CAF
DOCUMENTOS DE TRABAJO

CAF
WORKING PAPERS

PEOPLE OR PLACES: THE CAUSES OF SPATIAL INCOME DIFFERENCES IN BRAZIL

N° 2010/04

May, 2010

Clemens, Michael
Ortega, Daniel

CAF - Ave. Luis Roche, Torre CAF, Altamira. Caracas, Venezuela 01060

© CAF, 2010 por Clemens, Michael y Ortega, Daniel. Todos los derechos reservados. Pequeñas secciones del texto, menores a dos párrafos, pueden ser citadas sin autorización explícita siempre que se cite el presente documento.
Los resultados, interpretaciones y conclusiones expresados en esta publicación son de exclusiva responsabilidad de su(s) autor(es), y de ninguna manera pueden ser atribuidos a CAF, a los miembros de su Directorio Ejecutivo o a los países que ellos representan.
CAF no garantiza la exactitud de los datos incluidos en esta publicación y no se hace responsable en ningún aspecto de las consecuencias que resulten de su utilización.

PERSONAS O LUGARES: LAS CAUSAS DE LAS DIFERENCIAS ESPACIALES DEL INGRESO EN BRASIL
Clemens, Michael y Ortega, Daniel
CAF Documento de trabajo N° 2010/04
Mayo, 2010

RESUMEN

Este estudio explora las implicaciones empíricas de un simple modelo de heterogeneidad espacial de ingresos, un modelo en que la heterogeneidad espacial surge de diferencias en los lugares donde trabajan los individuos y de las diferencias entre los propios individuos. Usando data a nivel micro de cuatro ediciones del Censo de Brasil (desde 1970 hasta 2000), estimamos una prima asociada al lugar de residencia de entre 20% y 80% de los salarios, después de controlar por diferencias observables y no observables entre migrantes y no migrantes. Derivamos implicaciones de tres causas potenciales de esta relación: causalidad de ingresos por lugar, selección de migrantes y externalidades sobre no migrantes. La evidencia es incompatible con la noción de que la selección o externalidades son los principales determinantes de la relación entre movimiento e ingresos. Esto sugiere que incluso dentro de las fronteras de un país, grandes diferencias de productividad pueden surgir y persistir en el tiempo.

Palabras clave: heterogeneidad espacial, migración, Brasil

PEOPLE OR PLACES: THE CAUSES OF SPATIAL INCOME DIFFERENCES IN BRAZIL
Clemens, Michael y Ortega, Daniel
CAF Working paper N° 2010/04
May, 2010

ABSTRACT

This paper explores testable implications of a simple model of spatial heterogeneity in earnings, a model in which this spatial heterogeneity arises both from differences in the places that people work and differences among people themselves. Using micro data from four waves of the Brazilian Census (from 1970 to 2000), we estimate a place premium of between 20% and 80% of wages after controlling for observable and unobservable differences between migrants and non migrants. We derive testable implications of three different potential causes of this relationship: causation of earnings by location, migrant selection, and externalities on non migrants. A range of evidence is incompatible with the notion that selection or externalities are principal determinants of the movement - earnings relationship. This suggests that even within country's borders, very large productivity differences may arise and be persistent over time.

Michael Clemens
Center for Global Development
mclemens@cgdev.org

Daniel Ortega
Instituto de Estudios Superiores de Administración
Corporación Andina de Fomento
dortega@caf.com

Keywords: spatial heterogeneity, migration, Brazil

People or Places: The causes of spatial income differences in Brazil

Michael A. Clemens
Center for Global Development

Daniel Ortega
Instituto de Estudios Superiores
de Administración, and
Corporación Andina de Fomento

May 30, 2010

Abstract: This paper explores testable implications of a simple model of spatial heterogeneity in earnings, a model in which this spatial heterogeneity arises both from differences in the places that people work and differences among people themselves. Using micro data from four waves of the Brazilian Census (from 1970 to 2000), we estimate a place premium of between 20% and 80% of wages after controlling for observable and unobservable differences between migrants and non migrants. We derive testable implications of three different potential causes of this relationship: causation of earnings by location, migrant selection, and externalities on non-migrants. A range of evidence is incompatible with the notion that selection or externalities are principal determinants of the movement-earnings relationship. This suggests that even within a country's borders, very large productivity differences may arise and be persistent over time.

JEL codes J61, O15, R23

1 Introduction

The value of workers' labor depends crucially on where they work. Location matters most in developing countries. Workers typically earn much more in the richest state than the poorest state: five times more in Brazil (IBGE 2009), nine times more in India (GoI 2009). What causes these differences is an important outstanding question in development economics.

It is an important question for two reasons. It is important for research because any answer leads to other questions: If differences in income arise primarily from differences in places, the puzzle is why more people do not migrate. If income differences arise primarily from differences in people—because the most able workers congregate, or because migration impoverishes non-migrants—the puzzle is why so many people do migrate. The question is important for policy as well, because if differences in income arise primarily from differences in places, measures to encourage or discourage internal migration could have first-order effects on individual and national income.

Here we address this question with a simple model and microdata from Brazil. We document that migrants earn much more than observably identical non-migrants, and ask which theoretical mechanisms are capable of producing this pattern. The data allow us to test the predictions of three different mechanisms. First, some trait of the destination locality could raise migrants' productivity. Second, people with better earning ability could be more likely to migrate. Third, migrants' emigration could lower the earnings of non-migrants. Each of these mechanisms has observable implications. The data do not broadly support the predictions of the second and third mechanisms. This suggests that spatial income differences in Brazil are primarily caused by differences in place rather than differences in people—that is, they mostly represent a "place premium" (Clemens, Montenegro, and Pritchett 2008).

We focus on Brazil for several reasons. First, Brazil's regional income disparities and internal migration flows are large. Second, the census of Brazil contains detailed information on income

and has done so for decades, unlike many developing countries. Third, there are no explicit policy barriers to internal migration in Brazil, unlike in other countries like China and Russia. Finally, a substantial literature has discussed the economic effects of migration in Brazil.

2 The correlation between distance moved and earnings

Our first step is to measure the relationship between movement and earnings in Brazil. We start with census data on earnings and location, regress earnings on standard observable traits, and calculate a wage residual for each worker. This allows us to explore semiparametrically the relationship between distance moved and earnings for observably identical workers.

2.1 *Data*

The data on workers comes from nationally-representative, individual-level, cross-sectional data from four censuses of Brazil, taken in 1970, 1980, 1991, and 2000, and released by the Integrated Public Use Microdata Series (IPUMS) project at the Minnesota Population Center. A random sample of three million individual records was taken from each census to ease computational burden.[1] Of these, people outside the age range 15-64 are dropped. Each record gives the state of residence and state of birth of each worker, as well as the state of residence five years prior to the census (for the 1991 and 2000 censuses only).

The records do not indicate the exact distance between each person's place of residence and place of origin. We approximate the distance between a worker's origin and current place of residence by a measure of the state-to-state distance between the state of origin and state of residence.[2]

[1] This corresponds to 60.6% of the 4,953,759 individual records contained in the full 1970 census file, 51.1% of 5,870,467 records in the 1980 file, 35.2% of 8,522,740 records in the 1991 file, and 29.6% of 10,136,022 records in the 2000 file. We drop people born on the Fernando de Noronha islands.

