort of relative specialization, let s_j be national employment in industry j and define the Relative Specialization Index,

$$RZI_i \equiv \max_j(s_{ij}/s_j).$$

5615 The submarine sector employs only a small share of the national workforce, and also a small share of the Providence workforce. However, if the submarine sector's share of Providence employment is large relative to its national share, we will conclude that Providence is relatively specialized in submarine construction.

Measuring diversity is trickier. What does it mean to say that one set of jobs is
 5620 “more different than another?” For example, consider, the two sets of jobs,

{2 mechanics, 1 brain surgeon}

{1 butcher, 1 baker, 1 candlestick maker}

Which set is more diverse?

The industry standard for answering this question is the “Herfindahl Index”,

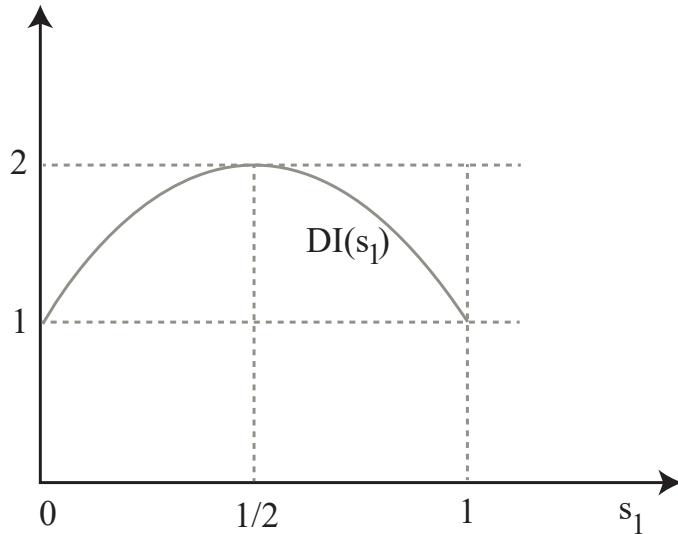
$$DI_i = \frac{1}{\sum_j s_{ij}^2}.$$

To see how the Herfindahl Index works, evaluate it when there are just two industries,
 $j = 1, 2$. In this case, we have $s_{i2} = 1 - s_{i1}$ and,

$$\begin{aligned} DI_i &= \frac{1}{\sum_{j=1}^2 s_{ij}^2} \\ &= (s_1^2 + s_2^2)^{-1} \\ &= (s_1^2 + (1 - s_1)^2)^{-1} \\ &= (2s_1^2 + 1 - 2s_1)^{-1}. \end{aligned} \tag{9.1}$$

5625 Figure 9.1 plots equation (9.1) as s_1 ranges from 0 to 1. With just two sectors, the index takes its largest value when employment is equally divided between the two sectors and takes its minimum values when employment is concentrated in a single sector. This intuition generalizes. With many sectors, the Herfindahl index increases as employment is distributed more evenly between sectors. So “most diverse” means

Figure 9.1: Herfindahl index with two industries



Note: When there are only two sectors, the sector 1 share determines the sector 2 share, and so we can evaluate the Herfindahl index if we know just the one share. The x-axis shows the share of employment in sector 1. The y-axis plots the resulting value of the Herfindahl index. The Herfindahl diversity index takes its maximum value when employment is evenly divided between the two sectors.

5630 employment is uniformly distributed across all sectors.

The corresponding relative diversity index is also sometimes useful,

$$RDI_i = \frac{1}{\sum_j (s_{ij} - s_j)^2}.$$

To understand the relative diversity index, consider what happens as a city's employment in every sector approaches the national average. As this happens, $(s_{ij} - s_j)$ goes to zero for every sector j and so the relative diversity index approaches infinity. For 5635 the relative diversity index, a city is maximally diverse when all employment shares equal the national average. For the regular diversity index, a city is maximally diverse

Table 9.1: Most and least specialized and diversified US cities in 1992

Rank	City(sector)	Specialization		Diversity	
		RZI	City	RDI	City
1	Richmond, VA(tobacco)	64.4	Cincinnati, OH	166.6	
2	Macon, GA(tobacco)	55.0	Oakland, CA	161.2	
3	Lewiston, ME(Leather)	49.6	Atlanta, GA	159.4	
4	Galveston, TX (petroleum)	49.1	Philadelphia, PA	151.4	

Note: *Table based on Duranton and Puga [2000].*

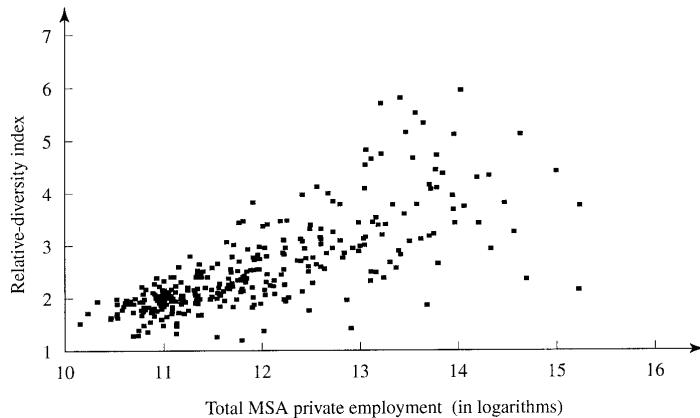
when all sector employment shares are equal.

Table 9.1 lists the four most diverse and the four most specialized US MSAs in 1992 on the basis of the relative diversity and relative specialization indexes. Richmond 5640 Virginia was the most specialized MSA and, not too surprisingly, was specialized in tobacco and tobacco products. Each of the other highly specialized MSAs is also a small town that is well known as the center of a particular industry. Cincinnati was the most diversified city in 1992. Cincinnati was about the 20th largest MSA in 1990 and is not specialized in any particular industry. Similarly, the other highly diversified 5645 cities are all large cities that are not well known producers of particular things. From Table 9.1 we take our first fact about systems of cities; specialized and diversified cities coexist.

Table 9.1 also suggests that bigger cities tend to be more diversified than small cities; the diversified cities are all large and the specialized cities are all small. Figure 5650 9.2 shows that this is true more generally. The horizontal axis in this figure is the log of MSA population, and the vertical axis is the Relative Diversity index. Large cities tend to be more diversified than small.

In fact, all cities have a large share of employment in non-traded sectors, e.g.,

Figure 9.2: Urban industrial diversity and population



Note: The figure describes US MSAs in 1992. The x-axis in the figure is the log of MSA employment. The y-axis is the Herfindahl index of diversity. The figure shows that bigger cities tend to be more diverse. Figure from Duranton and Puga [2000], ©Sage Publications.

services, so all cities are pretty diversified. However, MSAs with populations above
 5655 500k tend to have larger shares of their employment in business services; finance,
 insurance, and real estate, and smaller shares in manufacturing.

MSAs with populations between 50-500k tend to have larger employment shares in
 “mature industries” like “textiles”, less in “new industries”, like “instruments”. Big
 cities are relatively specialized in new industries.

5660 All together, there is evidence for cities as nurseries of firms and new industrial
 processes. New firms start in large, diverse, expensive places and once the process
 is established, migrate to smaller, less diverse, less expensive places. Tesla famously
 moved from the San Francisco to Austin in 2021. While the move was sometimes
 reported as Tesla’s repudiation of California’s political and regulatory environment,
 5665 this sort of move is a common part of the lifecycle of firms. Firms commonly develop

their products in big, expensive, diverse cities, and then move to smaller cities once they have figured everything out and want to make a lot of output cheaply.

We now turn to a consideration of how cities change over time. Table 9.2 lists the five largest US MSAs in 1997. The first column of the table presents the city's rank, and in parentheses, the change in its rank between 1977 and 1997. None of the four largest cities changed rank during this period, and Philadelphia, the fifth ranked city moved down one place to sixth by 1997 (replaced by San Francisco). New York has been the largest city in the country since colonial times, and Chicago was the second largest city from the end of the 19th century until the 1980s when it was overtaken by Los Angeles. The city size distribution, at least in terms of population rankings, is stable.

The second and third columns of table 9.2 report each cities ranking in terms of total employment in the manufacture of apparel and of transportation equipment (SIC codes 23 and 37), with changes in ranking in parentheses. Relative to total population, total employment by sector changes rapidly. Between 1977 and 1997, New York and Los Angeles switch place as the largest and second largest employer of apparel workers, Washington fell 7 places in the same ranking, and New York fell from being the third largest employer of makers of transportation equipment to 23rd. Cities change their relative populations slowly over time, but the way those populations are employed changes more rapidly.

Finally, note that a subset of the literature on agglomeration investigates the relationship between diversity, specialization and productivity. This literature generally concludes that knowledge intensive activities tend to be more productive in diverse cities, while established processes are more productive in specialized cities, that is, in

Table 9.2: Rankings and changes for the five largest US MSAs

Rank in 1977 (change in rank from 1977 to 1997)	Total population	Apparel	Transportation Equipment
New York	1(0)	1(+1)	3(+20)
Los Angeles	2(0)	2(-1)	2(0)
Chicago	3(0)	11(-3)	8(+4)
Washington	4(0)	13(+7)	29(+8)
Philadelphia	5(+1)	3(+3)	13(-3)

Note: *The first number in each column gives the MSAs 1977 ranking among all 1977 MSAs in the us. The parenthetical number gives the change in rank between 1977 and 1997.*

Table based on Duranton [2007].

5690 factory towns.

9.2 Zipf's law

Zipf's law is really a hypothesis about the size distribution of cities, sometimes also known as the Rank-Size Rule. Formally, Zipf's law states that the size distribution of cities follows a Pareto distribution with exponent equal to one. That is, if city's size is N and the rank of the city in the set of cities under consideration (usually a country) is $r(N)$, then

$$\ln r(N) = \ln A - \zeta \ln N. \quad (9.2)$$

ζ is called the “Zipf's coefficient”, and Zipf's law states that equation (9.2) holds and $\zeta = 1$.

We can now see why this relationship is often called a rank-size rule. Let N_1 and

5700 N_2 be the populations of the largest and second ranked cities. Evaluating equation (9.2) for the largest city, we have

$$\ln 1 = \ln A - \ln N_1^\zeta. \quad (9.3)$$

Similarly, for the second largest city, we have

$$\ln 2 = \ln A - \ln N_2^\zeta \quad (9.4)$$

Subtracting equation (9.4) from (9.3) gives

$$\ln \frac{1}{2} = \ln \left(\frac{N_2}{N_1} \right)^\zeta.$$

Exponentiating, we have

$$\frac{1}{2} = \left(\frac{N_2}{N_1} \right)^\zeta. \quad (9.5)$$

5705 Therefore, when $\zeta = 1$, equation (9.2) requires that the second ranked city be half the size of the largest. If we apply this same logic to the first and third ranked city, we have that the third ranked city is one third the size of the largest city, and so on as we work our way through from largest to smallest.

This is an odd property, and not one that we would expect to be true. Here, 5710 the obscure way in which we originally state the result in equation (9.2) shows its usefulness. We can easily test whether equation (9.2) holds for the observed size

distribution of cities by conducting the regression,

$$\ln r(N) = \ln A - \zeta \ln N + \varepsilon, \quad (9.6)$$

and then checking if we get $\zeta = 1$, or more precisely, checking if we fail to reject that $\zeta = 1$.

5715 Figure 9.3 reports the results of this regression using data describing the 135 largest US MSAs in 1990. The x -axis in this figure is the log of MSA population (in thousands) and the y -axis is the log of each city's ranking. New York City is the largest city. Its rank is equal to one, and the log of its rank is $\ln 1 = 0$. The population of the New York MSA in 1990 is about 17.8 million, or 17,800 thousands.

5720 Because $\ln 17,800 \approx 9.8$, the marker representing New York is (barely visible) at $(x, y) = (9.8, 0)$ in the figure. The second largest MSA, Los Angeles, is just above and to the left.

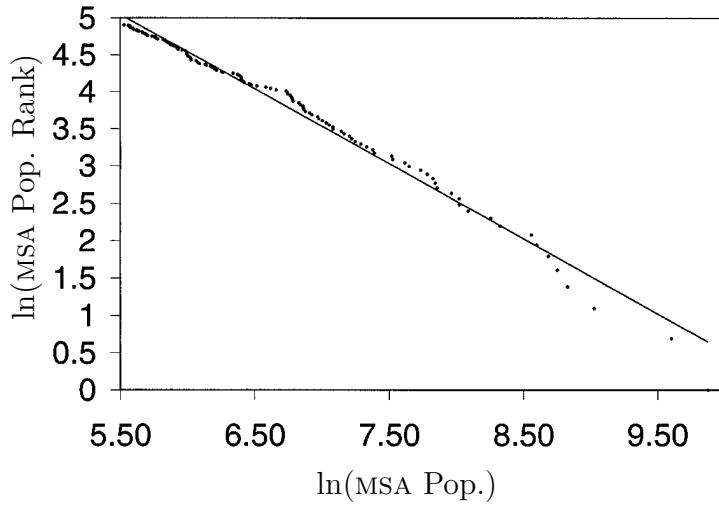
Figure 9.3 looks pretty good for Zipf's law. The data, 135 pairs of log population and log rank, line up neatly along the line

$$\ln r(N) = 10.53 - 1.005 \ln N.$$

5725 Here, the coefficient on $\ln N$ is ζ from equation (9.2), and at 1.005, it is close to one, as required by Zipf's law. Figure 9.3 is remarkable. The data all lines up almost perfectly along a line with slope minus one. This level of orderliness is just not something that we see often.

This looks pretty good for Zipf's law, and it is hard not to suspect that this 5730 figure tells us something important about how cities grow. Moreover, because urban

Figure 9.3: Zipf's law in US data in 1990



Note: Plot shows the log of MSA population rank on the y-axis and the log of population on the x-axis. Thus, New York City, the largest, first ranked city, is the data point at the bottom right of the plot. The trend line is the result of the regression in equation (9.6). The equation of this regression line is $\ln r = 10.53 - 1.005 \ln N$. That is, the estimated value of ζ is close to one, as required by Zipf's law. Figure from Gabaix [1999], ©Oxford University Press.

growth is pretty clearly central to the process of economic growth and development, it surely also tells us something important about the process of economic growth more generally. Stating this a little more precisely, figure 9.3 suggest that we can reject any model that describes how cities and economies grow if it does not predict a city size distribution that satisfies Zipf's law (and so looks like figure 9.3). Because Zipf's law is such a specific prediction, this should let us eliminate a lot of models.
5735

Gibrat's law is a cousin of Zipf's law, and it is a hypothesis about how cities grow. Gibrat's Law states that the each city grows at a random multiplicative rate in every period, regardless of city size. Ioannides and Overman [2003] perform a

5740 careful analysis of city level data in the US from 1900 to 2000 and conclude that, in their sample of large US cities, Gibrat's law and Zipf's law both seem to hold. In a remarkable paper, Gabaix [1999] shows that if that cities grow by a random multiplicative share, i.e. Gibrat's law, and in addition, the size of each city is bounded below at some strictly positive minimum, then the resulting distribution of city sizes
5745 satisfies Zipf's law.

Some caution is in order here, however. Figure 9.3 really looks to good to be true. In fact, it results from plotting one function of city size, log of population, against a decreasing function of size, rank, on the other. This rules out some of the disorderliness we see in figures like figure 8.9 by construction. This relationship *must* be decreasing. Worse still, while Zipf's law seems to hold pretty well for large cities
5750 in the US, it does not seem to hold for smaller cities in the US nor for many other countries in the world.

Holmes and Lee [2010] ask two questions about Zipf's law. First, whether it holds if we extend consideration to US cities smaller than 135 largest MSAs. Second,
5755 whether figure 9.3 could be an artifact of how MSA boundaries are drawn. Maybe the remarkable linearity that we see in figure 9.3 is telling us about the rule that the US Census Bureau uses to define MSAs, rather than about how cities grow? To investigate both questions, Holmes and Lee check whether the rank size rule applies to 85,287 six mile squares drawn on a regular grid covering the continental US in 2000.

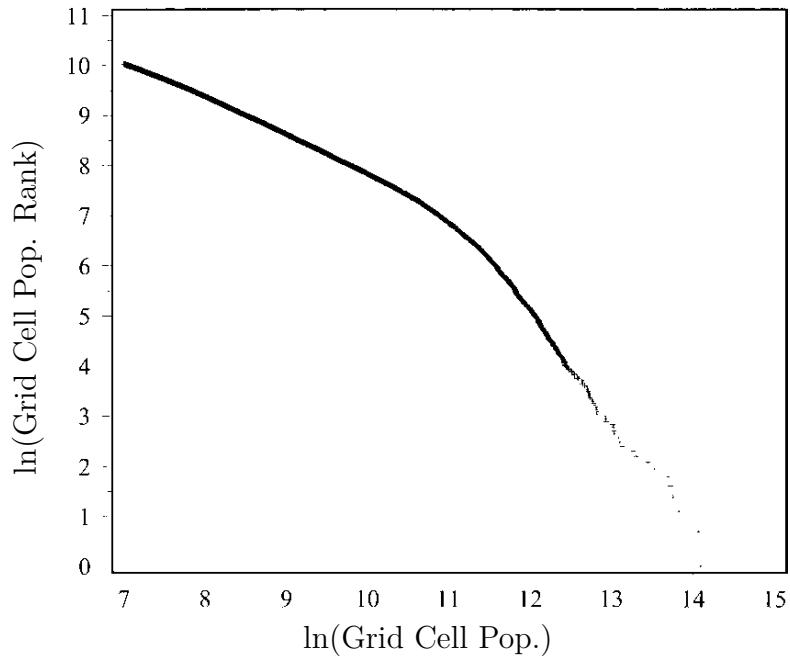
5760 Where figure 9.3 describes the relationship between the log of the size and the log of the rank of the largest 135 MSAs in 1990, figure 9.4 describes the relationship between the log of the size and the log of the rank of 23,974 six mile square grid cells in the US with a population above 1000 in 2000.