[2] The census files give the municipality of residence for 1980, 1991, and 2000, but only state of residence is available for all years 1970-2000. None of the census files give municipality of birth or municipality of previous residence; only the state is given.

To minimize the error this introduces, we measure state-to-state distances as the great-circle distances between the population-weighted centroids of each state pair (Head and Mayer 2002; Mayer and Zignago 2006). The latitude, longitude, and population of n most populous cities in each state come from the MaxMind (2008) World Cities Database, where $n = min\{n', 25\}$ and n' is the number of cities in each state described by MaxMind. We convert latitude and longitude into radians and calculate the distance from city k' in one state of each pair to city l' in the other state of the pair as $dist_{k}^{'}l' = \mathcal{E} \times (\arccos[\sin(\lambda_{k}^{'})\sin(\lambda_{l}^{'}) + \cos(\lambda_{k}^{'})\cos(\lambda_{l}^{'})\cos(\Lambda_{l}^{'} - \Lambda_{k}^{'})])$, where λ is latitude, Λ is longitude, and $\mathcal{E} = 6372.795$ km is the radius of the Earth. The weighted distance between state k and state l is then estimated as $dist_{kl} = \sum_{k' \in k} \left[\frac{(\pi_{k}^{'})}{P_k}\right] \sum_{l' \in l} \left[\frac{(\pi)_{l}^{'}}{P_l}\right] dist_{k}^{'}l'$, where π is city population and P is the sum of the populations of the n cities used in each state.

2.2 *Controlling for observable differences*

We wish to compare the earnings of observably identical workers who have moved different distances from a state of origin k to a state of destination l. The observable traits we use are state of origin, state of destination, educational attainment, age, sex, and urban or rural residence. In the basic analysis we estimate $k \times l$ separate regressions, one for each state dyad, of the form

$$\ln y_i = \beta_0 + \boldsymbol{S_i}\boldsymbol{\beta_1} + \boldsymbol{A_i}\boldsymbol{\beta_2} + \beta_3 w_i + \boldsymbol{\beta_4} \boldsymbol{r_i} + \varepsilon_i \qquad (1)$$

where y_i is total income from all sources in the previous month, $\boldsymbol{S_i}$ is a vector of dummy variables for years of schooling, $\boldsymbol{A_i}$ is a vector of dummy variables for age, w_i is a dummy for female, $\boldsymbol{r_i}$ is a vector of dummy variables for race, the β's are parameters or vectors of parameters to be estimated, and ε_i is a residual. We include both migrants and non-migrants in

the regression; non-migrants simply have the same state of origin and destination. We only include people with nonzero income between the ages of 15 and 64 inclusive.

The regression specification is very flexible. Running $k \times l$ separate regressions allows all personal characteristics to have difference returns for each state dyad. For example, it does not assume that the returns to being a São Paulo born woman resident in São Paulo are the same as the returns to being a Piauí born woman resident in São Paulo. And the use of dummies avoids more restrictive assumptions about the functional form of the schooling-earnings and age-earnings relationships.

Below we analyze the relationship between the wage residuals ε_i and distance from the origin, both parametrically and nonparametrically. In the basic analysis "origin" means state of birth, but in alternative specifications "origin" means state of residence five years before the census.

2.3 *Earnings rise with distance moved*

Figure 1 shows semiparametric regressions of wages on distance moved, in each of the last four censuses of Brazil. The black line shows all workers in the census. The horizontal axis is an estimate of how far each person lives from his or her birthplace. It is the population-weighted distance between each person's state of birth and state of residence, with zero for those who reside in their state of birth.

The vertical axis of the figure shows the earnings residual for that person in a regression (1) including other people born in the same state as that person and residing in the same state as that person, relative to all those residing in the same state. It is converted from log earnings back to arithmetic multiples of earnings, where 1.0 is the earnings of an observably identical non-migrant. That is, the vertical axis is $e^{\varepsilon_{ikl} - \bar{\varepsilon}_l}$, where ε_{ikl} is the residual from (1) for person i living in state k and born in state l, and $\bar{\varepsilon}_l$ is the average of those residuals for people living in state l.

In all years, earnings are higher for those who move out of their states of birth relative to observably identical workers (same age, education, gender, and race) who do not move. The slope of this relationship is similar across a broad range of distances. Workers living over 2000km from their states of birth earn roughly 60-80% more than workers living in their states of birth. This hold in 1970, 1990, and 2000; the relationship is slightly flatter in 1980. The four waves of the census depicted here show that these large productivity differences associated to places of work within Brazil not only exist but are very persistent over long periods of time.

This correlation is a distinct fact; its interpretation is more complex. The remainder of this paper generates and tests predictions arising from different theoretical mechanisms capable of generating the distance-earnings correlation.

3 Three models of the distance-earnings relationship

We will explore testable implications of a simple model of spatial heterogeneity in earnings, a model in which this spatial heterogeneity arises both from differences in the places that people work and differences among people themselves. Below we explore three separate theoretical mechanisms for the distance-earnings correlation: that movement affects earnings, that migrants are selected on correlates of earnings, and that migrants impose costs on non-migrants. Each has testable implications.

3.1 *The effect of movement on earnings*

First, suppose that all workers are identical and that no worker's movement has an external effect on any other worker (both will be relaxed later). A worker's productivity depends only on the quality of the place he or she works, p , which varies across a one-dimensional country. This local quality could reflect the quality of economic institutions, geographic conditions, or anything else particular to the locality that affects workers' productivity. Workers are paid their

marginal productivity. The worker starts out at one end of the linear country and must decide whether or not to move, and if to move, how far to move.

The worker has one work opportunity at the origin, where local conditions have quality $p_0 \sim U[0,1]$. There are additional opportunities further away along the line with a density of δ opportunities per unit distance. Define units such that there is one opportunity per unit distance: $\delta \equiv 1$. At each distance x from the origin there is thus one new opportunity where the quality of local conditions is $p_x \sim U[0,1]$, independently and identically distributed. A map of these opportunities might look like Figure 2.

Consider the worker's movement decision as a simple optimal stopping problem. Moving further away from home increases the chance that the worker will find excellent local conditions for work, but suppose that each unit of movement obliges the worker to bear a cost c such that $0 \leq c < 1$. The worker's income y is simply equal to the quality of local conditions, thus $y(p_0) = p_0$.