Figure 9.4 does not look like figure 9.3. It shows that there is a break in the rank-size relationship for six mile square pixels that divides pixels with log population larger and smaller than 11. In fact, Holmes and Lee [2010] estimate that for the larger pixels, the slope of the rank-size relationship is -1.94, and that for the smaller pixels, it is -0.75. This is less favorable for Zipf's law.

The statement of Zipf's law does not give us any guidance about the spatial units to which it should apply. Should it be MSAs, municipalities, water management districts, or six mile square cells? In the absence of some basis for saying that Zipf's law should apply to some spatial units and not others, if we see Zipf's law apply to MSAs and not six mile square cells, we have no way to know which case we should treat as the test of the law. Thus, taken together, figures 9.3 and 9.4 are contradictory evidence and do not let us reach a conclusion.

Soo [2005] has more bad news for Zipf's law. This paper estimates regression (9.6) for the 73 countries where the author could get population data at the metropolitan or city level. For most countries, he finds that ζ is different from one, so the rank-size rule does not hold. Of the 73 countries in his sample, there are about 20 where the Zipf coefficient is clearly less than one and about 30 where it is clearly greater than one. This leaves only about 20 where it is close to one.

Together with the findings in Holmes and Lee [2010], the balance of evidence suggests to me that Zipf's law is not a law at all, and figure 9.3 is just one of those occasional events where a more or less random process ends up looking very orderly.

Figure 9.4: Zipf's law for 6×6 mile cells in the US

Note: Plot of rank size rule exactly corresponding to the 135 largest MSAs used in figure 9.3, except that the spatial units come from the set of all 23,974 six mile square grid cells in the US with a population above 1000 instead of the 135 largest MSAs. Holmes and Lee estimate that for cells with log population above 11, the line of best fit has slope -1.94, and for cells with less population, the line of best fit has slope 0.75. Figure from Holmes and Lee [2010].

5785 9.3 Some basic facts about systems of cities #2

Recalling the way we defined Zipf's law, we see that even if Zipf's law rarely holds, we can use the expression that defines it to define a useful index describing how centralized or decentralized is a *system* of cities.

Using equations (9.2) and (9.5), we saw that if $\zeta = 1$ then the second largest city 5790 is half the size of the largest, the third is one third the size of the largest, and so on.

What happens $\zeta > 1$? To illustrate ideas, suppose $\zeta = 2$. In this case, (9.5) becomes,

$$\frac{1}{2} = \left(\frac{N_2}{N_1} \right)^2.$$

Reorganizing gives

$$\frac{\sqrt{2}}{2} = \frac{N_2}{N_1}.$$

Because $\frac{\sqrt{2}}{2} > \frac{1}{2}$, this means that the second largest city is more than half as large as the largest city, and using the same logic, the third city is more than one third as large as the largest, and so on. Therefore, if $\zeta = 2$, population is less concentrated in the largest cities than if $\zeta = 1$. More generally, as ζ gets larger, population is spread more evenly across cities, and as it gets smaller, it is more concentrated in the largest cities.
5795

If we notice that ζ has a minus sign in front of it, this means that as population decentralizes in a system of cities, ζ increases and the slope of the relationship between log rank and log population becomes more negative, i.e., steeper. Thus, we can look at a graph like figure 9.3 for two countries, and by comparing the slope of the relationship between log rank and log population, learn which of the two countries has its people more concentrated in its largest cities. The country with the flatter negative relationship has its population more concentrated.
5800
5805

Once we recognize that we can use ζ as a measure of how concentrated people are in the largest cities in a country, we are led immediately to two other questions. First, are there other indexes that we can use to measure this sort of concentration? Second, given some index of concentration like ζ , can we explain why some countries

5810 have their people more concentrated in their largest cities, and others do not? The answer to both questions is “yes”, and there is a small literature organized around these two questions.

There are a number of competing ways to measure how concentrated population is in a country’s largest cities. Most simply, “primacy” is the share of population in 5815 a country’s largest city. Slightly more difficult is just the variance of city size. When variance is zero, all cities are the same size, and as it increases, on averages city sizes are more different from each other. Finally, the Herfindahl index, described earlier in this chapter, can be applied to city sizes. As for the employment shares we discussed earlier, the Herfindahl index takes its maximum value when all cities in a country are 5820 the same size, and decreases when city sizes are more diverse. There is no particular reason to prefer one of these indexes over the other, they all measure more-or-less the same thing, but what research there is on city size distributions has relied primarily on the Zipf coefficient, ζ , and primacy.

Rosen and Resnick [1980] considers a sample of 44 countries around 1970 and 5825 finds that the Zipf’s coefficient decreases and then increases with country area, and increases with per capita GDP. The Zipf’s coefficient decreases with the extent of railway mileage density, but the effect is small enough that it cannot be distinguished from zero. Rosen and Resnick [1980] repeats their analysis using primacy as the outcome, and confirms the results they obtained for Zipf’s law.

5830 A caveat is required here. Rosen and Resnick [1980] establishes that there is a relationship, for example, between per capita GDP and the Zipf’s coefficient, but does not establish a direction of causation. It could be that the centralization of population decreases because GDP rises. On the other hand, the direction of causation could go

the other way.¹ This issue arises in all of the papers in this small literature, and so
5835 we need to interpret all of these results with the same caution.

Ades and Glaeser [1995] repeat the Rosen and Resnick [1980] investigation of primacy using a sample of 85 countries. They find that the population share of a primate city increases under dictatorships. They also find that primate share decreases with trade, with better developed road networks, and with larger shares of agricultural
5840 employment. Soo [2005] updates the work of Rosen and Resnick [1980] and using a sample of 44 countries. He finds that a higher density of roads and higher per capita GDP is associated with a higher Zipf's coefficient (and so a more even distribution of city population sizes). Soo [2005] also finds that the Zipf's coefficient decreases in countries with a history of political instability.

5845 Finally, Ioannides et al. [2008] investigate the extent to which information and communications technology are related to the arrangement of people across cities within a country using the Zipf's coefficient. They find that the Zipf's coefficient increases in magnitude as the log of telephone lines per capita increases. That greater telephone access decentralizes the population suggests an important role for trans-
5850 portation costs.

Summing up, our understanding of what determines how people are distributed across cities within a country is still pretty rudimentary. There is suggestive evidence that people concentrate in a smaller number of cities as; transportation costs (broadly defined) increase; as trade is less important or more costly; and, as government is more
5855 dictatorial. However, for all of these results, we can't have much confidence in the direction causation.

¹Indeed, [Henderson, 2003] estimates the relationship between primacy and GDP and finds that GDP is largest at an intermediate value of primacy. The same caveat applies here, too.

9.4 Path dependence and the locations of cities

We have so far considered the sizes and specializations of cities, but nothing of their locations. Cities tend to arise where there is some natural advantage. Harbors are the
5860 obvious example. New York City is where it is because it offers a safe natural harbor for ocean going ships, and is at the mouth of the Hudson River, which connects the city by water most of the way to Chicago.

Beyond this, the literature suggests that cities arise in a place as a result of some local advantage that is important *at the time of their founding*. However, because
5865 cities are so persistent, and technology changes so fast, when viewed through the long lens of history, these location choices start to look pretty random.

Bleakley and Lin [2012] demonstrate this for cities in the Southern US. Their paper begins with the observation, first made more than 100 years ago, that

[i]n the interior[South] the principal group of trade centers ... were those
5870 located at the head of navigation, or “fall line,” on the larger rivers. To these points the planters and farmers brought their output for shipment, and there they procured their varied supplies... It was a great convenience to the producer to be able to sell his crop and buy his goods in the same market. Thus the towns at the heads of navigation grew into marked
5875 importance ... (Philips (1905) quoted in Bleakley and Lin [2012])

That is, the fall line is a geological feature that was big enough to require traders go ashore and to carry their boats and goods around it, but not otherwise noteworthy. During the 18th and 19th century, towns arose at the portage sites where people would trade the goods conveniently already unpacked and on shore. Bleakley and Lin show

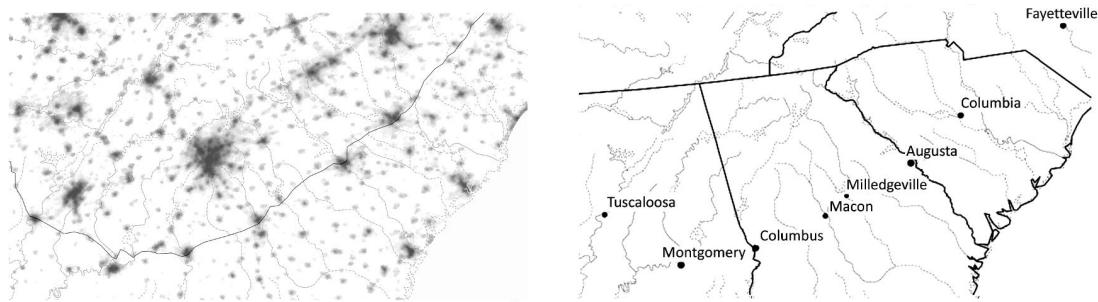
5880 that many of these towns persist today.

This result is easily visible on the two maps shown in figure 9.5. Both panels show the same area, most of the southern states of Alabama, Georgia and South Carolina, and show rivers as light gray lines. The left panel shows lights at night. Here, darker areas indicate places that are more brightly lit, and thus the location and extent of 5885 cities in the region. The thin positively sloped black line traces out the fall line. The right panel is a map showing state boundaries as heavy black lines and fall line cities as named dots. Note that state borders coincide with rivers in the cases of Augusta and Columbus.

Many modern cities occur at places where a river crosses the fall line. This is 5890 a striking result. Since these towns were founded, the importance of moving goods on inland waterways has declined dramatically even as the technology for doing so changed to no longer involve carrying boats around the fall line. And yet these towns remain as concentrations of economic activity, 140 years after the railroad came to dominate freight carriage and 80 years after trucks began to displace rail.

Henderson et al. [2018b] makes much the same point. They show that developing 5895 world cities tend to be closer to ports, while developed world cities tend to be closer productive agricultural land. This seems to reflect much the same intuition as we see in Bleakley and Lin [2012]. Cities in the developed world tend to be older than in the developing world, which means that they were built in a time when it was more 5900 difficult and expensive to move agricultural produce. Cities in the developing world were built when the cost of moving agricultural produce is much lower, and so the advantage of being near productive agricultural land is outweighed by the advantage of proximity to ports on the coast, and the international trade that these ports facilitate.

Figure 9.5: The fall line in two maps of the US South.



Note: *Left image is lights at night in 2003. In this image the thin black line describes the “fall line”, a minor geological feature just large enough to require colonial era traders to unload their boats and carry their goods and boats around it. The right panel shows locations of modern cities. Many cities started along the fall line, before railroads, are still important places today. Figure reproduced from Bleakley and Lin [2012], ©Oxford University Press.*

That is, the locations of cities in developed world countries are determined in part by their proximity to good agricultural land, even though agricultural is now a small share of all economic activity in most developed countries.

5905

Michaels and Rauch [2018] also reach a similar conclusion. Both England and France were colonized by Rome early in the first millennium C.E., and most cities during the time of Roman rule were Roman garrison towns. When the Roman empire fell in the fourth century C.E., England fell into disorder and cities in England were abandoned. This did not happen to the same extent in France. As a result, modern French cities are more likely to be built on Roman foundations, while English cities are more likely to date to the period from the period after about 650 C.E. This effectively gave England a chance to ‘reset’ the locations of its cities and this reset was not available in France. As a result, we see that English cities are more likely

5915

than French cities to be near an inland waterway and less likely to be near the coast. Again, we see that the locations of cities seem to be determined by factors that were important at the time of their founding, convenient places for a Roman garrison town, and proximity to an inland waterway, even if those factors are no longer important.

5920 9.5 Systems of cities

We have established several facts about cities. The overall distribution of city sizes more-or-less follows Zipf's law, at least for the largest cities. We know something about patterns of sectoral diversity and specialization across cities, and about how rapidly city size and sectoral specialization changes. Finally, we know a little bit 5925 about why cities are where they are. We'd like to develop a theory that lets us organize all of this. The first step is a model of "systems of cities", in which to think about how a population distributes itself across many potential city locations. One of the first formal statements of this problem is due to Henderson [1974], and the discussion here loosely follows his paper.

5930 To start, we need a simple description of how the costs and benefits of cities vary with their size. For this purpose, "benefits" are "output" and are subject to increasing returns to scale in production, "costs" are only commuting costs and the opportunity cost of labor (or maybe more accurately, the opportunity cost of rural residence).

We borrow the notation we used to describe agglomeration economies in Chapter 5935 8. Output is Y , population is N , and each person supplies one unit of labor so that

labor and population coincide. Total output is,

$$Y = AN^{1+\sigma}, \quad (9.7)$$

where σ is the wage elasticity of city size, the agglomeration effect.

The labor market is competitive and the wage is,

$$w = AN^\sigma. \quad (9.8)$$

Recall, that workers in a competitive labor market ignore their effect on aggregate population, and so a competitive labor market leads to w_i being the average product of labor, not its marginal product.

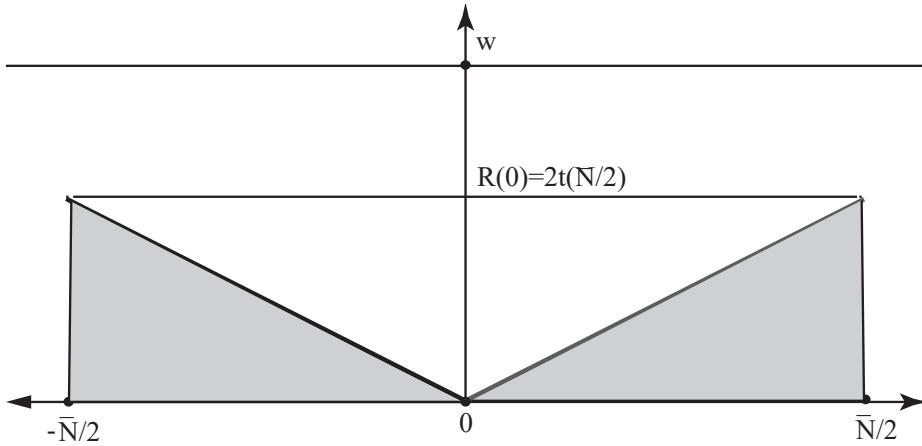
Next, recall the monocentric city model without housing from Chapter 1. People consume a fixed amount $\bar{\ell}$ of land, and to lighten notation, $\bar{\ell} = 1$. Agricultural land rent, \bar{R} , is equal to zero. Together, this requires that the length of the city is equal to its population or $N = 2\bar{x}$ (see Chapter 1).

Commuting costs are t per unit distance so the cost to commute from location x is $2t|x|$. Total commute costs are the sum of this cost over all occupied locations. Figure 9.6 illustrates this area. We would like to calculate both total and average commuting costs.

To calculate aggregate commute costs for a city of size N , integrate the shaded area in figure 9.6. That is,

$$TC(N) = 2 \int_0^{N/2} 2txdx = \frac{t}{2}N^2. \quad (9.9)$$

Figure 9.6: Calculating total commute cost in the monocentric city model



Note: Horizontal axis is displacement from the CBD. Vertical axis is dollars. The diagonal lines have slope $+/- 2t$ and describe the cost to commute to the center from each location. We calculate total commute cost for the city by evaluating the size of the shaded area.

Alternatively, recall that the area of a triangle is $1/2 \times \text{width} \times \text{height}$. This gives

$$TC(N) = 2 \times \frac{1}{2} \times \frac{N}{2} \times 2t\left(\frac{N}{2}\right) = \frac{t}{2}N^2.$$

It follows that average commuting cost is

$$AC(N) = \frac{TC(N)}{N} = \frac{t}{2}N. \quad (9.10)$$

We can now consider two different processes for assigning people to cities varies.

- 5955 The first is spatial equilibrium. This is familiar. We want to know how many cities we will have, and how large, if we consider allocations of people across cities such that no one wants to move and wages are determined competitively. The second is the planner's solution. Here we imagine a fictional planner who can dictate locations and wants to allocate people across cities to maximize a measure of aggregate welfare, 5960 $W(N)$, that we define below.

9.5.1 Spatial equilibrium

Start with spatial equilibrium. Make the following three assumptions. First, reservation consumption is \bar{c} and reservation utility is $\bar{u} = u(\bar{c})$. This is what a rural household gets. We imagine that the pool of rural residents is large compared to the population of cities. There is no crowding or returns to scale in the rural areas, so all rural residents get the same payoff, no matter how many of them there are. Second, all urban households have the average commute cost for their city. Third, aggregate land rent is divided evenly between all city residents (instead of going to absentee landlords). These are simplifying assumptions. They relieve us of having to keep track of individual incomes, rents, and commute distances.

Using these assumptions, we can calculate household consumption as a function of city size. Consumption is the difference between wages and average commute costs,

$$c_E(N) = w(N) - AC(N).$$

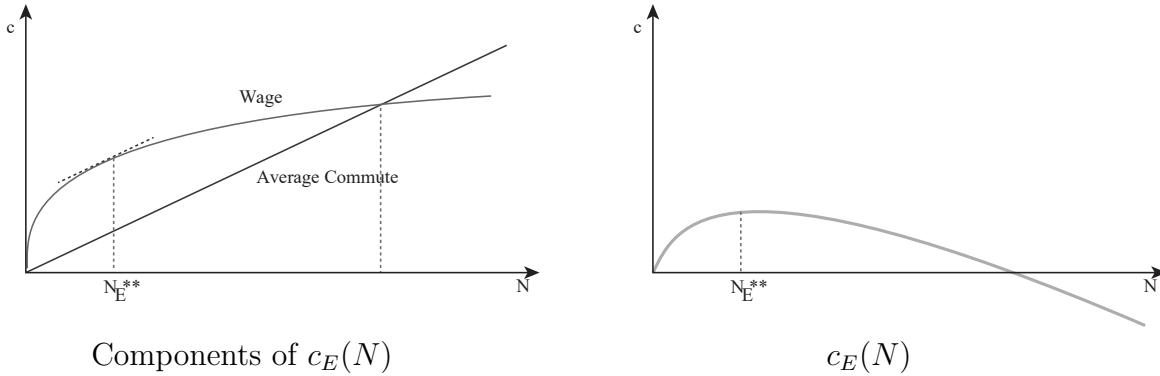
Substituting from equations (9.8) and (9.10), this becomes,

$$c_E(N) = AN^\sigma - \frac{t}{2}N. \quad (9.11)$$

That is, the consumption of a city resident is equal to the average product of labor

minus average commute cost.