The best local quality that the worker discovers in x units of movement from the origin is given by a standard statistic of the uniform distribution: $y_x \equiv E[p_x | p_x = \max_{i \leq x}\{p_i\}] = \frac{x}{1+x}$. If the worker is to move, then, the optimal distance is the distance at which the marginal increase in the expected value of the best local condition is equal to the marginal cost of moving: $dy_x/dx = c$, so

$$x^* = \frac{1}{\sqrt{c}} - 1 \qquad (2)$$

Workers thus move x^* units if the expected value of the local conditions that distance from the origin, net of the cost of moving, exceeds the value of conditions at the origin. That is, workers move if $V_x \equiv y_{x^*} - cx^* > p_0$, or

$$p_0 < 1 + c - 2\sqrt{c} \qquad (3)$$

Thus workers who begin in places where p_0 is sufficiently high do not move and earn $y = p_0$, and those who begin in places where p_0 is sufficiently low move x^* units from the origin and earn $y_{x^*} = 1 - \sqrt{c}$. For those who move, the effect of movement on earnings is to raise them by a factor of y_{x^*}/p_0. The factor by which earnings are raised per unit distance moved is therefore $\gamma \equiv \frac{y_{x^*}}{\frac{y_0}{x^*}}$, or

$$\gamma = \frac{\sqrt{c}}{p_0} \tag{4}$$

If we assume that the number of movers is small relative to the whole population, then (4) also reflects the slope of the observed *correlation* between the relative earnings gain by movers per unit distance moved.

So far, the model with homogeneous workers yields several predictions:

1. *Earnings rise with distance moved*, because by (4), $\gamma > 0$ as long as $c > 0$.
2. *The higher are movement costs, the steeper is the earnings gain per unit distance moved*, since by (4), $d\gamma/dc = 1/2p_0\sqrt{c} > 0$.
3. *With higher movement costs, a smaller fraction of workers move,* since $p_0 \sim U[0,1]$ and by (3) the fraction of workers who move is $f = 1 + c - 2\sqrt{c}$, thus $df/dc < 0$.
4. *With higher movement costs, those who do move, move shorter distances*, since by (2), $dx^*/dc < 0$.

3.2 *Migrant selection*

Migrant selection can explain part of spatial differences in earnings if we relax the assumption that all workers are identical. Suppose that each worker has an inborn quality $q \in [0,1]$ that affects her productivity in any given place. As in Kremer's (1993) "O-ring" theory of production, the degree to which quality affects productivity is determined by a parameter $\theta \in [0,1]$ reflecting the complexity of the task performed, so that $y_x = q^\theta \left(x/(1+x)\right)$. Now, equation (2) becomes

$$x^* = \sqrt{\frac{q^\theta}{c}} - 1 \tag{5}$$

Analogously to (3), workers will move if $y_{x^*} - cx^* > q^\theta p_0$, or

$$p_0 < 1 + \frac{c}{q^\theta} - 2\sqrt{\frac{c}{q^\theta}} \tag{6}$$

Equation (4), the factor by which earnings are raised per unit distance moved, now becomes

$$\gamma = \frac{\sqrt{c}}{p_0 q^{\frac{3}{2}\theta}} \tag{7}$$

All four of the previous predictions remain true, but we acquire additional predictions from this slightly richer model:

5. *Higher quality workers are more likely to move than lower quality workers*, since by (6), the fraction who leave $f = 2\sqrt{\frac{c}{q^\theta}} - \frac{c}{q^\theta}$, and thus $df/dq > 0$ provided $q^\theta > c$.
6. *Higher quality workers who do move, move further than lower quality workers*, since by (5), $dx^*/dq > 0$.

But there is an important difference between the present case and the case of homogenous workers: Now, the observed correlation between movement and earnings may not represent the *effect* of movement on earnings. To see why, note that condition (6) implies that workers move if they have quality $q > \bar{q} \equiv \left(c/(1+\sqrt{p_0})^2\right)^{1/\theta}$, and stay at the origin otherwise. Now let $\overline{q}$ be the average quality of those who do move and $\underline{q}$ be the average quality of those who do not leave; that is, $\overline{q} \equiv E[q|q > \bar{q}]$ and $\underline{q} \equiv E[q|q \leq \bar{q}]$. The observed earnings multiple per unit distance moved is $\hat{\gamma} \equiv \frac{y_{x^*}(\overline{q})}{y_0(\underline{q})/x^*(\overline{q})}$ where $y_0(\underline{q}) = \underline{q}^\theta p_0$, thus

$$\hat{\gamma} = \frac{\sqrt{c}}{p_0 \underline{q}^\theta \overline{q}^{\theta/2}} \tag{8}$$

but by (7), the *effect* of movement on the earnings of movers is $\gamma = \sqrt{c}/p_0 \overline{q}^{\frac{2}{2}\theta}$, and it is always true that the observed relationship between movement and earnings overstates the effect of movement on earnings: $\hat{\gamma} > \gamma$.

Intuitively, this is because higher quality workers both earn more and have a greater tendency to move. At greater distances from the origin we observe greater earnings, both because those who moved have increased their earnings by finding better local conditions and because those who move are more productive than those who do not move.

3.3 *External effects on non-migrants*

Earnings could also differ across space if migrants impose negative externalities on non-migrants. Return to the earlier assumption that workers are homogeneous. Suppose that the act of departure from the origin lowers earnings there, for example by eroding traditional institutions of production or insurance (e.g. Yap 1976; Glaeser and Redlick 2008; Gallego and

Mendola 2009). Average earnings at the origin are then some strictly decreasing function of the fraction of people who leave: $y(p_0) = p_0\phi(f)$ where $\phi(0) = 1$ and $\phi' < 0$. Now, the observed correlation between earnings and distance (4) becomes

$$\hat{\gamma} = \frac{\sqrt{c}}{p_0\phi(f)}. \tag{9}$$

As in the previous section, if anyone moves $(f > 0)$ then the observed correlation between earnings and distance exaggerates the true effect of movement on earnings: $\hat{\gamma} > \gamma$.

But this time the exaggeration arises by a separate mechanism. Suppose that all potential migrants decide noncooperatively, at time $t = 0$, whether or not to move at time $t = 1$. For migrants collectively, the average earnings gain caused by the ability to migrate is only $\gamma = \sqrt{c}/p_0$ from (4), because the counterfactual earnings for each worker if none of them could migrate would be p_0. But if anyone who wishes to migrate can do so, this lowers earnings for non-migrants to $p_0\phi(f)$, and observed earnings at the origin no longer serve as proper counterfactual earnings for the case of no migration. They are lower than true counterfactual earnings at the origin, so the observed earnings gap between destination and origin exaggerates the true effect of movement on earnings.

This model yields additional predictions:

7. *Average earnings are lower in high-emigration localities than in others, all else equal.* This is because $y(p_0) = p_0\phi(f)$ and $\phi' < 0$, with p_0 held constant.
8. *The correlation between distance and earnings overstates the effect of distance on earnings to the smallest degree for those who move the greatest distances.* This is

because $\bar{Y}/Y = 1/\phi(f)$, which implies $\left.\frac{d\left(\bar{Y}/Y\right)}{dx^*} = -(\phi(f))^{-2}\frac{d\phi}{df}\frac{df}{dc}\frac{dc}{dx^*} < 0\right.$, since we have already established $d\phi/df < 0$ and $df/dc > 0$, and by (2) we have $dc/dx^* < 0$.

Intuitively, prediction 8 comes from the fact that the negative externality caused by departure is independent of how far the migrant goes. Migrants who travel short distances do not raise their earnings by as much as migrants who travel long distances, but both do equal harm by leaving the origin, because this harm arises from their very absence. When migrants move short distances, a larger fraction of the earnings gap between migrants' earnings at the destination and non-migrants' earnings at the origin arises from the negative externality than when migrants move long distances.

4 Testing observable predictions of the models

The model shows how three different mechanisms could generate a positive relationship between the distance that migrant workers move and their earnings. First, workers searching for the best job opportunities across space can find better opportunities the further afield they are able to search. Second, those with the highest earning ability might move the furthest. Third, migrants' departure from the origin may exert a negative externality on non-migrants, lowering earnings at the origin.

Which of these three mechanisms is at work crucially determines the impact of movement on welfare. In the first mechanism, movement raises earnings for migrants and for the nation. In the second and third mechanisms, movement need not raise earnings for migrants or for the nation.

How much, of the observed correlation between movement and earnings represents the effect of movement on earnings? In other words, what is the degree of divergence between Y in (6)

and $\hat{\gamma}$ in (7)? The ideal empirical method to assess the difference would be to randomly apply exogenous movement and compare treatment and control—a method which is neither available nor desirable.

As described above, only workers with quality above a certain threshold decide to move. Denote the range of quality above the threshold by q^h and below by q^l, so $q^h = \{q \mid q > q^*\}$. The simple wage difference between workers who migrate ($M = 1$) and workers who don't ($M = 0$) is given by:

$$WAGE\_GAIN = E[W \mid M = 1, q^h] - E[W \mid M = 0, q^l]$$

which can be expressed as:

$$WAGE\_GAIN = E[W \mid M = 1, q^h] - E[W \mid M = 0, q^h] + E[W \mid M = 0, q^h] - E[W \mid M = 0, q^l]$$

the first two terms represent the wage gain due to migration, and the last two, the part of the simple difference in wages of migrants and non-migrants that is due to the fact that those who migrate have different quality than those who do not, this would be the bias in estimation procedures that do not account for such selection:

$$WAGE_{GAIN} = MIGRATION_{EFFECT} + SELECTION_{EFFECT} \quad (10)$$

Since all non-migrants have quality in the range q^l, it is not possible to estimate $E[W \mid M = 0, q^h]$ in order to quantify the selection effect. We therefore choose to look at alternative settings where either the selection effect is likely to be small (and $\hat{\gamma} \approx \gamma$), or where we can contrast the prediction of a pure selection effect with the data. This approach allows us to approximate bounds for the size of the migration effect, or the regional disparities in productivity.

The ideal scientific method to assess the effect γ of movement on earnings would be to randomly encourage movement. Randomized encouragement of domestic migration can be both ethical and practical (Katz, Kling, and Liebman 2001; Chowdhury, Mobarak, and Bryan 2009), but is only available very rarely and at great cost. A useful alternative is to identify empirical facts we would expect to observe if the correlation between movement and earnings arose primarily from the effect of movement on earnings $(\hat{\gamma} \approx \gamma)$, but which we would not expect to observe if the correlation arose primarily from selection or externalities $(\hat{\gamma} \neq \gamma)$, and vice versa. The model suggests four such facts.

4.1 *Comparing the distance-earnings relationship across occupations*

First, if migrant selection is a major determinant of the correlation between movement and earnings, this correlation should be more pronounced among migrants in occupations where interpersonal differences more greatly affect productivity.[3] This is suggested by (8), which implies $d\hat{\gamma}/d\theta > 0$.

In other words, the observed correlation between movement and earnings should overstate the effect of movement on earnings to the greatest degree for those executing the most complex tasks. The most menial and mechanical work (low θ) holds little prospect for interpersonal differences to greatly affect productivity, so earnings among workers in these occupations should display substantially less correlation with movement. If, on the other hand, places principally determine workers' productivity in all occupations $(\hat{\gamma} \approx \gamma)$, theory suggests we should expect a strong correlation between movement and earnings in all occupations.

[3] The International Standard Classification of Occupations 1988 designation of Elementary Occupations "covers occupations which require the knowledge and experience necessary to perform mostly simple and routine tasks, involving the use of hand-held tools and in some cases considerable physical effort, and, with few exceptions, only limited personal initiative or judgment. The main tasks consist of selling goods in streets, doorkeeping and property watching, as well as cleaning, washing, pressing, and working as labourers in the fields of mining, agriculture and fishing, construction and manufacturing." Source: http://www.ilo.org/public/english/bureau/stat/isco/isco88/publ4.htm

Figure 1 juxtaposes the movement-earnings correlation for workers in elementary occupations (in red) with that for all workers (in black). The relationships are similar. In 1970, elementary workers moving the greatest distances (over about 1500km) earn a premium about equal to that earned by workers in general. In 1980-2000, elementary workers moving the greatest distances earn a premium larger than that earned by workers in general. In all years, the premium for elementary workers moving shorter distances is less than that earned by workers in general, but only barely less.

Table 1 carries out more parametric tests of the same relationship. In each census year, the table shows the simple correlation between distance and earnings after controlling for the same basic observable traits. The next column then introduces an interaction term between distance moved and a dummy for elementary occupation. If selection on unobserved earnings determinants were lower in elementary occupations, and if such selection were a major determinant of the distance-earnings correlation, introduction of this interaction term would cause the coefficient on simple distance to rise substantially. But this is not the case. In no year does it rise to a degree that is statistically or economically significant. In two of the census years it falls, to a degree that is statistically but not economically significant. The distance-earnings relationship is substantially the same for elementary and non-elementary occupations.

Table 2 analyzes wage gains for male workers born in the Northeast who move to either São Paulo state or Rio de Janeiro state using the 2000 Census data. With all workers in the sample, the ratio of a migrant's wage to an observably identical non-migrant is about 1.57. The second row restricts the sample to include only workers in elementary occupations. The ratio falls to 1.44, suggesting that a small portion of the 1.57 ratio may arise from selection on unobservable wage correlates. But when the sample is restricted to those who have lived in their current state of residence for 20 years or more, the ratio rises to 1.75. In other words, the full-sample estimate of 1.57 slightly overestimates the wage gain of migrating from the Northeast to São Paulo or Rio due to selection on unobservables in occupations with high returns to unobservable skill, but slightly underestimates the wage gain because it includes recent

migrants who are not yet established. The fourth row bears out this conclusion by restricting the sample to elementary occupations *and* those who have lived in their current state of residence for 20 years or more: the ratio is 1.60.

If indeed migrant selection on unobservable determinants of earnings is substantially greater for non-elementary occupations than for elementary occupations, this evidence is incompatible with migrant selection being a major determinant of the distance-earnings correlation. This suggests that the selection effect of (9) is not very big, and may in fact in some cases be negative.

4.2 *Comparing migrants and non-migrants at the destination*

Second, if migrant selection is a major determinant of the correlation between movement and earnings ($\hat{\gamma} \neq \gamma$), we would expect to observe that migrants to any given destination should be more productive than the people who originated in that same place and did not move. This emerges from (6), which implies that a worker moves if

$$q > q^* = \left(\frac{\sqrt{c}}{1+\sqrt{p_0}}\right)^{\frac{2}{\theta}} \qquad (10)$$

again provided that $q^{\theta} > c$. Intuitively, greater migration costs (higher c), worse conditions at the origin (lower p_0), and more complex tasks (higher θ) all raise the relative tendency for the best workers to leave.