Note that land rent does not appear in equation (9.11). Because each household has the same commute costs, each pays the same rent (which is trivially equal to the average rent). Because each household receives an equal share of the total land rent,

Figure 9.7: Graphical evaluation of $c_E(N)$ 

Note: Left panel plots AN^σ as the concave gray line, and $\frac{t}{2}N$ as the straight, positively sloped black line. The right panel plots their difference. That is, $AN^\sigma - \frac{t}{2}N$. Recalling equation (9.11), the right panel plots how the consumption level of an average urban resident varies with city size.

rent paid and rent revenue received cancel out.

In a spatial equilibrium, we choose N so that no one wants to move between the countryside and the city. This means that in spatial equilibrium, we have

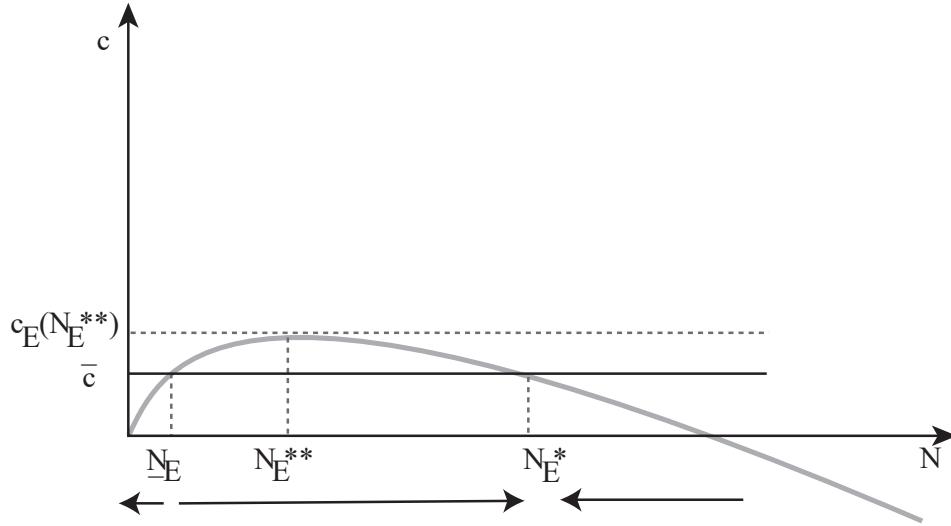
$$c_E(N) = \bar{c}.$$

Substituting from equation (9.11) gives,

$$\bar{c} = AN^\sigma - \frac{t}{2}N. \quad (9.12)$$

Equation (9.12) describes equilibrium city size. The right hand side gives the average consumption of an urban resident as a function of city size. The left hand side gives the consumption of a rural resident. When city size satisfies this equation then no

Figure 9.8: Illustration of spatial equilibrium for a single city



Note: This figure illustrates equilibrium city size. The curved line describes urban consumption, $c_E(N)$, from figure 9.7 and equation (9.11). The black horizontal line gives the rural consumption level, \bar{c} . An equilibrium occurs in each of the two places where these two lines cross. Arrows illustrate how we expect the system to evolve in response to a perturbation away from either of the two equilibria. For the smaller of the two equilibria, N_E , a perturbation causes a divergence from the equilibria. An equilibrium city of this size is unstable. For the larger of the two equilibria, N_E^* , the city returns to N_E^* after a shock. This is a stable equilibrium.

one wants to move and we have an equilibrium city.

Equation (9.12) is a non-linear equation and solving it analytically for city size is not easy. Fortunately, a graphical solution is straightforward and is illustrated in figures 9.7 and 9.8. The left panel of figure 9.7 plots the two components of the left side of equation (9.12), and the right panel plots the difference between these two components, that is, $c_E(N) = AN^\sigma - \frac{t}{2}N$. This curve describes the consumption level of a city resident as the size of the city varies, with N_E^{**} the city size that maximizes

urban consumption.

Taking an equation much like equation (9.11) as the starting point for their analysis
 5995 Au and Henderson [2006] use data describing Chinese cities to estimate the relationship between city size and worker output net of commuting. They find that this relationship has the same concave form as the one described by equation (9.11). As stylized as this model is, it seems to be capturing the basic facts about how cities operate.

6000 Figure 9.8 illustrates equilibrium city size. The curved line in the figure reproduces the graph of $c_E(N)$ from figure 9.7. The horizontal line gives \bar{c} , the rural consumption level. An equilibrium occurs when the two lines intersect.

In figure 9.8 there are two possible equilibrium city sizes, \underline{N}_E and N_E^* . This is not the only possibility, however. As \bar{c} increases, \underline{N}_E and N_E^* get closer and closer to each
 6005 other, until, when $\bar{c} = c_E(N^{**})$, there is exactly one solution. This unique equilibrium occurs when city size takes the value that maximizes urban consumption, N_E^{**} . If \bar{c} increases even further, there is no way for the city to match rural consumption levels, and the city is abandoned.

We are most interested in the general case where there are two solutions, but we
 6010 would like to know which solution will occur. There is some reason to think that we will only observe the larger of the two possible equilibria, N_E^* . To see why, suppose we start at the smaller of the two equilibria, \underline{N}_E and that the city experiences a small decrease in its population, ε . Then $c_E(\underline{N}_E - \varepsilon) < \bar{c}$, so after the decrease, the city is worse than the countryside. So once one person leaves, everyone leaves. On the other
 6015 hand, if the perturbation is an increase, then $c_E(\underline{N}_E + \varepsilon) > \bar{c}$. Now, rural residents migrate to the city until $N = N_E^*$. Thus, regardless of the sign of the perturbation,

once a city of size \underline{N}_E gets a shock to its population, it never returns to its original size. In this sense, a city of size \underline{N}_E is “unstable”. By a similar argument, the larger of the two equilibrium city sizes, N_E^* is stable. Thus, we rule out the smaller equilibrium,
 6020 \underline{N}_E and focus on the unique *stable* equilibrium city size, N_E^* .

This model gives us a foundation for thinking about the size and number of cities, but the existence of multiple equilibria is a problem. The model is missing some mechanism to winnow the set of possible spatial equilibria and make a unique prediction about what will happen. The reliance on stability to resolve this problem is standard, but problematic. Implicitly, notions of stability involve time and households
 6025 that change locations over time. But this means that rational households should be choosing an optimal location at each time. They should be making *many* choices of location, not just one, as in our static model. Justifying the use of stability to reduce the set of possible equilibria in this context, though common, requires logical
 6030 gymnastics.

We have so far discussed the size of a single city in equilibrium. We would like to think about how an equilibrium system of cities is populated. How many cities will there be, and how large?

To start, consider the case when there are initially zero cities. What does this
 6035 model suggest will happen? Notice that in equation (9.11), for any \bar{c} , I can find a city size sufficiently small that city residents are worse off than rural residents. This means that if we start from a world with no cities, then no one will have an incentive to move to cities. The first resident of a city is always worse off than if they stayed in the countryside. We need some form of collective action to get cities.

6040 On the other hand, if the pool of rural people is arbitrarily large, then any number

of cities can exist in equilibrium. As long as all cities are of size N_E^* , we can have any number of them. In this case, no household can improve their lot. All cities and the countryside give exactly the same payoff.

If the pool of rural people is finite, then the problem is even more complicated.

In this case, if all rural people urbanize then cities can be any size, as long as this size is between $[N^{**}, N_E^*]$. Such a configuration has three properties. First, because all the cities are the same size, they all give the same payoff, and so no household wants to switch cities. Second, by inspection of figure 9.8, for city sizes in the interval $[N^{**}, N_E^*]$ the payoff in the city is at least as good as the rural payoff, so no city resident wants to move to the country. Finally, city sizes in the interval $[N^{**}, N_E^*]$ are stable.

We are now in an even worse mess than when we faced the problem of two possible equilibria in a single city. When the set of rural households is finite, the model is predicting that many different systems of cities are possible in equilibrium, all of them full of stable cities. As long as everyone lives in a city and the city size is in the interval $[N^{**}, N_E^*]$, it's an equilibrium system of cities. When the set of rural households is infinite, any number of cities can occur in equilibrium, as long as they are all size N_E^* .

We need some way to choose among all of these equilibrium systems of cities.

Henderson's solution to this problem is to imagine a new type of agent, the "real estate developer", who chooses city sizes to maximize the profit from developing the city. We discuss this after we work out the planner's problem below.

Before we turn to the planner's problem, note a second problem with our equilibrium systems of cities. In equilibrium, all cities have to be the same size and all cities

6065 (mechanically) contain only one sector of employment. This is not consistent with common sense, much less with the stylized facts established earlier in this chapter. Both of these problems are easy to solve.

To see how we can adjust the model to allow for cities of different sizes, consider a system consisting of two cities with output in the two cities given by,

$$Y_1 = A_1 N_1^{1+\sigma}$$

6070 and

$$Y_2 = A_2 N_2^{1+\sigma},$$

where the subscript indicates the city under consideration. Otherwise, each of the two cities is a copy of the case that we have already considered. By varying A_i we shift $c_E(N_i)$ up and down, and as we shift $c_E(N_i)$ up and down, the equilibrium city size increases or decreases. This means that, holding \bar{c} fixed, I can choose A_i to 6075 adjust the equilibrium size of city i . By choosing A_1 and A_2 carefully, I can get two equilibrium cities of more or less arbitrary sizes. Generalizing to many cities, we can (for example) choose the A_i 's so that the resulting system of cities conforms to Zipf's law.

Generalizing to allow for multiple sectors within a city, e.g., fountain pens and 6080 police badges, is straightforward but complicated. It requires three steps. First, we allow each sector to have its own production function in each city. Second, we require a competitive labor market in each sector. Third, we let total commute cost vary with the sum across sectors of all workers in the city. Describing these changes formally

involves a lot of complicated notation, and little new intuition, so we'll skip it. If you
 6085 are interested in seeing the details worked out, look at Henderson [1974] or Au and
 Henderson [2006].

9.5.2 Planner's problem

Now consider the planner's problem. Assume the planner would like to maximize the
 surplus, $W(N)$, created by the city. Surplus is the value of output minus the cost of
 6090 commuting and the opportunity cost of labor. More formally,

$$W(N) = Y(N) - TC(N) - \bar{c}N. \quad (9.13)$$

Substituting from equations (9.7) and (9.9), gives

$$W(N) = AN^{1+\sigma} - t\frac{N^2}{2} - \bar{c}N.$$

To make this quantity as large as possible, the planner chooses city size, N_P^* , to solve

$$\max_N W(N) = AN^{1+\sigma} - \frac{tN^2}{2} - \bar{c}N. \quad (9.14)$$

Equation 9.14 is an unconstrained optimization problem in one variable. We solve

it by evaluating the first order condition, setting it equal to zero, and doing algebra.

6095 That is,

$$0 = \frac{dW(N)}{dN} = (1 + \sigma)AN^\sigma - tN - \bar{c}.$$

Rearranging, we see that the planner's optimal city size must satisfy,

$$\bar{c} = (1 + \sigma)A(N_P^*)^\sigma - tN_P^*. \quad (9.15)$$

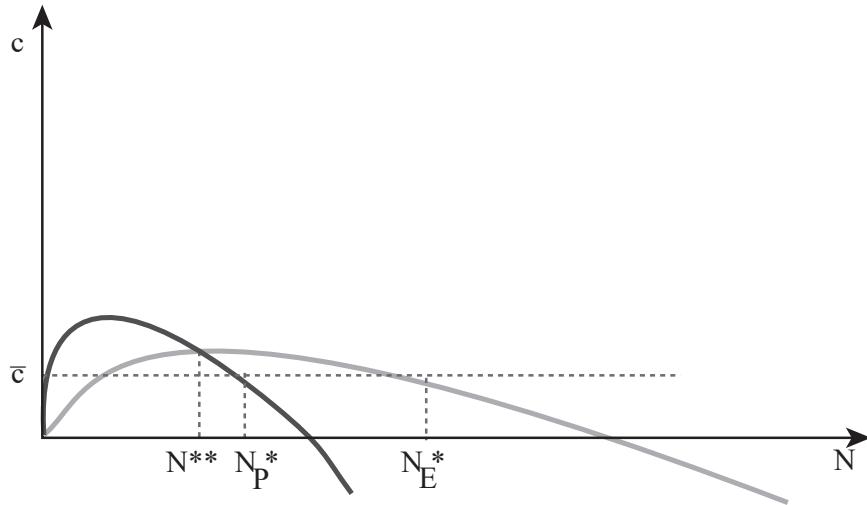
Like equation (9.12), equation (9.15) is difficult to solve analytically, but easy to solve graphically.

Figure 9.9 illustrates this solution. Repeating from figure 9.7 for reference, the light gray curve plots $c_E(N)$. The unique stable equilibrium city, N_E^* , occurs where this curve is equals \bar{c} . The high point on the light gray curve occurs at city size N^{**} . This is the city size that results in the largest possible urban consumption. The dark gray curved line describes the right side of the equation describing the planner's optimal city size, from equation (9.15). This curve lies above the light gray curve for city sizes smaller than N^{**} and below it for larger city sizes. The planner's optimal city size occurs where equation (9.15) is satisfied and the dark gray line crosses the horizontal line with height \bar{c} . This occurs for a city size N_P^* .² By inspection, the planner's optimum must be strictly smaller than the unique stable equilibrium city N_E^* , except in the special case when the equilibrium city size is $N_E^* = N^{**}$. In this case alone, the equilibrium city and the planner's optimum are the same size.

This requires two comments. First, figure 9.9 shows that, in general, equilibrium cities should be “too large” in the sense that they are bigger than what a surplus maximizing planner would choose. This seems surprising. Recall that when the labor market is competitive and there are agglomeration economies, as is the case here, then workers ignore the fact that by moving to a city they make everyone else more productive. Thus, on the basis of what we have done so far, we expect that in spatial

²We're going to ignore the second, much smaller optimal city size because it is unstable.

Figure 9.9: Illustration of optimal and equilibrium city size



Note: The light gray curve describes the curve $c_E(N)$, reproduced from figure 9.7. As in figure 9.7, the stable equilibrium city size, N_E^* . N^{**} is the city size that results in the largest possible difference between urban and rural consumption. The dark gray curved line describes the right hand side of equation (9.15). The planner's optimal city size occurs where equation (9.15) is satisfied. This occurs for city size N_P^* .

equilibrium, workers do not have strong enough incentives to move to cities. This should lead to cities that are “too small”.

This is not what happens. Why? When workers move to the city, the city expands, and the aggregate cost of commuting goes up. Just as agglomeration economies create public good problem, commuting also creates a public bad problem. As the city grows, access to the center becomes congested and it requires longer average commutes for the city to function. People ignore their effect on other people’s commute costs when choosing their residence, and this effect leads to cities that are too big.³ We learn

³Because of this, it is common to describe commute costs as “congestion”, independent of whether roads are congested or not.

6125 from figure 9.8 that the second effect dominates and we end up with cities that are too big.

Second, notice that in figure 9.9, the planner's optimal city size, N_P^* , is larger than the city size that maximizes the level of urban consumption, N^{**} . Why would the planner want to do anything other than maximize the level of consumption for city 6130 dwellers? Looking carefully at equation (9.13), the planner is not trying to maximize urban consumption levels. Rather, the planner is trying to maximize the total value created by the city. In figure 9.9, we see that the level of urban consumption, the light gray line, is nearly constant in a neighborhood of N^{**} . This means that the planner can add people beyond N^{**} with little effect on per capita urban consumption, but 6135 a big effect on the aggregate surplus created by the city. Hence, a planner interested in maximizing the total surplus created by the city allows a city that is a little bit larger than the size that maximizes the level of urban consumption.

9.5.3 Extension #1: Real estate developers

The systems of cities model we've developed above has two problems. First, even 6140 once we use stability to rule out multiple equilibrium sizes for any given city, there will generally be many different systems of cities that are consistent with spatial equilibrium. This means that, after all this trouble, the model really does not tell us what should happen. Second, equilibrium systems of cities all have the property that all cities should be too big, that is, bigger than the size that maximizes surplus. This 6145 means that there is "money on the sidewalk", if someone can just figure out how to pick it up.

To solve these problems, Henderson [1974] proposes that cities be developed and

managed by real estate developers. Real estate developers do three main things. The first is to develop cities. The second is pay residents to move to the city up until 6150 population just exceeds N . The third is to stop people from moving to the city once population reaches the optimal size. That is, N_P^* . In exchange, the real estate developer gets to charge all city residents (or at least those that arrive after the N th person) the difference between urban and rural consumption levels, $c_E(N_P^*) - \bar{c}$, to live in the city.

6155 As described, this sounds implausible. But it sometimes actually happens. Irvine Ca., with population almost 300k, was planned, built and managed by the Irvine company. More commonly, city councils act in much the same way as Henderson's real estate developers. City councils routinely restrict the ability of land owners to build residential buildings through zoning codes and municipal regulations, and they 6160 collect property taxes. These are the two main activities of Henderson's real estate developers.

At the time of this writing, housing prices in much of the developed world are high and housing seems to be scarce in many places. Such housing shortages are often attributed to overly bureaucratic city councils and to selfish, NIMBY, homeowners 6165 who oppose increases in residential density. The resulting prescription to fix the housing shortage is to limit the ability city councils to regulate building. If our model of systems of cities is right, what does it suggest about the wisdom of such policies?