At any given destination, then, we will observe a mix of low q workers that did not move, and high q workers who moved to that destination from elsewhere. This means that migrants at the destination must have higher q on average than non-migrants, if pre-migration mean q is the same at the origin and destination. For example, if workers move from place A to place B, the ratio of quality q between migrants and non-migrants at the destination is $\bar{q}^A/\underline{q}^B > 1$

provided that the pre-migration distribution of q is the same at A and B, and the effect of q on a worker's productivity is independent of location.

What if worker quality is partially linked to location? In addition to capturing aspects of worker quality that are portable—such as intelligence or drive—q could also capture aspects of quality that are not portable. These might include locally-specific technical knowledge or connections to locally powerful families. Even in this case, the average quality of migrants at the destination must exceed the average quality of non-migrants at the destination, as long as some nonzero portion of quality is portable, and provided the destination has sufficiently good local conditions p_0.

To see this, assume again that the distribution of pre-migration quality at locations A and B is the same, and the average for both distributions is $\bar{q}$. Suppose that only a fraction α of a worker's quality is portable to the destination. Migrants' average quality at the destination is then $\alpha\bar{q}^A + (1-\alpha)\bar{q}$, and non-migrants' average quality at the destination is $\underline{q}^B$. By (10), if local conditions at the destination are sufficiently attractive, which is to say that p_0 is sufficiently close to unity at the destination, then high-q workers do not selectively leave the destination B, and $\underline{q}^B = \bar{q}$. The ratio of quality q between migrants and non-migrants at the destination is $(\alpha\bar{q}^A + (1-\alpha)\bar{q})/\bar{q} > 1$, provided $\bar{q}^A > \bar{q}$ and $\alpha > 0$.

Positive selection of emigrants from A still requires that quality of migrants at the destination B exceed the quality of non-migrants, as long as some of the migrants' quality is portable and local conditions at B are sufficiently attractive. There would be no such difference if the place itself made the migrants (and the locally originating non-movers) equally productive ($\hat{\gamma} \approx \gamma$).

Figure 3 tests whether workers who move greater distances typically possess different unobservable earnings determinants than non-migrants at the destination. The left-hand side of Figure 3 simply reproduces the semiparametric regression for the year 2000 from Figure 1,

whose vertical axis compares migrants' earnings to non-migrants' *at the origin* after controlling for age, education, and gender. The right-hand side of Figure 3 conducts the same exercise but compares migrants' earnings to non-migrants' *at the destination*, again after controlling for observables. The left side shows that people who leave their states of birth earn more than observably identical people who do not leave; the right side shows that people who *arrive* in a destination state from greater distances do not typically earn more than observably identical people born at the destination.

Table 2 also compares migrants' performance at the destination to native's performance at the destination, in the specific case of migrants from the Northeast to São Paulo state or Rio de Janeiro state. The fifth row of the table compares migrants' wages at the destination to wages of observably identical natives of the destination. Migrants have lower wages than natives, a difference that is not substantially different within elementary occupations, and which all but disappears after 20 years of residence. This is evidence against positive selection on unobservable wage correlates at the origin.

This evidence suggests that, if migrant selection on unobservable earnings determinants is a substantial cause of the distance-earnings correlation, that selection occurs on traits that are not portable from one state to another. For example, if native intelligence is an important unobserved determinant of earnings in any location, and people born in each state are typically equally endowed with native intelligence, this evidence is incompatible with a high degree of migrant selection on native intelligence.

4.3 *Comparing the earnings of the same individual at two points in time*

The above evidence suggests that migrant selection on *portable* unobserved determinants of earnings is not a major cause of the distance-earnings relationship. But the relationship could also be caused by selection on unobserved determinants of earnings that are not portable. Suppose for example that a worker's unobserved traits would give her higher-than-typical

earnings at the origin, and that she is more likely than other workers to move to a location where all workers have the same earnings. Her departure would mechanically reduce average earnings at the origin, but would not affect average earnings at the destination. This would tend to produce higher average earnings at the destination than at the origin, even though the effect of movement on earnings for this particular worker was negative.

Panel data would allow us to address this question empirically. Among observably identical workers present in a given state at one moment of time, are those with greater earnings more likely to subsequently migrate? Panel data of this kind are very rare, though they have been analyzed by Beegle et al. (2010) in Tanzania and Dercon et al. (2009) in India.

Here we use data from a vast and unique panel dataset on the full universe of formal-sector workers in Brazil, the formal employment census (Relação Anual de Informações Sociais, or RAIS) dataset collected by the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística, or IBGE). To computational burden, we use the 1% random sample of individuals in the RAIS data constructed and described by Aguayo-Tellez, Muendler, and Poole (2010). Individuals are linked across time by tax identification number, observed in December of each year between 1987 and 2001. The records describe wage, gender, age, education, occupation, state, and municipality for each worker.

Two strong advantages of these data are the multiple observations of the same workers tracked over space and time, and the fact that the underlying data constitute the full universe of formal-sector workers. The principal disadvantage of the data is that it includes only workers who are employed by firms formally registered with the government. It thus omits informal salespeople, construction workers, subsistence farmers, and the like.

Table 3 analyzes the subset of those workers who are observed with positive earnings at three points in time: December 1990, December 1995, and December 2000. The top half of the table simple regresses ln wages on dummy variables for location or movement without any

additional controls; the bottom half of the table introduces a set of dummy variables for gender, age, and education level.

The first column of Table 3 restricts the sample to workers in the Northeast, São Paulo state, or Rio de Janeiro state in 2000, and regresses wages on a dummy variable for location in either São Paulo or Rio. The coefficient shows that formal-sector workers in São Paulo or Rio earn about 50% more than observably identical workers formal-sector in the Northeast. This finding, for wages of formal sector workers, is statistically indistinguishable from the premium shown in the first row of Table 2, which comprises all workers in the census and reflects total earnings rather than wages. This suggests that the analysis is not greatly distorted by restricting attention to the wages of formal-sector workers.

The second column of Table 3 repeats the same regression but limits the sample to workers who were working in the Northeast in 1995. The coefficient on the São Paulo-Rio dummy does not substantially change. The wage premium for Northeasterners who move from the Northeast to São Paulo-Rio by 2000 is roughly the same as the premium for all those who are in São Paulo-Rio in 2000 relative to those in the Northeast.

The third column of Table 3 conducts an analysis analogous to Figure 3, asking whether those who moved from the Northeast to São Paulo-Rio between 1995 and 2000 earn more than observably identical workers at the destination. As in Figure 3 and Table 2, they do not; the movement dummy is statistically insignificant.

Finally, the fourth column of Table 3 asks whether the 1995 earnings of those in the Northeast who would later move to São Paulo-Rio by 2000 were different than the 1995 earnings of workers in the Northeast who did not later move. This is a rare chance to observe migrants prior to migration, in a sample relatively free of selection within the realm of formal-sector employees. After controlling for basic observable traits, the movement dummy is statistically significantly different from zero, suggesting some positive selection of migrants from the

Northeast to São Paulo-Rio on unobserved wage determinants. This positive selection appears to explain roughly one third of the correlation between movement and earnings.

The last three columns of Table 3 repeat the exercise of columns 2-4, using 1990 as the base year instead of 1995. The results are substantively identical with one important exception: the statistical significance of the movement dummy in the final column disappears in the bottom half of the table. That is, earnings of workers in the Northeast in 1990 who would later move to São Paulo-Rio by 2000 are no different from the earnings of observably identical workers who did not end up moving. Here, very little of the roughly 45% premium for those who moved appears to be explained by selection on unobservables. For both the 1990 and 1995 base years, the large majority of the movement-earnings relationship is not explained by selection on unobservables.

4.4 *Comparing the distance-earnings correlation across distance*

We should expect certain patterns in the relationship between distance and earnings if that relationship is substantially generated by the fact that migrants' absence lowers the earnings of non-migrants. People who leave a state are absent from the origin, whether they go to a nearby state or the furthest state. If this mechanism were entirely responsible for the distance-earnings relationship, the gap between migrants' earnings and non-migrants' earnings would be equal for all distances of migration.

If this mechanism were *partially* responsible for the distance-earnings relationship, the gap between migrants' earnings and non-migrants' earnings should grow quickly with distance at short distances, and grow more slowly with distance at long distances. This is because, as in prediction 8, $d\left(\hat{\gamma}/\gamma\right)/dx^{*} < 0$. The degree to which the distance-earnings correlation overstates the effect of distance on earnings should be greatest at the shortest distances. If the distance-

earnings correlation does not markedly diminish with distance, then, we should doubt that the correlation arises primarily from external effects of migrants' absence on non-migrants.

We do not observe patterns like this in the data. The semiparametric regressions in Figure 1 do not show substantial diminishing returns to distance moved, in any year, up to 2000km. If a major determinant of the distance-earnings relationship in the parametric regressions of Table 1 arose from harm to non-migrants by migrants' departure, we would not expect large nonlinearities in this relationship. Those who moved 300km would inflict much more harm than those who moved 0km, but those who moved 1,300km would not inflict much more harm than those who moved 1,000km. But no such pattern emerges from the data. This is incompatible with the notion that the distance-earnings relationship substantially arises from damage inflicted by movers on non-movers.

4.5 *Comparing return-migrants to non-migrants*

Finally, we expect to observe that return migrants' productivity differs substantially from that of never-migrants if migrant selection principally determines the correlation between movement and earnings. If the place itself determines productivity we should expect little difference in productivity between return migrants and never-migrants. The simple model above could generate return migration in various ways. For example, if migration cost c varied across individuals and over time, a worker might find it optimal to work in another location for a certain period and work at the origin thereafter.

Whether return migrants would be more or less productive than never-migrants is ambiguous and depends on the return-selection process. Higher-quality workers are more likely to emigrate in the first place, by (10), so there are three possibilities: a) If both selection into out-migration and selection into return-migration are both positive, return migrants must be more productive than never-migrants. b) If selection into out-migration is positive and selection into return-migration is negative and small, return migrants must be somewhat more productive

than never-migrants. c) If selection into return-migration is negative and large, return migrants must be less productive than never-migrants.

What if return migrants are exactly as productive as non-migrants? This requires two conditions: First, return migration must be negatively selected, to offset the positive selection of initial emigrants. Second, the degree of negative selection in return migration must exactly equal the degree of positive selection in initial emigration. If either of these conditions is not met, and if migrant selection is substantially responsible for the distance-earnings correlation, then we would expect to observe that return migrants differ substantially in productivity from never migrants *in one direction or the other*. On the other hand, if we do not observe substantial differences in productivity between return migrants and non-migrants, then there are two possibilities. Either selection is important and the two selection effects happen to precisely offset one another, or migrant selection is generally a less important determinant of the observed distance-earnings correlation than the effect of places on earnings.

Table 2 contains suggestive evidence in this regard. It analyzes return migrants: people who were born in the Northeast and reside in the Northeast, but resided in either São Paulo or Rio five years before the census. Compared to non-migrants, observably identical return migrants earn substantially less. This suggests that either 1) selection unobservable wage correlates for out-migration in the first place was not strongly positive, or 2) selection on unobservable wage correlation for return migration is strongly negative, or 3) something about migration experience causes sharply lower earnings in the home region, or some combination of these.

5 Conclusions

We document a large a systematic correlation between internal migration and the earnings of the full universe of labor-force participants in Brazil. This relationship has been remarkably stable for 30 years. We explore three separate theoretical explanations for this correlation between distance moved and earnings: First, movement itself might cause an increase in

earnings. Second, workers with greater inherent determinants of earnings might be selected into migration. Third, workers who move could impose negative externalities on workers who do not move. We develop empirical tests of each theory.

Those tests provide evidence that is not compatible with selection or externalities being major causes of the distance-earnings relationship. First, the relationship is substantially identical for workers in occupations where unobserved traits are unlikely to greatly affect earnings. Second, migrants' earnings at the destination are very similar those of observably identical non-migrants. Third, panel data on formal-sector workers reveal that, prior to migration, those who later migrate have only somewhat greater earnings than observably identical workers who do not later migrate; the large majority of the distance-earnings correlation is unexplained by the magnitude of that selection. Fourth, the slope of the distance-earnings relationship does not materially vary over a broad range of distances moved. This is incompatible with a large negative externality on non-migrants imposed by migrants' departure. Fifth, suggestive evidence from return migrants shows little sign of strong positive selection in outmigration.

Together, these suggest that there are large differences in workers' earnings caused by workers' spatial location in Brazil. In other words, internal migration in Brazil is an investment with large returns. Workers may be unable to take advantage of those returns because of barriers to movement, such as credit constraints (Chowdhury, Mobarak, and Bryan 2009), or because of high costs of information and assimilation into urban slums. Together, these findings suggest that Yap (1976) was correct to assess that "Brazil's urban policies, therefore, are better oriented toward alleviating urban poverty than toward reducing migration to cities".

References

Aguayo-Tellez, Ernesto, Marc-Andreas Muendler, and Jennifer Pamela Poole (2010), "Globalization and Formal-Sector Migration in Brazil", *World Development*, forthcoming.

Beegle et al. *Review of Economics and Statistics*

Chowdhury, Shyamal, Ahmed Mushfiq Mobarak, and Gharad Bryan (2009), "Migrating Away from a Seasonal Famine: A Randomized Intervention in Bangladesh", Human Development Research Paper 2009/41 (New York: United Nations Development Program).

Clemens, Michael A., Claudio E. Montenegro, and Lant Pritchett (2008), "The Place Premium: Wage Differences for Identical Workers across the U.S. Border", Working Paper 148 (Washington, DC: Center for Global Development).