9.5.4 Extension #2: Allowing cities of size zero

6170 Looking at the expression for urban consumption, either in equation (9.12) or in figure 9.8, we see that urban consumption goes to zero as city size goes to zero. This is

really an artifact of the particular functional form we've chosen for production, and seems a little odd. A person living alone in a city should probably be as productive as a person living alone in the countryside.

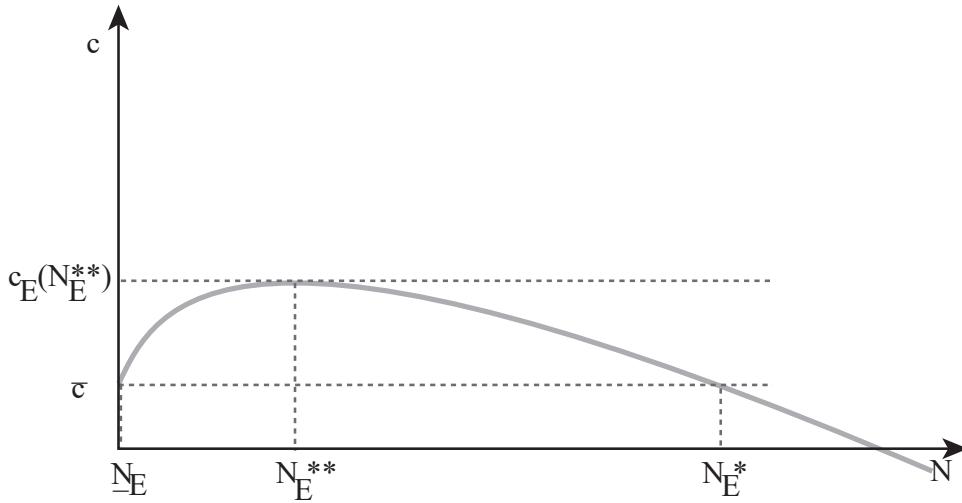
Suppose, that we fix this problem by adjusting $c_E(N)$ so that $c_E(0) = \bar{c}$, as illustrated in figure 9.10. Then, using the same logic as we developed in figure 9.8, we still have two possible equilibrium city sizes, 0 and N_E^* , and only the larger of the two is stable. However, with this change, we rule out equilibria where no one lives in a city. If a single person city gets the same payoff as the countryside, then we do not need collective action for a city to start. A single rural resident can start a city, but this single person city will be unstable and we expect it to grow to size N_E^* .

This suggests the following dynamic story for the evolution of a system of cities. Once a city reaches size N_E^* , a rural residents starts a new city of size zero. Once this happens, the existing city splits and people divide themselves between the two cities. This process repeats when both cities grow to size N_E^* . Therefore, by adjusting the shape of the $c_E(N)$ curve so that cities of size zero have the same payoff as the countryside, we can get city creation in equilibrium system of cities with the need for real estate developers. Obviously, the earlier caveat about dynamics also applies here, too.

9.6 Conclusion

We began this chapter with the following questions. What does the size distribution of cities look like? How does it change over time? What are patterns of sectoral specialization? Do cities specialize in different activities? Does this change over

Figure 9.10: Equilibrium city size when very small cities are possible



Note: *Equilibrium city size when a person in a city of size zero is exactly as productive as someone in the countryside. In this case, there are two possible equilibrium city sizes, $N_E = 0$ and N_E^* . Only the larger equilibrium is stable.*

time?

These questions turn out to be straightforward to answer, at least on the basis of
 6195 US data. Cities are very different from one another. Some, usually small, specialize in the production of one or a few products. Others, usually larger, make almost everything. These larger cities tend to be home to more business services, that is, white collar jobs, and to more knowledge intensive activities generally. In particular, firms often start in big diversified cities before moving to smaller specialized cities.

6200 The size distribution of cities is stable over time, particularly for larger cities, and in many countries, approximately follows Zipf's law. That is, the second largest city is half the size of the largest, and so on. The way that city residents are employed is more ephemeral. Cities with large employment in one sector often have much smaller

employment in this sector a generation later, and conversely.

6205 The locations of cities appears to be determined, at least in part, by the natural advantage of places at the time they were founded. For example, US cities in the South are located along the fall line, while French cities are relatively likely to be on the foundations of Roman garrison towns. Given the persistence of cities, the conditions that precipitated the founding of a city at a particular spot, thus, often 6210 come to seem unimportant. Over a long time frame, the locations of (many) cities seems random.

We have two main tools to explain these patterns, Gabaix's model of random growth and Henderson's model of systems of cities. Gabaix's model of random growth shows that Zipf's law is implied if the growth rate of cities is random and cities cannot 6215 shrink below a certain size. The evidence for both Zipf's law and random growth rates is mixed, but seems defensible for larger cities in many countries.

Henderson's model of systems of cities explains exactly the same thing as Gabaix, but instead of describing the size distribution of cities as a statistical process, he starts from stylized facts about cities and develops an equilibrium model. In particular, 6220 Henderson assumes that productivity is increasing returns to scale in cities, and the costs of operating a city, mainly the costs of getting people back and forth to the center, are also increasing in city size. Together, these two facts suggest that urban consumption first rises and then falls as city size increases. This supposition finds at least some empirical support.

6225 Taking this inverted U shape for the relationship between city size and urban consumption as given, we can begin to think about spatial equilibrium in a system of cities. This leads to three main conclusions. First, we need to be concerned that "no

“cities” is an equilibrium. Since this is contradicted by observation, it should cause us to be suspicious of the model. We can fix this problem by tinkering with relationship between city size and consumption, or by allowing real estate developers. Both fixes seem defensible. Second, we can match any distribution of city sizes and specialization in this model by tinkering with the productivity terms, A_i , and by allowing for multi-sector production technologies. This is good, but does not illuminate why cities are different sizes or are specialized in different sectors. Finally, multiple equilibria are pervasive in this systems of cities model. This is a problem without a solution. The model does not make unique predictions, and so it is hard to use it to explain what we observe.

The Gabaix and Henderson models of systems of cities make some progress on explaining the size distribution of cities. However, they do not speak to why cities specialize as they do or about why cities are located where they are. There has been some research on these questions as well, although the level of technical difficulty places them beyond the reach of this book.

For the interested reader, Duranton and Puga [2001] describes and explains the coexistence of diversified and specialized cities. Rossi-Hansberg and Wright [2007] develops a model of endogenous city growth which explains both Zipf’s law and sectoral specialization, an unambiguous improvement on what we have accomplished here. Rossi-Hansberg and Wright [2007] relies on Henderson’s real estate developers to avoid the possibility of multiple equilibria that we saw emerge in the simpler framework developed above. Fajgelbaum and Gaubert [2020] consider the sorting different types of people across cities and occupations, and consider how equilibrium changes when we allow transfers of income from one city to another. Such transfers are a

routine feature of public finance and implicitly ruled out by the models discussed in this chapter.

None of these models has anything to say about *where* cities locate. “Central Place Theory”, due to Walter Christaller in 1930 posits that cities will arise in a hierarchy. Large business and market centers will be surrounded by smaller regional market cities which are themselves surrounded by agricultural villages. While this is not an equilibrium model, it is intuitively appealing. Hsu et al. [2014] make progress in thinking about how central place theory can emerge as an equilibria. Related to this, De Palma et al. [2019] examine how systems of cities can emerge and where they will locate, and how these patterns vary with the strength of agglomeration economies and the cost of commuting.

Problems

1. Suppose that the shares of three industries in city i are

$$S_{i1} = \frac{1}{10}, S_{i2} = \frac{1}{10}, S_{i3} = \frac{8}{10}$$

- (a) Evaluate the Herfindahl index for this city.
- (b) Suppose the share of each industry in national employment is $\frac{1}{3}$. Evaluate the relative specialization for city i , RZI_i .
2. In this problem we will repeat the derivation of $c_E(N)$ for a circular city. Recall that household consumption is the difference between wages and average commute costs, or

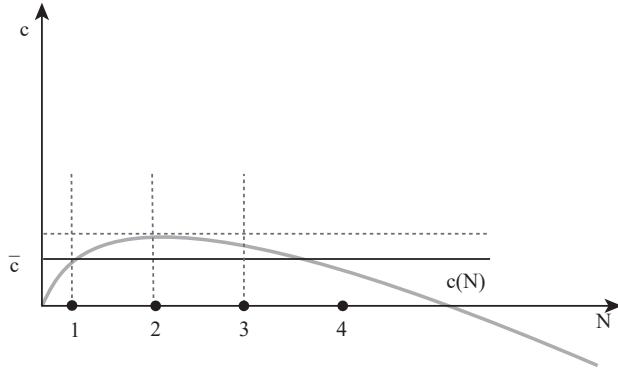
$$c_E(N) = w(N) - \frac{TC(N)}{N} = AN^\sigma - \frac{TC(N)}{N}$$

- (a) First, assuming that $\bar{l} = 1$, and the city extends to \bar{x} , what is the population of the city?
- (b) Recall that for an individual located at x , commuting costs are $2t|x|$. What is the total commuting cost of this city?
- (c) What is the average commuting cost?
- (d) Plug the average commuting cost into the formula for $c_E(N)$. Plot $c_E(N)$.

3. Consider the relationship between city size and consumption as given below.

Note that the utility for being the only person in a city is zero.

Figure 9.11:



- (a) Suppose we must assign 6 people to cities. What possible equilibrium configurations can be maintained? Are there any unstable equilibria?
- (b) What would a real estate developer do? Why?

Chapter 10

Sorting, Voting with Your Feet, and a Simple Hedonic Model

6285 People are different in different places. This is not just the Sesame Street maxim that “everyone is different”. More than 20% of the adult population of Alabama was medically obese in 2000. For Colorado, that share was less than 14%. Similarly, 43% of adults in San Francisco had college degrees in 2000 vs 11% for Danville, Virginia.

6290 This raises two questions. First, are people different because the places cause them to be different, or because people with different attributes and predispositions sort into different places? The answer to this question seems to be “yes”. Sometimes places change people and sometimes people sort. The second question follows from the first. If people are choosing where to live on the basis of their own heterogeneous tastes and attributes, what does this imply for the incentives facing local governments
6295 and firms?

10.1 Sorting versus causation: Why are people different in different places?

Sprawl and obesity: Body Mass Index (BMI) is defined as the ratio of a person's weight in kilograms to their height in meters squared, kg/m^2 . BMI is easy to observe, 6300 is predictive of health problems associated with being overweight, and is widely used as an indicator of whether people are or are not overweight. The threshold for being "obese" is a BMI of 30 or above.

The Behavioral Risk Factor Surveillance System (BRFSS) is a large, representative cross-sectional survey of US adults, that asks survey respondents, among other things, 6305 to report their height, weight, and county of residence. This allows the calculation of BMI. Figure 10.1 reports on these data and illustrates the rate of adult obesity by state. Darker colors indicate lower rates of obesity and lighter colors indicate higher rates. Alabama is worst with an adult obesity rate between 20% and 24%. Colorado is best at between 10% and 14%.

6310 This map invites us to ask whether differences in obesity rates arise because people in different states have different propensities for obesity, or if different states have different propensities to cause obesity. That is, do cross-state differences in the obesity rate reflect the sorting of people, or do they reflect different propensities of states to cause obesity?

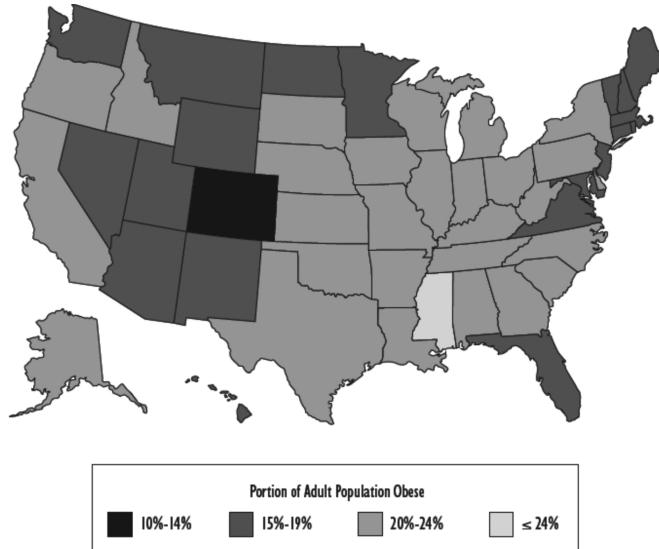
6315 Ewing and McCann [2003] attempt to address this question. Their hypothesis is that urban sprawl is to blame. To check this, they must first measure "urban sprawl". To do this, they use census data to construct a county level index of sprawl as a weighted average of variables describing population density and the organization of

census blocks. Because census blocks often follow streets, this is an easy, if indirect
6320 way of learning about how streets and neighborhoods are organized. They then ask how county level obesity, measured with the BRFSS, varies with their measure of urban sprawl. This shows that people who live in sprawling counties are more likely to be heavy. In fact, the BRFSS measures many other health related outcomes and habits, and Ewing and McCann find many of these are also worse in more sprawling
6325 places.

While these results are intuitively appealing, we imagine people moving to the suburbs and snacking on donuts as they drive from one strip mall to the next, they do not really help us to resolve the problem of sorting versus causation. Do sprawling places make people obese, or do the people who move to sprawling places have a
6330 different propensity for obesity?

Eid et al. [2008] attempts to answer this question. To do this, they construct panel data reporting individual BMI and neighborhood by year. Let i and t index person and year and let BMI_{it} be person i 's BMI in year t . This is the outcome of interest. In addition, let x_{it} be a list of individual characteristics, like age, gender and education. Finally, consider two measures of how sprawling person i 's neighborhood is in year t . First, measure sprawl with the “share of undeveloped land within 1km of residential address” and denote this measure $Sprawl_{it}$. Second, let $Mixed-use_{it}$ be the “count of retail establishments within 1km”. Development in which retail and residential land use are mixed is often regarded as the opposite of sprawl, and so this
6335 variable should have the opposite effect on obesity as sprawl.
6340

Figure 10.1: Adult obesity rates in the US in 2000



Note: *Figure reports the share of the population that is medically obese, that is, with a BMI > 30, by state in 2000. Obesity rates are different in different places. Figure reproduced from Ewing and McCann [2003].*

Eid et al. estimate two main regressions. The first is,

$$\text{BMI}_{it} = \beta x_{it} + \gamma_1 \text{Sprawl}_{it} + \gamma_2 \text{Mixed-use}_{it} + u_{it}. \quad (10.1)$$

The parameters of interest are γ_1 and γ_2 . These parameters describe relationship between sprawl and mixed use and BMI. Because this regression contains individual demographic controls, it estimates γ_1 and γ_2 by comparing the BMI of people with similar demographic characteristics who live in neighborhoods with different levels of sprawl or mixed use.
6345

Here is what they find,

$$\text{BMI}_{it} = \beta x_{it} + 0.46 \text{Sprawl}_{it} - 3.95 \text{Mixed-use}_{it}, \quad (10.2)$$

with the coefficients on both sprawl and mixed use measured precisely. Whether measured by an increase in nearby retail or an increase in nearby undeveloped land,
6350 people who live in more sprawling places are heavier. This approximately replicates the results in Ewing and McCann [2003] using Eid et al.'s quite different data.

Eid et al.'s second regression begins with a more general description of the process determining BMI,

$$\text{BMI}_{it} = c_i + \beta x_{it} + \gamma_1 \text{Sprawl}_{it} + \gamma_2 \text{Mixed-use}_{it} + u_{it}. \quad (10.3)$$

Equation (10.3) differs from (10.1) in the inclusion of the term c_i .

6355 The estimation of equation (10.2) is difficult to interpret. The parameters of interest, γ_1 and γ_2 could reflect a causal effect of neighborhood on obesity, or they could reflect the sorting of people with a propensity to be heavy into sprawling places. Equation (10.3) allows a more precise description of this problem. The term c_i is person level unobserved propensity to be heavy. If we can figure out how to either
6360 estimate c_i , or purge it from our estimating equation, we will have taken a big step towards answering the sorting versus causation question.

Eid et al. begin by taking first differences,

$$\begin{aligned} \text{BMI}_{it} &= c_i + \beta x_{it} + \gamma_1 \text{Sprawl}_{it} + \gamma_2 \text{Mixed-use}_{it} + u_{it} \\ \underline{\text{BMI}_{it-1}} &= \underline{c_i + \beta x_{it-1} + \gamma_1 \text{Sprawl}_{it-1} + \gamma_2 \text{Mixed-use}_{it-1} + u_{it-1}} \\ \implies \Delta \text{BMI}_{it} &= \beta \Delta x_{it} + \gamma_1 \Delta \text{Sprawl}_{it} + \gamma_2 \Delta \text{Mixed-use}_{it} + \Delta u_{it} \end{aligned} \quad (10.4)$$

First differencing means all time-invariant individual characteristics drop out, c_i in particular. This regression compares how much BMI changes for people who move to neighborhoods with more versus less sprawl or mixed-use.

Notice that this is exactly the same approach taken in equation (8.7) to estimate the effect of the number of nearby inventors on inventor productivity, so we have been through this argument before in a different context.

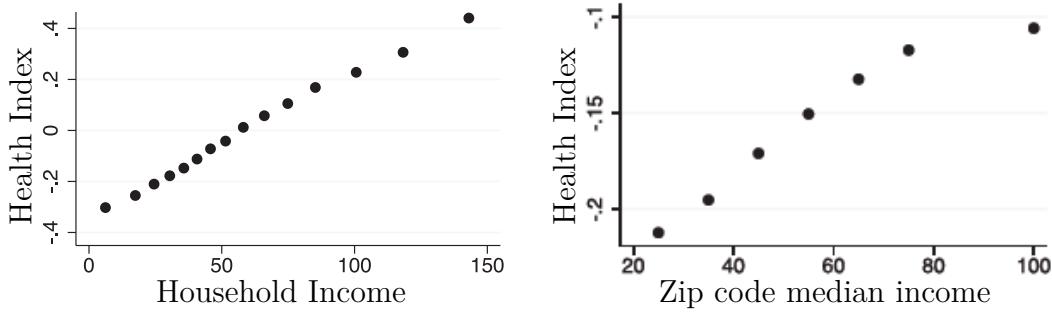
When Eid et al. estimate equation (10.4), they find that

$$\Delta \text{BMI}_i = \beta \Delta x_i - 0.04 \Delta \text{Sprawl}_i + 0.50 \Delta \text{Mixed-use}_i, \quad (10.5)$$

and (as in equation (10.2)) the coefficients on sprawl and mixed use are estimated precisely.

Two things about this estimate stand out. First, the effect of sprawl on obesity is actually slightly negative and the effect of mixed use on obesity is slightly positive. That is, if we control for the unobserved individual propensity to gain weight, c_i , by first differencing, the effect of sprawl and mixed use on obesity is close to zero or slightly negative. Second, the estimated effects in equation (10.2) and (10.5) are quite different. In equation (10.2) we see that people are heavier in more sprawling places. In equation (10.5), we don't. This suggests that effect of sprawl on obesity

Figure 10.2: Average healthiness of household purchases and store offerings by income



Note: *Left: x-axis is household income, y-axis is health index of grocery purchases, dots report meant health index of grocery purchases for households in each income bin.*

Wealthier households buy healthier groceries. Right: x-axis is zipcode median income, y-axis is mean healthfulness of groceries sold by a store, dots give mean healthfulness of stores by zipcode median income. Stores in wealthier neighborhoods sell healthier food.

Figures reproduced from Allcott et al. [2019], ©Oxford University Press.

in equation (10.2) is because people in more sprawling places tend to have different

6380 unobserved propensities to be heavy.

That is, the observed difference in obesity across places reflects the sorting of people with different propensities to gain weight into sprawling places, not a causal effect of sprawl on weight.

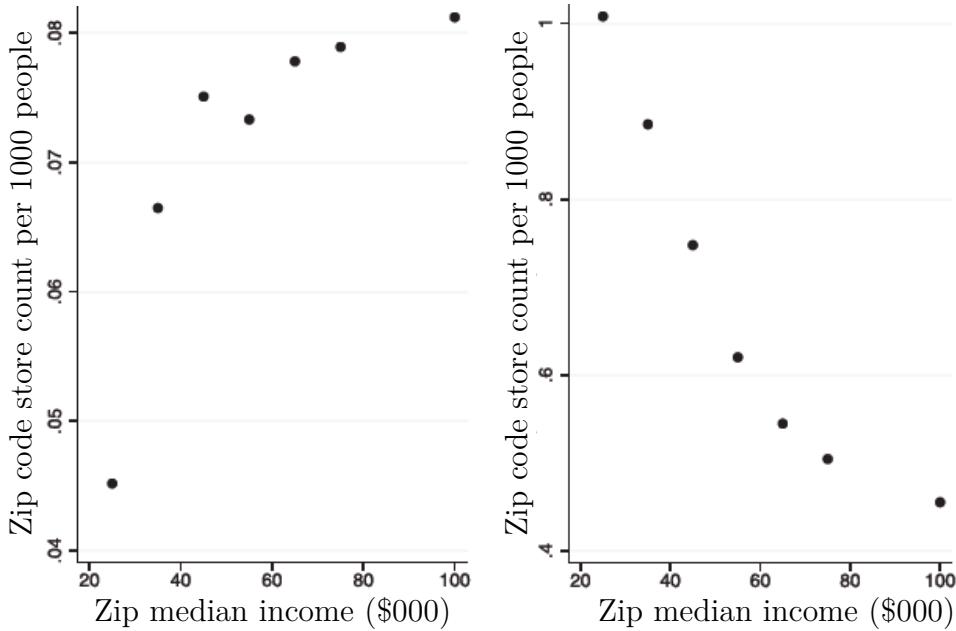
Diet and food deserts: Allcott et al. [2019] look at how the availability of big

6385 grocery stores affects diet. Less formally, if you live in a “food desert” does the entry of a grocery store change your diet?

We just investigated the relationship between a landscape characteristic, sprawl, and an outcome related to behavior, obesity. In this case, the available evidence suggests that landscape does not have a causal outcome on behavior, but that people inclined to the behavior sort into particular landscapes. We here repeat the analysis

6390

Figure 10.3: Food deserts and income



Note: *Counts of large grocery and convenience stores per 1000 residents for an average zipcode in each income bin between 2004-2016. Large groceries are those with 50 or more employees and small groceries have fewer. Wealthy neighborhoods have more big grocery stores. Poor neighborhoods have more convenience stores.* Figure reproduced from Allcott et al. [2019], ©Oxford University Press.

for a different landscape attribute, proximity to a large grocery store, and a different outcome, a healthy diet. We would like to determine whether people in poor neighborhoods have worse diets because they have worse stores, or do poor neighborhoods have worse stores because the people in these neighborhoods demand worse groceries?

6395 Allcott et al. [2019] tackle this question. They begin with data that matches people and their grocery purchases to stores and the set of products available at these stores.

Grocery stores assign each product a Universal Product Code, or UPC. These

codes are extremely detailed. Literally, each distinct product in a store has its own code. Checkout scanners record these codes, grocery order by grocery order, and in some cases, match these order-by-UPC data to individuals. Such data allows Allcott et al. to track the grocery purchases of households across stores and over time. By aggregating purchases within a store, they can also describe the characteristics of an average grocery order sold at the store.

The next step is to define “healthy” and “food desert” quantitatively. The “Healthy Eating Index” is a widely used index for assessing the healthfulness of an individual’s diet. This index consists of a weighted sum of different components of a person’s diet. Loosely, a diet where a high fraction of calories comes from fresh fruits and vegetables and whole grains gets a good score, and diet where a high share of calories comes from sugar, saturated fats, and highly processed foods gets a low score. By applying this index to household grocery orders and to total store sales, Allcott et al. can track the healthfulness of groceries purchased by a household or sold by a store.

The left panel of figure 10.2 shows the relationship between household income and the healthfulness of household grocery purchases. The x -axis of this figure reports household income, and the y -axis reports the health index value of grocery purchases. The dots in the figure report mean healthfulness of household grocery purchases by income. Households with higher incomes buy healthier groceries. The right panel of figure 10.2 shows that people who live in poor neighborhoods confront less healthy offerings in their local grocery stores. The x -axis in this figure reports mean zip code income and the y -axis reports the store level healthy eating index. Each dot gives the mean health index value for stores in zip codes in each income bin. The figure is unambiguous. The stores in poorer zip codes sell less healthy groceries. These are

the basic data that give rise to the conjecture that living in a food desert is harmful.

Figure 10.3 suggests a mechanism by which food deserts operate on diet. The x -axis in both panels is zip code median income. The y -axis in both figures is the count of stores per person. On the left, it is the count of large grocery stores, those that employ more than 50 people, and on the right it is grocery stores that employ fewer than 50 people. These size classifications map approximately into full service grocery stores and convenience stores. The pattern in the two figures is clear. There are lots of convenience stores in poor neighborhoods and not many large grocery stores, and the opposite in wealthier neighborhoods. If people living in poor neighborhoods want to buy healthy groceries nearby, they will struggle to do so.

This suggests that we could improve the diets of people living in poor zip codes if we could just find a way to open large grocery stores nearby. Allcott et al.'s data allows them to examine exactly this experiment. 23% of their sample households live in zipcodes without a large grocery store. These are households living in food deserts. For many of these household, a new large grocery store enters within a 15 minute drive of their location during the study period. If the availability of a large grocery store full of healthy groceries is an important determinant of the healthfulness of household grocery purchases, then we should see a change in the shopping behavior of these households.

Figure 10.4 shows the results of this experiment. In each of the three panels, the x -axis reports the number of quarters since the opening of the new large grocery store, with negative numbers indicating prior periods. For households in food deserts, each panel shows how an aspect of household grocery expenditures changes around the time when a large grocery store opens nearby.

In the top panel we see that the share of household expenditures at the chain to which the new store belongs increase by about 3%. Food desert residents respond to the availability of the large new grocery just as we think they should, they shop there.
6450 So far so good. The middle panel shows how the food desert residents change their total expenditure at large grocery stores around the time the new store opens. This plot shows no change. Food desert residents shop more at the new store, but they shop less at other large grocery stores, so there is no net change in the expenditure share at large grocery stores. The bottom panel shows the change in the health index of grocery purchase for food desert households around the time of the new store opening. Consistent with the fact that expenditures at large stores stays constant, so to does the health index.

Summing up, poor households are much more likely to have poor access to healthy foods than wealthier households and they are also more likely to have poor diets.
6460 However, on average, changing their access to healthy foods does not change the health index of their grocery purchases. Just as we saw with obesity and sprawl, the attributes of the place do not affect peoples' behavior around grocery purchases. Rather, poor people who are predisposed to purchase unhealthy groceries tend to sort into the same neighborhoods. On average, opening grocery stores does little to
6465 change their behavior.

Test scores and school districts: In the cases of obesity and diet, the case for the importance of sorting is made indirectly. The hypothesized effect of landscape, whether sprawl or proximity to groceries, operates in the cross-section, but not once we allow for the possibility that the locations of people or stores can change. Because

6470 people don't seem to change their behavior when their landscape changes, we conclude that the sorting of people into places must be responsible for the differences we see in the cross-section. In this sense, the evidence, thus far, for sorting is indirect.

6475 Bayer et al. [2007] gives more direct evidence. In the US, public school choice is determined by residential location. Children can generally only attend public schools in the same school district as their house. This creates an incentive for parents who value school quality to choose houses in school districts with good schools, exactly the intuition we exploited to value school quality in Chapter 1. Bayer et al. construct data reporting mean fourth grade test scores, household demographics, and household location for a large sample of households in the San Francisco Bay Area in 1990. These 6480 data allow Bayer et al. to examine how demographic characteristics vary at school district boundaries where the quality of available schools changes.

6485 Figure 10.5 reports their results. In each panel, the x -axis reports distance from the school district boundary in miles, with positive distances indicating displacement into the better of the two districts and conversely. In every panel, the mean value of the y -axis variable is reported in bins 0.02 miles wide, relative to the value for the bin that extends from the boundary to -0.02, the first bin outside the better district. The top left panel shows the increase in test scores when we cross from the worse to the better district. That there is a large increase is a mechanical consequence of the way the figures are constructed, the better district of every border-pair is shown on 6490 the right side of the figure.

The remaining three panels have the same structure as the top left, but report on different outcomes. Clockwise from the top right, these outcomes are; mean income, share black, and share of residents with college degrees. In each case, we see obvious

Table 10.1: US MSAs with highest and lowest shares of college graduates in 2000

Highest		Lowest	
MSA	Share college	MSA	Share college
San Francisco, CA	0.44	Mansfield, OH	0.12
Washington, DC	0.42	Vineland-Millville-Bridgerton, NJ	0.12
Colombia, MO	0.42	Visalia-Tulare-Porterville, CA	0.11
Madison, WI	0.41	Danville, VA	0.11

Note: *Reproduced from Moretti [2004].*

changes when we cross from the worse to the better district. On average, the residents
6495 of the better school districts are better educated, have higher incomes, and are more likely to be white. Because fourth grade tests scores are unlikely to affect the race, education, and income, of current adult residents, it seems likely that these differences reflect the sorting of people with different attributes into different locations.

The fact that demographics change at the border calls into question the border
6500 design methodology we developed to value school quality in Chapter 1. Because demographics vary the border it may be that differences in house prices across school district boundaries reflect households' willingness to pay for their neighbors' attributes in addition to (or instead of) their willingness to pay for school quality. This possibility was invisible in the framework we developed in Chapter 1 because the model underlying this framework did not allow people to be heterogeneous. Bayer et al. generalize the framework from Chapter 1 to allow households to differ from each other in many ways. This leads them to conclude that school quality is much less important for explaining housing price changes across municipal borders than in previous estimates.

6510 **Neighborhood and human capital:** Table 10.1 reports the share of college graduates in the four MSAs with the highest share of graduates and the four with the lowest share of graduates. The differences are dramatic. The four highest have college shares above 40% while the four lowest are 12% or less. The two cities with the highest share are San Francisco and Washington DC, both destinations for the highly educated, while the next two are small towns that are home to large colleges. The 6515 four MSAs at the bottom of the standings are small rust belt or agricultural cities. It is not obvious whether the variation in college share reflects the sorting of educated people into certain MSAs, or if certain MSAs are better at educating their residents.

Because the accumulation of human capital is so important a determinant of 6520 income and well being, understanding the extent to which places affect human capital is also important. However, the relationship between place and human capital is also famously difficult to study. The Moving to Opportunity experiment was a large scale randomized control trial conducted in order to resolve it.

The Moving To Opportunity experiment, or MTO, was administered by the US De- 6525 partment of Housing and Urban Development between 1994-8 in five cities; Chicago, NY, LA, Boston and Baltimore. In my opinion, it is one of the high points of social science research over the past two generations.

The experiment was designed to determine whether, and how much, the lives of the very poor could be affected if they moved from their current poor neighborhoods into 6530 more affluent neighborhoods. To be eligible for the experiment, households needed to; have children, live in public housing projects or other subsidized housing, and live in a neighborhood where the poverty rate was above 40% (a very poor neighborhood).

Eligible households that agreed to participate were randomly assigned to one of

two treatment groups or a control. In the first “Experimental” treatment, households
6535 were offered a voucher for rent that they could use only if they moved to a neighbor-
hood with a poverty rate less than 10%. The rent voucher paid the difference between
30% of the household’s income and market rent (up to a ceiling). The Experimental
voucher was an opportunity to move out of public sector housing into a unit of the
household’s choosing, but only if the unit was in a much less impoverished neighbor-
hood. If households did not use the voucher, they were permitted to stay in their
6540 current unit. At this time, residents of public or subsidized housing were expected to
pay 30% of their income in rent, and so moving did not directly affect income net of
rent.

In the second experimental treatment, households were offered an identical voucher
6545 to the one received in the first treatment, but without the location restriction. This
voucher has the same structure as the one on which one of the largest US housing
subsidies is based, the “Section 8” program, and so it is called the “Section 8” treat-
ment. As for the Experimental treatment, if households did not use the voucher, they
were permitted to stay in their current public sector housing.

6550 The remaining households were assigned to the control group. Households in this
group were permitted to stay in their current housing. They did not receive a voucher,
although they did receive counseling intended to help them move if they wanted to.

In total, 4604 households and 15,892 people participated in the experiment. Of the
15,892 people, 11,276 were children. A typical participating household was headed
6555 by a young Hispanic or black single mother without a high school diploma, and was
very poor.

Initial analyses of the experiment conducted two years afterward found little effect

on the test scores of children, though treated children were healthier [Katz et al., 2001]. Several years later [Chetty et al., 2016] reanalyzed the MTO experiment to 6560 investigate long run effects on children in affected households. Chetty et al. linked data for the children in the MTO experiment to tax data for the period 2008-2012, 10-14 years after the experiment. During this 2008-2012 window, the children of treated households ranged in age from 13 to 27, and so it is possible to examine of how the experiment affected adult incomes.

6565 Let y_i denote an outcome for child i . Chetty et al. consider many outcomes, but I restrict attention just to annual earnings between 2008-12. Next, define two indicator variables. The first, EXP_i takes the value one if child i 's household is offered the Experimental treatment and zero otherwise. The second, $S8_i$ takes the value one if child i is offered the Section 8 treatment and zero otherwise.

6570 The main estimating equation is,

$$y_i = \alpha + \beta_E^{ITT} EXP_i + \beta_S^{ITT} S8_i + \varepsilon_i. \quad (10.6)$$

Chetty et al. also worry about city specific and child specific observables, but I'm suppressing this to keep things a little simpler.

The interpretation of this regression is both precise and simple. Consider the children of a control households. For these children, both EXP_i and $S8_i$ are zero. 6575 For this subsample, equation (10.6) consists of just the constant and the error. In this case, α evaluates to the mean of y_i for the children of control households.

Next consider the children of households receiving the experimental treatment. For these children, EXP_i is one and $S8_i$ is zero. For this subsample, equation (10.6)

is just $\alpha + \beta_E^{ITT}$ and the error. In this case, $\alpha + \beta_E^{ITT}$ evaluates to the mean of y_i for
 6580 treated children.

Because we have already evaluated α using the sample of Control children, we can take the difference between Experimental and control group means to get β_E^{ITT} . Thus, β_E^{ITT} is the mean difference between Experimental and control group children, exactly the quantity of interest. A similar logic shows that β_S^{ITT} must be the mean
 6585 difference between Section 8 and Control children.

The second column of table 10.2 reports results of an estimate of equation (10.6) when the outcome variable is the poverty rate in child i 's neighborhood one year after the experiment. The average neighborhood poverty rate for children in the control group is about 50%. For children offered the Experimental treatment it is 17% lower,
 6590 about 33%. This is about the same as for children in the Section 8 group.

These sorts of estimates are often called an “intent to treat” effect, or ITT. It’s the effect on the people you randomly select for an experimental treatment. Intent to treat effects are the average reduction in neighborhood poverty for children whose households are offered a voucher, *regardless of whether they actually move*. Column
 6595 one of table 10.2 reports the share of children whose households took up each of the offered vouchers. Among children in households offered the Experimental treatment, about 47% used the voucher and moved. Among children in households offered the Section 8 treatment, this number is about 66%.

The 17% reduction in neighborhood poverty rate experienced by an average child
 6600 receiving the Experimental treatment reflects the movement of just less than half of treated children. This means that the reduction in poverty for a child whose household was induced to move by the Experiment must be larger, about $17.05/0.47$, or about

35%.¹ This effect is sometimes called the effect of treatment on the treated, or TOT. The third column of table 10.2 reports this value. We see that among households who
6605 accepted the offered voucher, households receiving the Experimental treatment saw a much larger reduction in poverty share than those receiving the Section 8 treatment. This is just what should have happened. The experiment worked the way it was supposed to.