Dahl 2002 *Econometrica*

Dercon et al. "Migration, Risk Sharing, and Subjective Well-Being"

Gallego and Mendola 2009 XXXXXXXXXXXXXXXXXXXXXXXXX

Glaeser and Redlick 2008 XXXXXXXXXXXXXXXXXXX

GoI [Government of India] (2009), *Economic Survey 2008-2009* (Delhi: Ministry of Finance). Table 1.8, page A13. http://indiabudget.nic.in/es2008-09/esmain.htm, accessed Dec. 15, 2009.

Head, Keith and Thierry Mayer, (2002), "Illusory Border Effects: Distance Mismeasurement Inflates Estimates of Home Bias in Trade," CEPII Working Paper No 2002-01 (Paris: Centre d'Études Prospectives et d'Informations Internationales).

IBGE [Instituto Brasileiro de Geografia e Estatística] (2009), *Contas Regionais do Brasil 2003-2007* (Brasília: Ministério do Planejamento). Tabela 4. http://www.ibge.gov.br/home/estatistica/economia/contasregionais/2003_2007/tabela04.pdf, accessed Dec. 15, 2009.

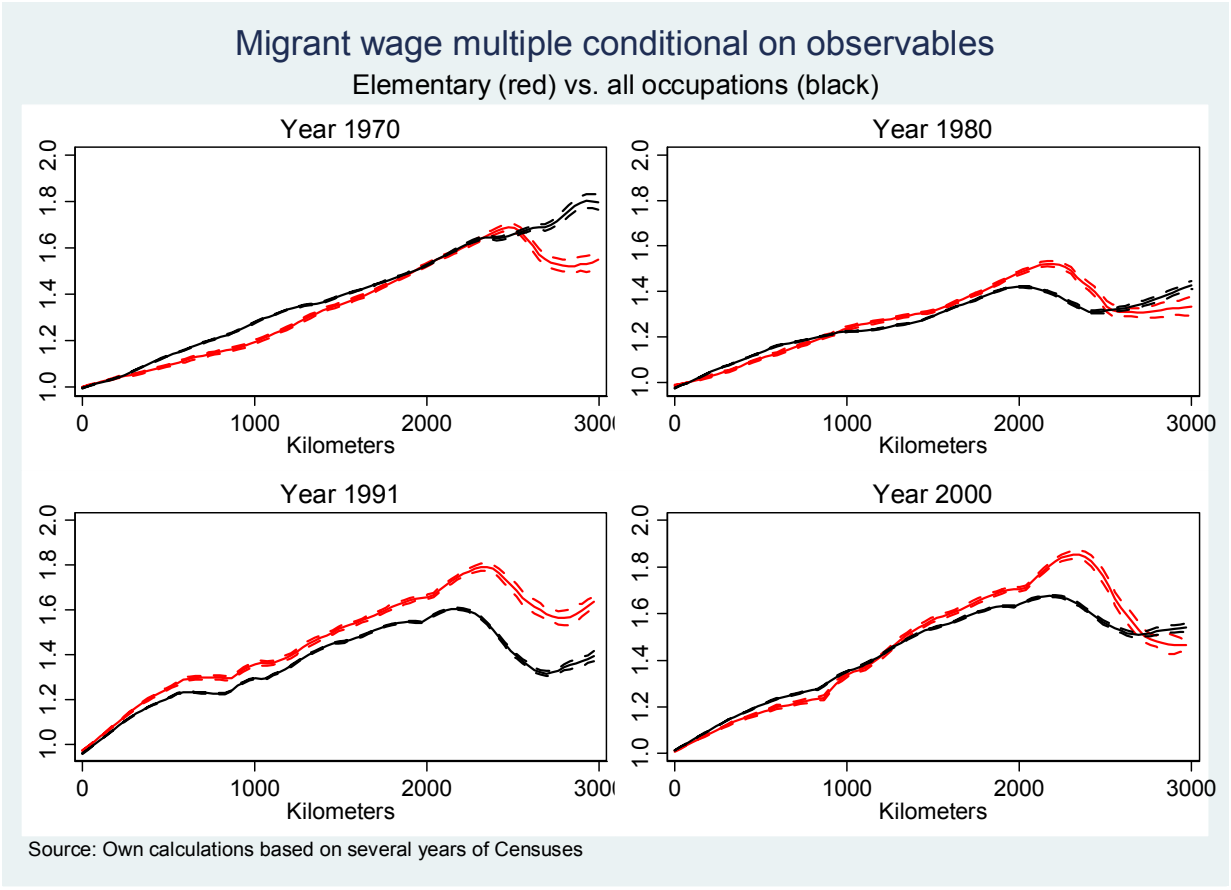
Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman (2001), "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment", *Quarterly Journal of Economics* 116 (2): 607-654.

MaxMind (2008), *MaxMind World Cities with Population*, updated September 24, 2007 (Boston: MaxMind LLC). http://www.maxmind.com/app/worldcities, accessed February 12, 2009.

Mayer, Thierry and Soledad Zignago, (2006), "Notes on CEPII's distances measures," CEPII Explantory Note (Paris: Centre d'Études Prospectives et d'Informations Internationales).

Yap, Lorene (1976), "Internal migration and economic development in Brazil", *Quarterly Journal of Economics* 90 (1): 119-137.

Figure 1: Earnings rise with distance moved from state of birth



Local linear regressions with 95% confidence bands, Epanechnikov kernel, bandwidth 250km.

Figure 2: Example of earnings landscape in the optimal stopping problem

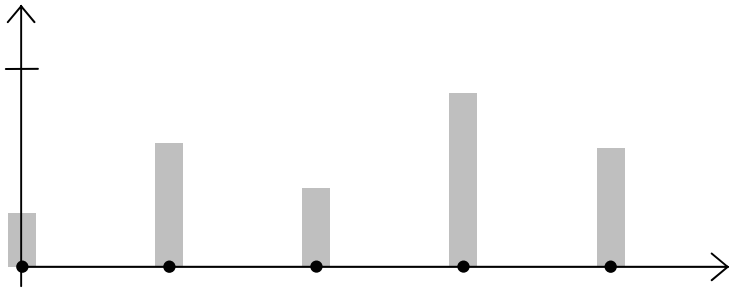
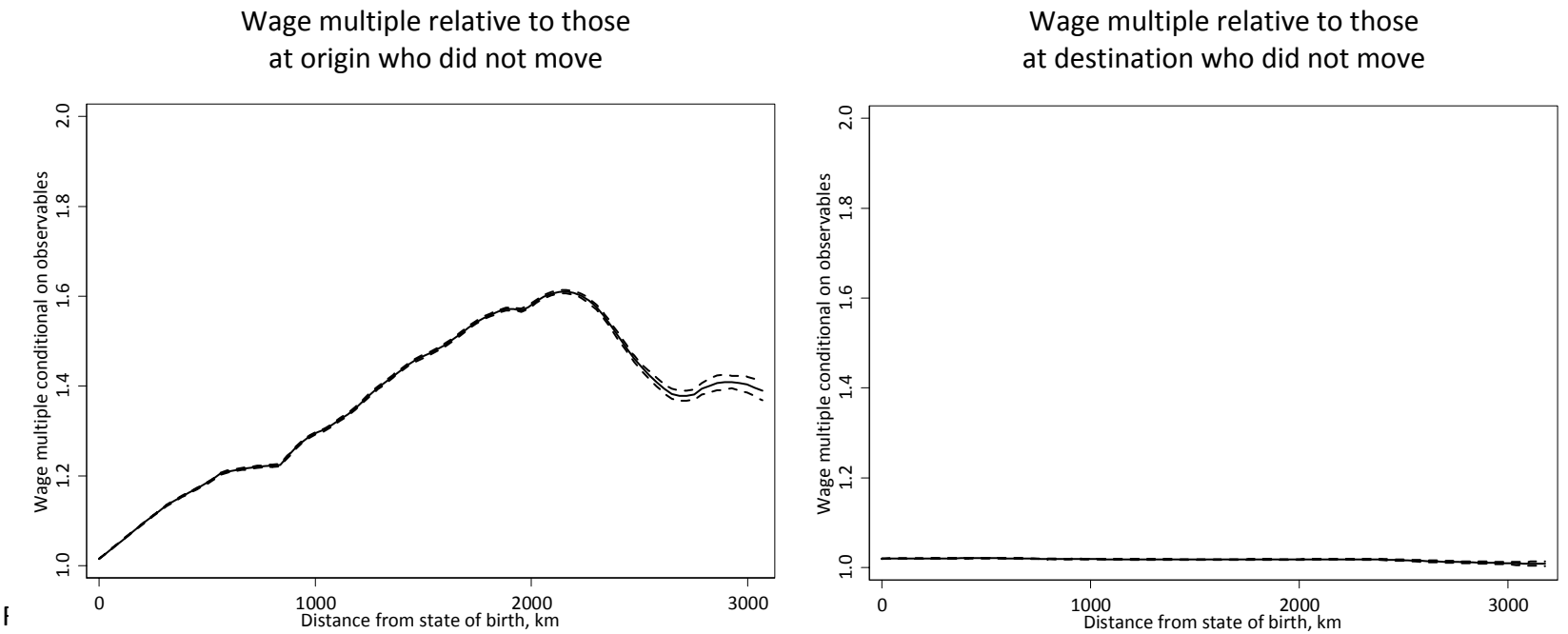


Figure 3: Little evidence of positive selection on portable unobserved wage determinants (2000)



Local linear regressions with 95% confidence bands in dashed lines, Epanechnikov kernel, bandwidth 250km. Vertical axis shows ratio of wage residual (after controlling for gender, age, and education level) for people who moved relative to the same residual for two difference reference groups. In the left-hand figure the reference group is people in the migrant's state of origin who are living in their state of birth; the right-hand figure the reference group is people in the migrant's state of destination who are living in their state of birth. All data are from the 2000 census.

Table 1: Parametric comparison of the distance-earnings relationship by occupation category

Year	1970		1980		1991		2000	
Distance from state of birth (000km)	0.167	0.168	0.143	0.136	0.177	0.169	0.143	0.142
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Elementary occupation		–0.133		–0.281		–0.174		–0.170
		(0.003)		(0.003)		(0.003)		(0.003)
Elementary occup. × Distance		–0.005		0.077		0.069		0.031
		(0.004)		(0.003)		(0.004)		(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	821,235	821,235	1,023,828	1,023,828	1,084,863	1,084,863	1,178,925	1,178,925
N pop.	25,774,341	25,774,341	40,117,295	40,117,295	54,563,214	54,563,214	68,607,161	68,607,161
R^2	0.523	0.525	0.456	0.462	0.435	0.438	0.449	0.450

Dependent variable is ln(total income). Standard errors in parentheses. Regressions weighted by census sampling weight. "Controls" are dummy variables for education levels, age groups, sex, and rural, plus constant term. "*N* pop." is the number of people in the population represented by the sample, scaled by sampling weight.

Table 2: Focus on migrants from the Northeast to São Paulo state or Rio de Janeiro state

Numerator	Denominator	Restriction	Ratio	95% conf. int.
Migrants (N = 20,217)	Non-migrants (N = 99,735)	None	1.569	(1.459, 1.687)
” (N = 2,445)	” (N = 7,970)	Elementary occup.	1.442	(1.141, 1.822)
” (N = 8,411)	” (N = 88,397)	20+ year resident	1.746	(1.542, 1.978)
” (N = 900)	” (N = 6,895)	Elementary occup. and 20+ yr. res.	1.598	(1.062, 2.405)
Migrants (N = 20,217)	Natives (N = 93,528)	None	0.874	(0.816, 0.936)
” (N = 2,445)	” (N = 5,757)	Elementary occup.	0.889	(0.709, 1.114)
” (N = 8,411)	” (N = 86,812)	20+ year resident	0.970	(0.859, 1.095)
” (N = 900)	” (N = 5,315)	Elementary occup. and 20+ yr. res.	0.973	(0.650, 1.456)
Return migrants (N = 933)	Non-migrants (N = 99,735)	None	0.766	(0.564, 1.040)

“Migrants” means born in Northeast, resident in São Paulo or Rio. “Natives” means born in São Paulo or Rio and resident in São Paulo or Rio. “Non-migrants” means born in Northeast, resident in Northeast, and did not live in São Paulo or Rio five years ago. “20+ year resident” means that the person has lived in current state of residence for 20 or more years. The “Northeast” is Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe and Bahia.

Table 3: Panel data on formal sector workers moving from the Northeast to São Paulo-Rio de Janeiro

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep. var.: ln wage in year:	2000	2000	2000	1995	2000	2000	1990
Sample observed at	NE/SP/Rio	NE/SP/Rio	SP/Rio	NE	NE/SP/Rio	SP/Rio	NE
Migrants' prior location	—	NE 1995	NE 1995	—	NE 1990	NE 1990	—
<u>No controls</u>							
In SP or Rio, 2000	0.488***	0.422***			0.315***		
	(0.008)	(0.049)			(0.033)		
Moved to SP or Rio			–0.075	0.041		–0.189***	–0.235***
by 2000			(0.049)	(0.057)		(0.032)	(0.036)
R-squared	0.055	0.005	0.000	0.000	0.005	0.001	0.002
<u>With controls</u>							
In SP or Rio, 2000	0.506***	0.500***			0.454***		
	(0.007)	(0.039)			(0.027)		
Moved to SP or Rio			0.011	0.152***		–0.038	0.026
by 2000			(0.037)	(0.044)		(0.025)	(0.030)
R-squared	0.359	0.290	0.335	0.306	0.288	0.335	0.265
Observations	65277	15949	49107	16157	16043	49107	16399
of which migrants	—	353	353	353	658	658	658

Sample held constant between "no controls" and "with controls" regressions in each column. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Controls are for gender, age, and education: a dummy for female, 8 dummies for highest level of education completed (primary incomplete, primary complete, middle school incomplete, middle school complete, high school

incomplete, high school complete, college incomplete, college complete or above, omitted base group illiterate), and 7 dummies for age ranges (15-17, 18-24, 25-29, 30-39, 40-49, 50-64, 65 and over, omitted base group 10-14).