The terminology “treatment on the treated” means something a little subtle. In
6610 the MTO, assignment to Control, the Experiment, or Section 8, was randomized. However, conditional on assignment to a treatment arm, the decision to move is not random. It is made by optimizing households. This raises the possibility that households who move in response to the treatment are different from those who do not. Thus TOT is not generally the same as the effect of moving a child at random.
6615 More troubling for our purpose, is the possibility that the households who select into moving in the Section 8 treatment are somehow different than those who select into moving from the Experimental treatment. This is a limitation of even this *actually* random experiment.

We have established that MTO worked as it should; it induced randomly selected
6620 households to move to better neighborhoods. Thus, if we look at what happens to affected children, we can learn the effects of these better neighborhoods.

Table 10.3 presents results like those in table 10.2, but where the outcome variable is annual income for children between 2008 and 2012. During this period, sample children will range in age from 13 to 27, so not all of them will be in the workforce,

¹This calculation of TOT effects is only approximate; using the numbers in the first row of table 10.2, $17.05/0.47 = 35.77 \neq 35.96$. This difference is not rounding error. The actual calculation used in table 10.2 is the result of a more complicated regression technique whose details are beyond the scope of this book.

Table 10.2: Impacts of MTO on voucher take-up and neighborhood poverty rates for children under 13 when treated.

	Voucher take-up	Poverty rate, one year post	
		ITT	TOT
Experiment vs. Control	47.66	-17.05	-35.96
Section 8 vs. Control	65.80	-17.88	-22.57
Control group mean		50.23	50.23

Note: *Reproduced from Chetty et al. [2016].*

6625 but many will be. The control group mean is 11,270\$. For children whose households were offered the Experimental treatment, this number is 1624\$ larger. Because this increase reflects the changed earnings of only the 47% of the children who moved because of the experimental voucher, the increase in income for children who moved is about double, at 3477\$. This is about a 31% increase. The second row reports
6630 the corresponding estimates for children whose households received the Section 8 treatment. These estimates are smaller than those for the Experimental vouchers.

And thus we have our answer. The children of households randomly induced to move to better neighborhoods have better labor market outcomes. Because we have manipulated treatment experimentally, this cannot be a consequence of sorting. It must be caused by the neighborhood. The effect also seems to be large. Comparing
6635 the TOT estimates from tables 10.2 and 10.3, we see that an average child experiencing a 35% decrease in neighborhood poverty experiences about a 31% increase in income. Dividing, this suggests that each 1% reduction in neighborhood poverty rate contributes about 1% to average adult annual earnings.

6640 It is interesting to note that the conclusions of the MTO study seem to line up qualitatively with the de la Roca and Puga [2017] finding from Chapter 8. That is,

Table 10.3: Impacts of MTO on adult earnings for children under 13 when treated.

	Annual earnings, 2008-12	
	ITT	TOT
Experiment vs. Control	1624	3477
Section 8 vs. Control	1109	1723
Control group mean	11270	11270

Note: *Reproduced from Chetty et al. [2016].*

that spending time in a big Spanish city increases a person's wage. It looks like where you live also affects human capital accumulation, at least in Spain and the US.

Sorting versus Causation, summing up: People are different in different places.

6645 Sometimes, these differences appear to reflect sorting. Among the examples we've discussed, this looks to be the case for the relationship between obesity and sprawl, and diet and food deserts. It is even clearer that people sort to be near particular neighborhood amenities like good schools and people like themselves. But sometimes, places actually change people. This seems to be the case for the relationship between 6650 neighborhood poverty rates and adult incomes. City size probably also causes changes in peoples' productivity.

So, people are different in different places, and sometimes it's because people sort and sometimes it's because places change people. If you want to know which is happening in any particular case, you need to implement a quasi-experimental 6655 (or experimental) research design to check. We don't have any theory to help us understand which characteristics are subject to change by place of residence, and which are not. This is a question that researchers have not even begun to address.

10.2 Spatial equilibrium and Tiebout sorting

People are different in lots of important and interesting ways and they (sometimes) 6660 sort across locations on the basis of these differences. Some of the place specific attributes that people sort on are provided or influenced by local governments, e.g., sprawl, public schools, grocery stores. Each of these place specific goods is a public good.

This invites two questions. First, is this complicated process consistent with our 6665 basic notion of spatial equilibrium? (Yes, but not too surprisingly, it's a little messy). Second, how does the possibility of sorting affect the incentives for local governments have to provide "local public goods".

Many of the services that local governments provide look like public goods; police, fire, schools, roads, transit, water, electricity, gas, trash collection. These goods are 6670 not "excludable". That is, it's hard or impossible to deny them to anyone in the service area. This makes it difficult to charge people prices that reflect the marginal cost of the services. These goods are typically financed with property taxes, but there are lots of other possibilities; sales taxes, excise taxes on cars and other property, or taxes on gasoline.

6675 This creates a problem. The collection of tax revenue and the provision of local public goods are not as connected as for private goods. If a restaurant gives me a bad meal, I can choose not to go back. If the city doesn't pick up my trash, I can't withhold my property tax payment.

In a classic paper Tiebout [1956] makes the argument that people can "vote with 6680 their feet" and move away from municipalities that do not provide good value for money, that is, good public services per dollar of tax collected. Once we allow this

sort of mobility, the provision of local public goods looks more like a private good. If a restaurant is bad, I don't go back. If a municipality is bad, I move away. This means that any municipality that does not offer a bundle of public services that consumers 6685 want at a price (in taxes) that is competitive with other nearby municipalities, is going to see all of its residents move away.

This has three interesting implications. First, we should expect to see optimal provision of local public goods. If you compare this to the pure theory of public goods, this is a remarkable and neat conclusion. Second, the effect of a marginal change in the 6690 property tax rate on real estate prices should be zero. Why? If services are provided optimally, then their marginal cost exactly equals their marginal value to households, and so the change in public services and the change in property tax exactly offset each other. Third, we should expect to see municipalities specialize in serving populations with different tastes for public goods. There should be communities that have high 6695 taxes and lots of public services, and communities that have low taxes and fewer services.

Tiebout wrote his paper in 1956, before the profession began to depend on mathematical models. To understand how the Tiebout's model works a little better, it's helpful to think about it in the context of a spatial equilibrium model with heterogeneous agents. 6700

The following discussion is based on Chapter 8 of Sieg [2020], which offers a nice example of how to do this. To begin, let's think about a population of agents with preferences over consumption, housing, and a local public good, and a set of municipalities that offer different combinations of taxes and local public goods.

6705 To begin, suppose there are just two municipalities, indexed by $i = 0, 1$. Let G_i

be the level of the local public good in municipality i , w be income, and p_i the price of housing in municipality i . People have different incomes, w , but are otherwise the same. For completeness, let c and h be consumption and housing.

Describe preferences with an indirect utility function,

$$V(G, p, w, \alpha) = - \left[\frac{\alpha}{G^\rho} + \left(\frac{p^\beta}{w} \right)^\rho \right]. \quad (10.7)$$

6710 $\beta \in (0, 1)$ measures the strength of the taste for housing. $\rho > 0$ describes the willingness to substitute between the public good and the other goods, and $\alpha > 0$ measures the strength of the preference for public goods.

The indirect utility function in equation (10.7) is complicated looking but has intuitive properties. First, with α and ρ positive, $\frac{\alpha}{G^\rho}$ is decreasing in the level of public good G , so with the negative sign in equation (10.7), this means that V is increasing in G , just as it should be. A similar argument shows that V is decreasing in the price of housing, p_i , and increasing in income, w , also just as it should be. So, V decreases in the price of housing, and increases in w and G . Figure 10.6 illustrates two indifference curves in (p, G) space. Utility is increasing as rent decreases and the 6715 public good increases, and so the indifference curve V_2 has a higher level of utility than V_1 .

Tiebout sorting models assume that indirect utility functions satisfy a “single crossing” property. This property requires that indifference curves in (p, G) space get steeper as w increases. This is illustrated in figure 10.7, here the household with 6720 the dashed indifference curve has higher income than the household with the solid indifference curve.

To get some intuition about the single crossing condition, fix a household's consumption of housing. In this case, an increase in p of Δp translates to a decrease in c of Δph . Thus, the slope of this indifference curve tells us the willingness to trade c for G . Under the single crossing condition illustrated in figure 10.7, as w goes up, indifference curves get steeper, so people will trade more c per unit of G . That is, richer people will pay more for housing. So in this case, the single crossing condition just says that, all else equal, richer people will pay more for housing.

Suppose we have a set of people whose wages are uniformly distributed between $[\underline{w}, \bar{w}]$ (if you don't know what this means, it is close to "there are the same number of people with each possible wage" but adjusted to the fact that $[\underline{w}, \bar{w}]$ is a continuum). We would like to divide these people between two municipalities offering (p_1, G_1) and (p_0, G_0) .

We started talking about taxes and public goods. But this model is about house prices and public goods. Implicitly, the house prices in this model are after tax. That is, $p_i = (1 + \tau_i)\tilde{p}_i$, where τ and \tilde{p} are the property tax rate and the before tax price of housing. If municipality 0 is a low service, low tax municipality, and municipality 1 is the a service high tax municipality, then we should have $p_1 > p_0$ and $G_1 > G_0$. That is, the high tax municipality 1 has higher (after tax) housing prices and more public goods than does municipality 0. Note that the model doesn't describe where p comes from, as the monocentric city model does. We consider this in the next section.

Suppose we can find w^* such that this household is exactly indifferent between the two municipalities. For this household, both available bundles of (p, G) lie on the same indifference curve, as in figure 10.2. Notice that such a marginal household may not exist. For example, if one municipality is badly managed and provides a low level

of public services, we could have $p_1 > p_0$ and $G_1 < G_0$. In this case, everyone should prefer municipality zero and no marginal household would exist.

Under the single crossing condition, all households with $w > w^*$ choose municipality 0 and conversely. This is illustrated in figure 10.8. The “preferred set” is down and to the right. We’ve just constructed a spatial equilibrium in which there are a continuum of different types of agents.⁶⁷⁵⁵

How does this model reflect Tiebout equilibrium? As a municipality offers “worse” combinations of (p, G) fewer agents will want to choose it.

If we take the Tiebout model seriously, then it has an interesting implication for the relationship between property taxes and real estate prices. If Tiebout sorting truly forces municipalities to compete with other to provide public services, then municipalities should provide public services just up to the point that the tax cost of those services is offset by the value of the municipal services. But the value of municipal services and property taxes are both capitalized into real estate prices.⁶⁷⁶⁰ This means that if public services are being provided optimally, and property taxes are being assessed optimally, then real estate prices should stay about constant when one or the other changes marginally. This is not to say that property taxes are not being capitalized into real estate prices. Rather, they are *both* being capitalized into real estate prices, but their effects are opposite, and if municipal services are being provided optimally, perfectly offset each other.⁶⁷⁶⁵⁶⁷⁷⁰

The evidence about this is mixed. For example, Bayer et al. [2007] and Black [1999] show that property markets capitalize school quality. In the context of the Tiebout model, this suggests that school quality is not optimally provided. Dachis et al. [2012] suggest that changes in the property tax rate in municipal Toronto

6775 were almost completely capitalized into real estate prices. This is consistent with city residents expecting that the change in property tax will have no impact on the level of municipal services. Rosen [1982] finds that house prices almost completely capitalize the property tax decrease that came with California's Proposition 13 tax cuts. This is also consistent with city residents expecting that the change in property
6780 tax will have no impact on the level of municipal services. Coury et al. [2021] find that constructing municipal sewer and piped water supplies in late 19th century Chicago increased land prices by about 60 times the cost of construction. Sewer construction was largely financed by property taxes, and it is hard to imagine that the increment to land prices from sewers was not dramatically larger than the increment to property taxes.
6785 This suggests that, in this case, municipal sewer service was under provided.

This is messy, but seems sensible. We have solid evidence that property prices reflect local service quality. This prediction is common to all of the models we've studied, and is backed up by many careful empirical studies, e,g, those surveyed in Chapter 2. However, the evidence above suggests that, while Tiebout sorting and
6790 "voting with your feet" sometimes disciplines local governments and drives them to provide efficient level of public services in a cost minimizing way, sometimes this does not work. We do not currently have a good understanding of when voting with you feet is effective, and when it is not.

10.3 The hedonic model

6795 The discussion of Tiebout sorting describes market equilibrium in an environment where there are a continuum of types of households and, potentially, a continuum of

levels of public good. This is more general than anything we've talked about so far. The continuous monocentric city model has one (or a few types) of households and a continuum of locations. The discrete monocentric city model has a continuum of
6800 types of households and discrete number of locations.

Thinking about markets with a continuum of types of households and a continuum of goods has obvious relevance to urban economics, e.g., households differentiated by income that choose from a large number of school districts or a continuum of house sizes. We can't really address this problem with the models we've studied so far.

6805 The model that permits us to consider this case is known as the "Hedonic model" [Rosen, 1974]. In addition to its usefulness for urban and spatial models, it is useful for thinking about any good characterized by a continuum of attributes, e.g., the speeds of cars or computers. Specifying the Hedonic model basically involves adding a production side to the model we used for Tiebout sorting.

6810 Chapter 23 of Sieg [2020] offers a nice example illustrating this model. Consider a housing market where houses are differentiated only by quality $z > 0$. Let $p(z)$ be the market price for a house of quality z . This is the equilibrium price schedule of interest. Notice that if z is, for example, distance to the CBD, then $p(z)$ is just the house price gradient that we've already considered.

6815 To simplify the problem, $p(z)$ is the price of one unit of quality z . For example, the price of one house of size or quality z . It is called the "hedonic price function". The solid, light gray line in figure 10.9 describes a simple hedonic price function.

The hedonic model extends to the case where households choose quality and quantity. It also extends to the case where z is a list of attributes, e.g., size, quality and
6820 distance to the center. However, once we move away from the simplest cases, these

models generally only have numerical solutions, and so they are not useful for illustrating ideas.

Let's begin by describing the supply of housing. There are a continuum of firms, really housing developers, who can produce exactly one unit of housing with any quality. Index firms by θ and suppose $\underline{\theta} < \theta < \bar{\theta}$. A firm of type θ produces housing with quality z at cost,

$$c(z, \theta) = \frac{z^\beta}{\theta}.$$

Here, β tells us how the marginal cost of quality evolves. If $\beta > 1$, then each successive increment to z is more costly than the one before, and conversely if $\beta < 1$. It is better to be a firm with a high θ than a low one. A higher θ reduces the cost of all types of housing, so θ indexes both firm type and productivity.

Firms take the hedonic price function as given and chose the quality of their single unit of output to maximize profits. That is, each firm solves,

$$\max_z \pi(z, p(z); \theta) = p(z) - \frac{z^\beta}{\theta}.$$

This is a maximization problem in a single variable, z . To solve it, we differentiate and set the derivative equal to zero. Rearranging this first order condition slightly gives,

$$\frac{dp}{dz} = \beta \frac{z^{\beta-1}}{\theta}.$$

Because we don't know the functional form of $p(z)$, we can't evaluate its derivative

and so we write $\frac{dp}{dz}$.

If we solve this expression for θ , we get,

$$\theta(z) = \beta \frac{z^{\beta-1}}{\left(\frac{dp}{dz}\right)}. \quad (10.8)$$

$\theta(z)$ tells us the types of profit maximizing firms that will optimally make houses
 6840 with quality z when facing hedonic the price function $p(z)$. Because each firm makes just one unit of housing, this is a supply schedule. It describes the number of units provided for each quality z

If we set $\pi(z, p(z); \theta)$ equal to a constant, we can construct iso-profit curves. The dashed black lines in figure 10.9 trace out a series of such iso-profit curves for
 6845 different values of θ , with higher values of θ to the right of lower values. The first order condition for the profit maximization problem requires that the iso-profit curves be tangent to the hedonic price function for every θ , as drawn in the figure.

Now consider the demand for housing. Households have income m and pay the hedonic price for their unit of housing. Households prefer houses of higher quality,
 6850 bigger z , but are heterogeneous in their taste for quality. Let γ index the set of types of households, and suppose that $0 < \gamma < 1$. More specifically, suppose households choose the quality of their house to solve,

$$u(z, p(z); \gamma) = \gamma z^\alpha + m - p(z).$$

The first term in this utility function tells us how households value z . Households with higher types γ value quality more highly. If $\alpha > 1$ then households value each
 6855 successive unit of quality more than the one before. On the other hand, if $\alpha < 1$ then

the marginal utility of quality is decreasing.

The household's problem is a single variable optimization problem, and we solve it by taking the derivative and setting it equal to zero. Rearranging this condition slightly, we have

$$\frac{dp}{dz} = \alpha\gamma z^{\alpha-1}.$$

- 6860 As for the firm problem, because we do not know the functional form for p , we write its derivative in general form. Rearranging, we have

$$\gamma(z) = \frac{\left(\frac{dp}{dz}\right)}{\alpha z^{\alpha-1}}. \quad (10.9)$$

This is a demand function. $\gamma(z)$ gives us the number of households of type γ who optimize with quality z when facing hedonic price function $p(z)$.

- 6865 If we set the household's utility function $u(z, p(z); \gamma)$ equal to a constant, we can draw indifference curves. These are illustrated in figure 10.9 as the solid black lines. The household first order condition requires that indifference curves be tangent to the hedonic price function, the light gray line in the figure. Because iso-profit curves are also tangent to the hedonic price function, this means that iso-profit and indifference curves are tangent to each other along the hedonic price function, or alternatively, 6870 that the hedonic price function is defined as the locus of points where iso-profit and indifference curves are tangent.

Now we need to choose $p(z)$ so that the market clears at every z . One natural

way for this to happen is if

$$\gamma(z) = \theta(z). \quad (10.10)$$

Recalling that $\theta(z)$ describes the number of firms supplying houses of type z , and
6875 that $\gamma(z)$ describes the number of households demanding houses of quality z , this is
 really just an ordinary market clearing equation, demand equals supply. The wrinkle,
 is that it is a market clearing condition for a market in which there are a continuum
 of varieties.

The market clearing condition in equation (10.10) implicitly relies on the “perfectly
6880 assortative matching” of household and firm types. That is, higher firm types, θ ,
 always sell to higher household types, γ . For this condition to hold, we are making
 an implicit assumption that there are the same number of firms and households of
 each type. Notice that lots of other matching rules are possible and the choice of
 matching rule is important for the equilibrium. The requirement that we choose a
6885 matching rule is the main extra assumption that we require in order to extend our
 intuition about equilibrium from conventional markets with one variety to the case
 at hand, with a continuum of varieties.

Substituting the expressions for demand and supply from equations (10.9) and
 (10.8) into the market clearing equation (10.10), we get,

$$\frac{\left(\frac{dp}{dz}\right)}{\alpha z^{\alpha-1}} = \beta \frac{z^{\beta-1}}{\left(\frac{dp}{dz}\right)}.$$

6890 If we rearrange this expression to get dp/dz by itself, we have

$$\frac{dp}{dz} = (\alpha\beta)^{\frac{1}{2}} z^{\frac{\alpha+\beta-2}{2}}.$$

We can solve this for $p(z)$ by integrating,

$$\int \frac{dp}{dz} dz = \int (\alpha\beta)^{\frac{1}{2}} z^{\frac{\alpha+\beta}{2}-1} dz.$$

Because z occurs on the right hand side as part of a polynomial, this integral just requires us to integrate a polynomial. Evaluating this integral, we get

$$p(z) = (\alpha\beta)^{\frac{1}{2}} z^{\frac{\alpha+\beta}{2}} + C$$

This gives us the hedonic price function such that if households and firms optimize,

6895 the market for z clears everywhere. This is just the light gray line in figure 10.9

Almost. We still need to pin down the constant of integration, C . To do this, we need one more piece of information. For example that $u(z, p(z); \gamma = 0) = 0$. This is like a free mobility condition. It says that we need to guarantee the worst household a payoff of zero (or they will leave).

6900 We've now managed to completely describe an equilibrium with a continuum of household and firm types. This is about as general a description of people and firms as you could ask for, so this is a real accomplishment. However, it also raises a problem. Figure 10.10 illustrates. This figure describes almost the same environment as we considered in figure 10.9, but with two changes. First, for clarity, firm iso-profit curves are dropped. Second, I show two indifference curves for a household that is

not very sensitive to changes in p and z . These are the solid black lines, and two indifference curves for a household that is more sensitive to changes in p and z . These are the dashed black lines.

Clearly, we can construct exactly the same hedonic price function from consumers of either type, and therefore, without more information than just the hedonic price function, we can't tell whether our price function describes the behavior of the more or less sensitive agents. In turn, this means that the welfare implications of changes to the equilibrium are also indeterminate. This is a well known problem with the hedonic model. They provide a description of how markets work when people and firms are different from each other in a flexible and realistic way, but this realism highlights problems evaluating welfare that do not appear in simpler models. This problem is the subject of an interesting but difficult literature, e.g., Bajari and Benkard [2005].

10.4 Conclusion

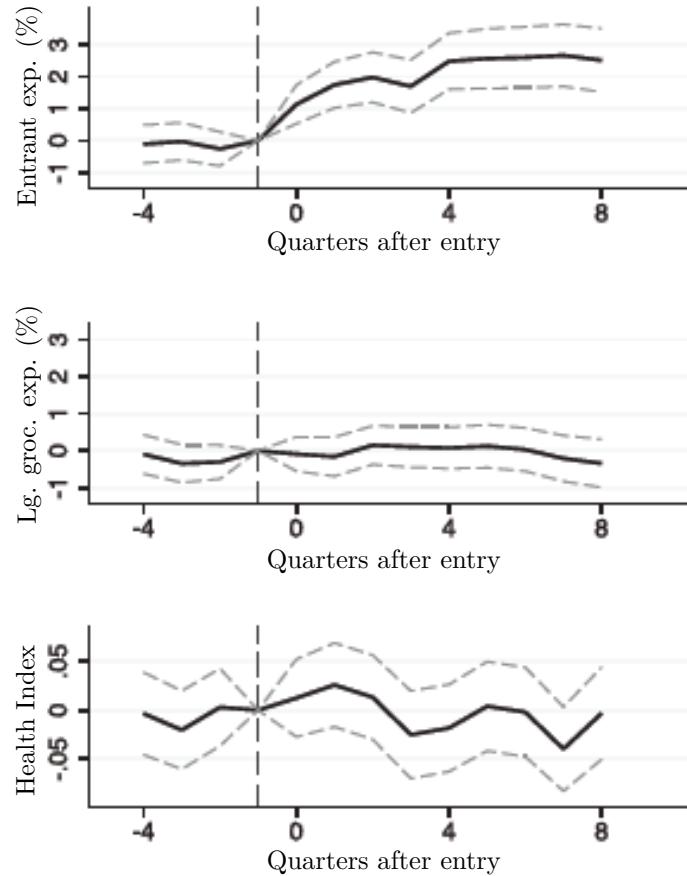
People are different in different places. Figuring out why is tricky for two reasons. First, because it's hard to tell whether differences result from sorting or from the action of places on people. Second, because it looks like sometimes differences are the result of sorting, and sometimes they are caused by differences in places. In the examples we discussed, it looks like obesity and bad diets are not caused by the characteristics of places, but reflect the sorting of people with propensities to be overweight or eat unhealthy diets. Similarly, it looks like affluent, educated and disproportionately white people sort into good school districts; certainly, the school district a person lives in does not affect the color of their skin. On the other hand,

the MTO experiment shows us that sometimes, at least, neighborhoods can change people. Growing up in a poor neighborhood causes adults to be less productive than growing up in a more affluent neighborhood. Similarly, time working in a big city looks like it makes adults more productive than time working in a smaller city.

These facts invite three questions. First, is there some logic governing the sorts of characteristics that are and are not affected by neighborhood characteristics? Second, when neighborhoods change people, what is the mechanism? For example, what is it about more affluent neighborhoods and bigger cities that makes people more productive workers? Finally, what are the implications of the complicated difference between people and the fact that they sort across locations on the basis of these differences, for how we think about spatial equilibrium, and consequently for how we evaluate policy interventions in situations where sorting seems important?

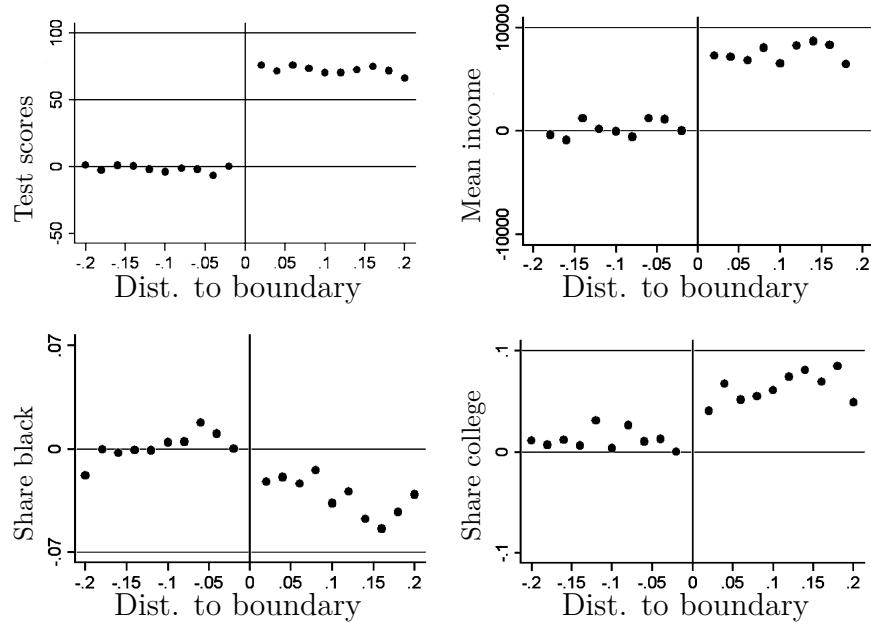
We know very little about the first two questions. If you observe that people in one place have one trait and those in another have a different trait, then, common sense aside, unless the particular situation has been the subject of an investigation, the literature has little guidance to offer as to whether cross-location difference results from sorting or from places changing people. We also know little about why particular places make people more or less productive. We do better on the third question. Both the Tiebout and hedonic model let us think carefully about the implications of sorting.

Figure 10.4: Changes after opening of new neighborhood large grocery in a food desert



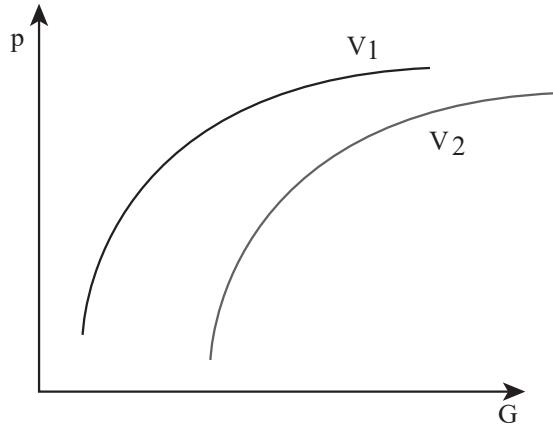
Note: *Top: Share of expenditure at grocery chains with large grocery stores that have opened within 15 minutes of household's zip code. New stores attract about 3% household grocery expenditure, on average. Middle: Total share of household grocery expenditure in large grocery stores for households with 15 minutes of newly opened large store. That this does share does not change suggests that new large stores divert expenditure from other large groceries rather than increasing total expenditures in large groceries. Bottom: Healthfulness of household food purchases for households living within 15 minutes of a new large grocery store. The health index of household grocery purchases does not change when a large grocery store opens near by. In all, these figures suggest that the healthfulness of grocery purchases by food desert residents does not respond to the availability of a large nearby grocery where healthier products are available. Figures reproduced from Allcott et al. [2019], ©Oxford University Press.*

Figure 10.5: Test scores and demographics at school district boundaries



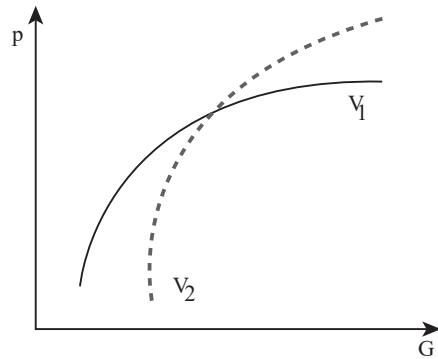
Note: *x-axis shows distance to the school district boundary in miles. Each panel shows how a particular trait varies with distance to the border. Positive values indicate displacement from the border into the better school district and negative values displacements into the worse school district. All figures report mean of y-axis variable relative to the bin extending from the boundary to -0.02, that is the bin just outside the better district. Top left shows mean test score increase across borders. This increase is by construction. The figure is constructed so that better districts are reported on the right hand side of the figure. Top right shows mean income. Bottom left shows black share of residents. Bottom right shows share of residents with a college degree. Because the demographics of adults are not directly affected by school district quality, these differences likely reflect the sorting of people into school districts. Figure reproduced from Bayer et al. [2007], ©University of Chicago Press.*

Figure 10.6: Indifference curves for the indirect utility function of equation (10.7).

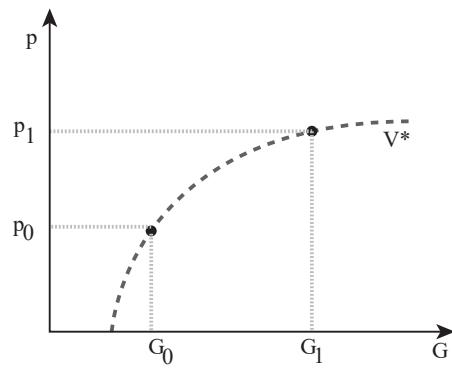


Note: Two indifference curves in (p, G) space. Households are better off as the price of housing, p falls, and the level of the local public good, G , rises, so the utility level for curve V_2 is greater than for V_1 .

Figure 10.7: Single crossing property

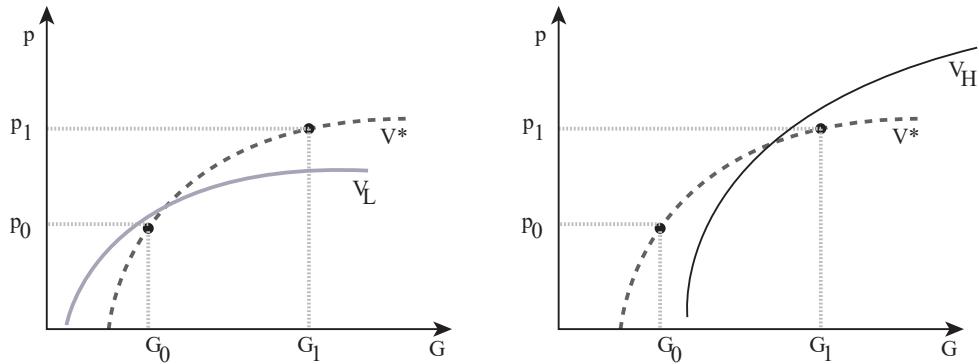


Note: Two positively sloped concave indifference curves, one for lower income V_1 and one for higher income V_2 . The V_2 curve is more-or-less a small counter-clockwise rotation of the V_1 curve, and the two curves cross in exactly one place. That the V_2 curve is everywhere steeper than the V_1 curve indicates a greater willingness to trade income for housing. Indifference maps like the one in this figure satisfy the single crossing property.



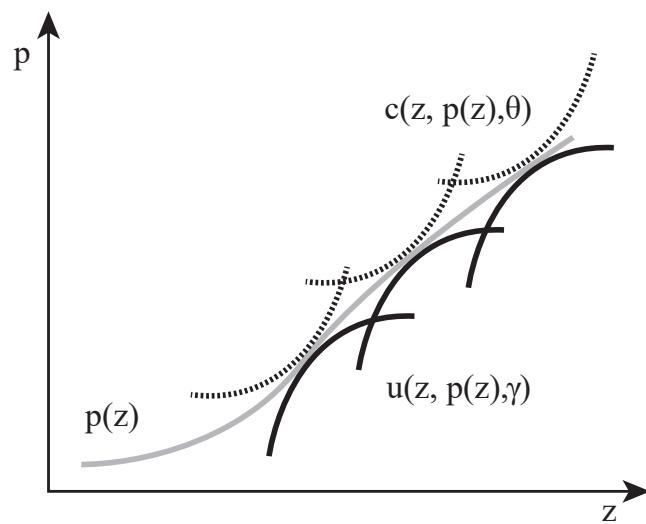
Note: If there are two municipalities, each offering a distinct bundle of prices and public goods, then we can generally find a household that will be indifferent between the two municipalities.

Figure 10.8: Tiebout sorting equilibrium



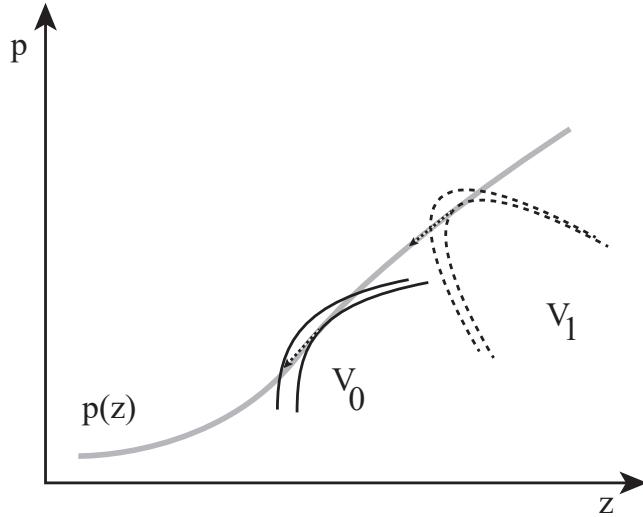
Note: Under the single crossing condition, all households with $w < \hat{w}$ choose municipality 0 and all households with $w > \hat{w}$ choose municipality 1.

Figure 10.9: Illustration of the hedonic equilibrium



Note: The light gray line illustrates the hedonic price function, $p(z)$. This function is easily observed. The solid black lines illustrate indifference curves for different consumers who each optimize at a different value of (p, z) . The dashed lines describe the behavior of different producers who also each optimize at a different value of (p, z) . In an equilibrium, the number of buyers and sellers must match at each pair (p, z) on the hedonic price function.

Figure 10.10: Hedonic price function with alternative preferences



Note: The light gray line illustrates the hedonic price function, $p(z)$. The solid black lines illustrate indifference curves for a consumer who's utility, V_0 , is not sensitive to the exact combination of p and z . For this consumer a small displacement along the price function, for example, does not change utility very much. The dashed black lines illustrate indifference curves for a consumer who's utility, V_1 is more sensitive to the exact combination of p and z . For this consumer utility is more sensitive to a small displacement along the price function. Importantly, both types of consumers can define the same hedonic price function. This means that, without more information about the economy than just the hedonic price function, it is not generally possible to assess how much utility changes when we change the equilibrium somehow.

Problems

1. The standard deviation of the sprawl index in Eid et al. (2008) is 0.281. Use this to evaluate the relationship between a one standard deviation change in sprawl and BMI (using the results from both the cross-sectional regression and the first-differences regression).
6950
2. Use Allcott et al. (2019) to determine, on average, by how much does a \$20,000 increase in household income change the grams of sugar added per 1,000 calories of food on your grocery store shelf? Explain.
3. (a) Using the right two panels of Figure 10.5, estimate the willingness to pay per point for a neighborhood with better test scores.
6955
(b) Using the same logic, estimate the value of a one percentage point higher college share, black share, and \$1,000 in neighborhood income.
(c) Using these four estimates, what is the implied price increase when you cross the boundary into a higher quality school district?
6960
(d) Compare this to the actual price increase. Your estimate should have been much bigger than what is observed. Why did this happen?
4. In this problem, we will work through an example of a hedonic model. Consider a housing market where houses are differentiated only by quality z . Let $p(z)$ be the market price for a house of quality z .
6965
(a) Let the firms cost function be $c(z, \theta) = \frac{z^{1/\beta}}{\theta}$. Set up the firms profit maximization problem. Derive the first order condition.

- 6970
- (b) Each household consumes unit housing to attain a utility $u(z, p(z); \gamma) = \gamma z^\alpha + m - p(z)$. Set up the households profit maximization problem. Derive the first order condition.
- (c) Let $p(z) = z^2$, $\alpha = 2$ and $\beta = \frac{1}{3}$. Assuming perfectly assortative matching, solve for the optimal housing quality that clears the market. What is the number of firms making houses with the optimal housing quality?

Bibliography

- 6975 Alberto F. Ades and Edward L. Glaeser. Trade and circuses: Explaining urban giants.
Quarterly Journal of Economics, 110(1):195–227, 1995.
- Gabriel M. Ahlfeldt and Jason Barr. The economics of skyscrapers: A synthesis.
Journal of Urban Economics, 129:103419, 2022.
- David Albouy. The unequal geographic burden of federal taxation. *Journal of Political Economy*, 117(4):635–667, 2009.
- 6980 David Albouy, Walter Graf, Ryan Kellogg, and Hendrik Wolff. Climate amenities, climate change, and American quality of life. *Journal of the Association of Environmental and Resource Economists*, 3(1):205–246, 2016.
- Hunt Allcott, Rebecca Diamond, Jean-Pierre Dubé, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell. Food deserts and the causes of nutritional inequality. *Quarterly Journal of Economics*, 134(4):1793–1844, 2019.
- Marcella Alsan and Claudia Goldin. Watersheds in child mortality: The role of effective water and sewerage infrastructure, 1880–1920. *Journal of Political Economy*, 127(2):586–638, 2019.

- 6990 Mohammad Arzaghi and J. Vernon Henderson. Networking off Madison Avenue. *Review of Economic Studies*, 75(4):1011–1038, 2008.
- Chun-Chung Au and J Vernon Henderson. Are Chinese cities too small? *Review of Economic Studies*, 73(3):549–576, 2006.
- David Autor. The faltering escalator of urban opportunity. In *Securing Our Economic Future*, Melissa S. Kearney and Amy Ganz (eds.), pages 108–136. Aspen Strategy Group, 2020.
- Patrick Bajari and C Lanier Benkard. Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of Political Economy*, 113(6):1239–1276, 2005.
- 7000 Nathaniel Baum-Snow. Did highways cause suburbanization? *The Quarterly Journal of Economics*, 122(2):775–805, 2007.
- Nathaniel Baum-Snow, Loren Brandt, J Vernon Henderson, Matthew A Turner, and Qinghua Zhang. Roads, railroads, and decentralization of Chinese cities. *Review of Economics and Statistics*, 99(3):435–448, 2017.
- 7005 Patrick Bayer, Fernando Ferreira, and Robert McMillan. A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638, 2007.
- Christopher R. Berry. Reassessing the property tax. Available at SSRN 3800536, 2021.

- 7010 Sandra E. Black. Do better schools matter? parental valuation of elementary education. *Quarterly Journal of Economics*, 114(2):577–599, 1999.
- Hoyt Bleakley and Jeffrey Lin. Portage and path dependence. *Quarterly Journal of Economics*, 127(2):587–644, 2012.
- Dan Bogart, Alan Rosevear, and Leigh Shaw-Taylor. Did turnpikes make roads better? Technical report, UC Irvine, working paper, 2024.
- Jutta Bolt and Jan Luiten Van Zanden. The maddison project: Collaborative research on historical national accounts. *Economic History Review*, 67(3):627–651, 2014.
- Leah Boustan, Devin Bunten, and Owen Hearey. Urbanization in american economic history, 1800–2000. *The Oxford Handbook of American Economic History*, 2:75, 2018.
- 7020 Leah Platt Boustan. Was postwar suburbanization "White Flight"? Evidence from the black migration. *Quarterly Journal of Economics*, 125(1):417–443, 2010.
- Leah Brooks and Zachary Liscow. Infrastructure costs. *American Economic Journal: Applied Economics*, 15(2):1–30, 2023.
- 7025 Jan K. Brueckner. The structure of urban equilibria: A unified treatment of the muth-mills model. *Handbook of Regional and Urban Economics*, 2(20):821–845, 1987.
- Gharad Bryan, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh. 7030 *Econometrica*, 82(5):1671–1748, 2014.

Gerald Carlino and William R. Kerr. Agglomeration and innovation. *Handbook of Regional and Urban Economics*, 5:349–404, 2015.

Raj Chetty, Nathaniel Hendren, and Lawrence Katz. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 106(4):855–902, 2016.

Colin Clark. Urban population densities. *Journal of the Royal Statistical Society. Series A (General)*, 114(4):490–496, 1951.

Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon. Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, 63(2):723–742, 2008.

Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon. The costs of agglomeration: House and land prices in French cities. *The Review of Economic Studies*, 86(4):1556–1589, 2019.

Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon. The production function for housing: Evidence from France. *Journal of Political Economy*, 2020.

Michael Coury, Toru Kitagawa, Allison Shertzer, and Matthew A. Turner. The value of piped water and sewers: Evidence from 19th century Chicago. 2021.

Victor Couture and Jessie Handbury. Urban revival in America, 2000 to 2010. Technical report, National Bureau of Economic Research, 2017.

Victor Couture and Jessie Handbury. Urban revival in America. *Journal of Urban Economics*, 119:103267, 2020.

- Victor Couture, Gilles Duranton, and Matthew A. Turner. Speed. *Review of Economics and Statistics*, 100(4):725–739, 2018.
- David M. Cutler and Edward L. Glaeser. Are ghettos good or bad? *Quarterly Journal of Economics*, 112(3):827–872, 1997.
- 7055 David M. Cutler, Edward L. Glaeser, and Jacob L. Vigdor. The rise and decline of the American ghetto. *Journal of Political Economy*, 107(3):455–506, 1999.
- Ben Dachis, Gilles Duranton, and Matthew A. Turner. The effects of land transfer taxes on real estate markets: Evidence from a natural experiment in Toronto. *Journal of Economic Geography*, 12(2):327–354, 2012.
- 7060 Morris A. Davis and François Ortalo-Magné. Household expenditures, wages, rents. *Review of Economic Dynamics*, 14(2):248–261, 2011.
- Jorge de la Roca and Diego Puga. Learning by working in big cities. *Review of Economic Studies*, 84(1):106–142, 2017.
- André De Palma, Yorgos Y Papageorgiou, Jacques-François Thisse, and Philip 7065 Ushchev. About the origin of cities. *Journal of Urban Economics*, 111:1–13, 2019.
- Jan De Vries. *European Urbanization, 1500-1800*. Routledge, 2013.
- Gilles Duranton. Urban evolutions: The fast, the slow, and the still. *American Economic Review*, 97(1):197–221, 2007.
- 7070 Gilles Duranton and Henry G. Overman. Testing for localization using microgeographic data. *Review of Economic Studies*, 72(4):1077–1106, 2005.

Gilles Duranton and Diego Puga. Diversity and specialisation in cities: Why, where and when does it matter? *Urban studies*, 37(3):533–555, 2000.

Gilles Duranton and Diego Puga. Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review*, 91(5):1454–1477, 2001.

⁷⁰⁷⁵ Gilles Duranton and Diego Puga. Urban growth and its aggregate implications. *Econometrica*, 91(6):2219–2259, 2023.

Gilles Duranton and Matthew A. Turner. Urban growth and transportation. *Review of Economic Studies*, 79(4):1407–1440, 2012.

⁷⁰⁸⁰ Jean Eid, Henry G. Overman, Diego Puga, and Matthew A. Turner. Fat city: Questioning the relationship between urban sprawl and obesity. *Journal of Urban Economics*, 63(2):385–404, 2008.

Glenn Ellison and Edward L. Glaeser. Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105(5):889–927, 1997.

⁷⁰⁸⁵ Reid Ewing and Barbara McCann. Measuring the health effects of sprawl: A national analysis of physical activity, obesity and chronic disease. 2003.

Pablo D Fajgelbaum and Cecile Gaubert. Optimal spatial policies, geography, and sorting. *Quarterly Journal of Economics*, 135(2):959–1036, 2020.

⁷⁰⁹⁰ Pablo D Fajgelbaum, Cecile Gaubert, Nicole Gorton, Eduardo Morales, and Edouard Schaal. Political preferences and the spatial distribution of infrastructure: Evidence

- from California's high-speed rail. Technical report, National Bureau of Economic Research, 2023.
- Masahisa Fujita and Hideaki Ogawa. Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics*, 12
7095 (2):161–196, 1982.
- Xavier Gabaix. Zipf's law for cities: An explanation. *Quarterly Journal of Economics*, 114(3):739–767, 1999.
- Miquel-Àngel Garcia-López. All roads lead to Rome... and to sprawl? Evidence from European cities. *Regional Science and Urban Economics*, 79:103467, 2019.
- 7100 Miquel-Àngel Garcia-López, Adelheid Holl, and Elisabet Viladecans-Marsal. Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities. *Journal of Urban Economics*, 85:52–67, 2015.
- Miquel-Àngel Garcia-López, Ilias Pasidis, and Elisabet Viladecans-Marsal. Suburbanization and transportation in European cities. *Journal of Economic Geography*,
7105 2024.
- Edward L. Glaeser and Joshua D. Gottlieb. The wealth of cities: Agglomeration economies and spatial equilibrium in the united states. *Journal of Economic Literature*, 47(4):983–1028, 2009.
- Edward L Glaeser and David C Maré. Cities and skills. *Journal of Labor Economics*,
7110 19(2):316–342, 2001.

Edward L. Glaeser, Jed Kolko, and Albert Saiz. Consumer city. *Journal of Economic Geography*, 1(1):27–50, 2001.

Edward L. Glaeser, Matthew E. Kahn, and Jordan Rappaport. Why do the poor live in cities? The role of public transportation. *Journal of Urban Economics*, 63(1):
7115 1–24, 2008.

Marco Gonzalez-Navarro and Matthew A. Turner. Subways and urban growth: Evidence from earth. *Journal of Urban Economics*, 108:85–106, 2018.

Arpit Gupta, Vrinda Mittal, Jonas Peeters, and Stijn Van Nieuwerburgh. Flattening the curve: pandemic-induced revaluation of urban real estate. *Journal of Financial Economics*, 146(2):594–636, 2022.
7120

Michael R. Haines. The urban mortality transition in the United States, 1800–1940. In *Annales de démographie historique*, number 1, pages 33–64, 2001.

John R. Harris and Michael P. Todaro. Migration, unemployment and development: A two-sector analysis. *American Economic Review*, pages 126–142, 1970.

7125 Stephan Hebllich, Stephen J Redding, and Daniel M Sturm. The making of the modern metropolis: evidence from london. *The Quarterly Journal of Economics*, 135(4):2059–2133, 2020.

J. Vernon Henderson. The sizes and types of cities. *American Economic Review*, pages 640–656, 1974.

7130 J. Vernon Henderson and Matthew A. Turner. Urbanization in the developing world: Too early or too slow? *Journal of Economic Perspectives*, 34(3):150–73, 2020.

J. Vernon Henderson, Tanner Regan, and A. Venables. Building the city: Urban transition and institutional frictions. *Processed London School of Economics*, 2018a.

J Vernon Henderson, Tim Squires, Adam Storeygard, and David Weil. The global
7135 distribution of economic activity: Nature, history, and the role of trade. *Quarterly Journal of Economics*, 133(1):357–406, 2018b.

Vernon Henderson. The urbanization process and economic growth: The so-what question. *Journal of Economic Growth*, 8(1):47–71, 2003.

Vernon Henderson, Ari Kuncoro, and Matt Turner. Industrial development in cities.

7140 *Journal of Political Economy*, 103(5):1067–1090, 1995.

Thomas J. Holmes. Localization of industry and vertical disintegration. *Review of Economics and Statistics*, 81(2):314–325, 1999.

Thomas J. Holmes and Sanghoon Lee. Cities as six-by-six-mile squares: Zipf’s law?

In Edward L. Glaeser, editor, *Agglomeration economics*, pages 105–132. University
7145 of Chicago Press, 2010.

Chang-Tai Hsieh and Enrico Moretti. Housing constraints and spatial misallocation.

American Economic Journal: Macroeconomics, 11(2):1–39, 2019.

Wen-Tai Hsu, Thomas J Holmes, and Frank Morgan. Optimal city hierarchy: A dynamic programming approach to central place theory. *Journal of Economic Theory*, 154:245–273, 2014.

Yannis M Ioannides and Henry G Overman. Zipf’s law for cities: an empirical examination. *Regional Science and Urban Economics*, 33(2):127–137, 2003.

Yannis M. Ioannides, Henry G. Overman, Esteban Rossi-Hansberg, and Kurt Schmidheiny. The effect of ICT on urban structure. *Economic Policy*, 23(54):201–242, 2008.

7155

Lawrence F. Katz, Jeffrey R. Kling, and Jeffrey B. Liebman. Moving to opportunity in Boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics*, 116(2):607–654, 2001.

7160

Stephen F. LeRoy and Jon Sonstelie. Paradise lost and regained: Transportation innovation, income, and residential location. *Journal of Urban Economics*, 13(1):67–89, 1983.

Robert E. Lucas. Externalities and cities. *Review of Economic Dynamics*, 4(2):245–274, 2001.

7165

Robert E Lucas and Esteban Rossi-Hansberg. On the internal structure of cities. *Econometrica*, 70(4):1445–1476, 2002.

Alfred Marshall. *Principles of economics*. Springer, 2013.

Neil Mehrotra, Matthew A. Turner, and Juan Pablo Uribe. Does the US have an infrastructure cost problem? Evidence from the interstate highway system. *Journal of Urban Economics*, 143:103681, 2024.

7170

Guy Michaels and Ferdinand Rauch. Resetting the urban network: 117–2012. *Economic Journal*, 128(608):378–412, 2018.

Peter Mieszkowski and Barton Smith. Analyzing urban decentralization: The case of Houston. *Regional Science and Urban Economics*, 21(2):183–199, 1991.

- Enrico Moretti. Human capital externalities in cities. In *Handbook of Regional and Urban Economics*, volume 4, pages 2243–2291. Elsevier, 2004.
- Enrico Moretti. The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, 111(10):3328–75, 2021.
- Kaivan Munshi. Community networks and the process of development. *Journal of Economic Perspectives*, 28(4):49–76, 2014.
- William D. Nordhaus. Geography and macroeconomics: New data and new findings. *Proceedings of the National Academy of Sciences*, 103(10):3510–3517, 2006.
- Nathan Nunn and Nancy Qian. The potato's contribution to population and urbanization: Evidence from a historical experiment. *Quarterly Journal of Economics*, 126(2):593–650, 2011.
- Jeff Nussbaum. The night New York saved itself from bankruptcy. *The New Yorker*, October 16, 2015.
- Hideaki Ogawa and Masahisa Fujita. Equilibrium land use patterns in a nonmonocentric city. *Journal of Regional Science*, 20(4):455–475, 1980.
- Oded Palmon and Barton A. Smith. New evidence on property tax capitalization. *Journal of Political Economy*, 106(5):1099–1111, 1998.
- Stephen J. Redding and Matthew A. Turner. Transportation costs and the spatial organization of economic activity. *NBER Working Paper*, (w20235), 2014.
- Jennifer Roback. Wages, rents, and the quality of life. *Journal of Political Economy*, 90(6):1257–1278, 1982.

- 7195 Kenneth Rosen and Mitchell Resnick. The size distribution of cities: An examination of the Pareto Law and primacy. *Journal of Urban Economics*, 8(2):165–186, 1980.
- Kenneth T. Rosen. The impact of proposition 13 on house prices in Northern California: A test of the interjurisdictional capitalization hypothesis. *Journal of Political Economy*, 90(1):191–200, 1982.
- 7200 Sherwin Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974.
- Stuart S. Rosenthal and William C. Strange. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics*, 85(2):377–393, 2003.
- Stuart S. Rosenthal and William C. Strange. Evidence on the nature and sources 7205 of agglomeration economies. In *Handbook of Regional and Urban Economics*, volume 4, pages 2119–2171. Elsevier, 2004.
- Esteban Rossi-Hansberg and Mark L.J. Wright. Urban structure and growth. *Review of Economic Studies*, 74(2):597–624, 2007.
- Richard Rothstein. *The color of law: A forgotten history of how our government segregated America*. Liveright Publishing, 2017.
- 7210 Allen J. Scott. World development report 2009: Reshaping economic geography, 2009.
- Holger Sieg. *Urban Economics and Fiscal Policy*. Princeton University Press, 2020.
- Paramita Sinha, Martha Caulkins, and Maureen Cropper. The value of climate amenities: A comparison of hedonic and discrete choice approaches. *Journal of Urban Economics*, 126, 2021.
- 7215

Kwok Tong Soo. Zipf's law for cities: A cross-country investigation. *Regional Science and Urban Economics*, 35(3):239–263, 2005.

Jacques-François Thisse, Matthew A Turner, and Philip Ushchev. Foundations of cities. *Journal of Urban Economics*, 143:103684, 2024.

7220 Kelsey L. Thomas, Elizabeth A. Dobis, and David . McGranahan. The nature of the rural-urban mortality gap (report no. EIB-265). Technical report, U.S. Department of Agriculture, Economic Research Service, 2024.

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

7225 Matthew A. Turner. Landscape preferences and patterns of residential development. *Journal of Urban Economics*, 57(1):19–54, 2005.

World Health Organization. Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines. 2017.