How Important is Economic Geography for Rural Non-Agricultural Employment? Lessons from Brazil

ERIK JONASSON
Department of Economics, Lund University, Sweden
erik.jonasson@nek.lu.se

and

STEVEN HELFAND
Department of Economics, University of California, Riverside
steven.helfand@ucr.edu

Abstract: By paying particular attention to the local economic context, this paper analyzes rural non-agricultural employment and earnings in non-agricultural jobs. The empirical analysis is based on the Brazilian Demographic Census, allowing for disaggregated controls for the local economy. Education stands out as one of the key factor for shaping employment outcome and earnings potential. Failure to control for locational effects can lead to biased estimation of the importance of individual and household-specific characteristics. The empirical results show that local market size and distance to population centers have a significant impact on both non-agricultural employment prospects and earnings. The impact, however, is quantitatively larger for employment.
Key Words: Rural non-agricultural employment, economic geography, Brazil.
JEL Codes: O18, Q1, R1.

Contributed Paper prepared for presentation at the International Association of Agricultural Economists Conference, Beijing, China, August 16-22, 2009

How Important is Economic Geography for Rural Non-agricultural Employment? Lessons from Brazil

1. Introduction

Rural non-agricultural activities have received increasing attention since the 1990s. The share of rural household income that stems from non-agricultural sources ranges from 35 percent in Asia to 40 percent in Latin America and 45 percent in Sub-Saharan Africa, emphasizing that the rural economy consists of much more than just agriculture (Reardon et al., 2001). Among the roles of the rural non-agricultural (RNA) sector are its potential to absorb an underemployed labor force and thereby slow rural-to-urban migration, to increase the income of the poor, and to contribute to national economic growth (Lanjouw and Lanjouw, 2001). These roles of the RNA sector, and particularly the potential to be a pathway out of poverty, have been recognized in rural development strategies in recent decades (Echeverría, 2000; World Bank, 2008).

The extent to which rural non-agricultural employment (RNAE) is able to reduce poverty depends on rural households' access to such employment and the income prospects in these activities. The accessibility and income prospects depend jointly on supply-side effects (individual and household characteristics), demand-side effects (characteristics of the relevant labor markets), and market participation costs. Household asset endowments on their own will not generate upward income mobility if there is insufficient demand for labor, or if market participation is very costly due to physical distance to markets and underdeveloped infrastructure.

In this paper we seek to assess the importance of supply, demand, and participation cost effects on an individual's probability of engaging in RNAE and on earned income in the RNA sector. The previous empirical literature has been concerned mainly with supply-side considerations. For this reason, we devote more attention to studying the role of participation costs and demand. Even though there is a consensus that location matters for the viability of the

RNA sector, the empirical support relies on indirect locational indicators which provide limited insight into the role that remoteness from markets and urban areas actually plays (Dirven, 2004).

To reach a deeper understanding of demand-side effects and the role of market-participation costs, we use a fuller set of variables than previous studies to describe the local economic geography. By utilizing data from the Brazilian Demographic Census of year 2000, we are able to test for the role of municipal-level economic factors such as local market size and distance to population centers. We show that demand-side factors and proxies for participation costs have strong effects on the probability of being engaged in RNAE. Geographical variables have a weaker and less consistent relationship with earnings. Our conclusions about the importance of the local economic geography stand up to a number of robustness checks that seek to address endogeneity and measurement concerns.

2. Conceptual background and related literature

As a residual concept, the rural non-agricultural sector contains a wide range of activities, including everything from low-return street-vending to well-paid jobs in the formal sector. Traditionally, RNAE has been considered largely dependent on backward and forward linkages to agriculture (Mellor, 1976). A significant share of Brazilian agriculture, however, is characterized by large, highly mechanized, export-oriented production. Thus, it is unclear how strong such local linkages are relative to countries with smaller farms, lower levels of technology, and weaker links to the world market. In this spirit, Graziano da Silva and del Grossi (2001) argue that the composition of the rural non-agricultural sector in Brazil often bears little relation to regional agricultural development, and that its dynamism depends more on the size of cities in a given region. This view is supported by the map of the Brazilian Southeast (Figure 1), which depicts the share of the rural labor force whose principal occupation is in RNAE in each municipality. Non-agricultural activities are more prevalent in the proximity of capital cities and highly urbanized areas. The pattern is most pronounced in the densely populated areas surrounding São Paulo, Rio de Janeiro, and Belo Horizonte. In these areas, RNAE is above 50%, whereas in some of the remote hinterlands, the share falls below 15%.

It is widely recognized that geographical location and local economic conditions matter for the employment outcome and earnings prospects of rural households. Previous studies have utilized a range of indicators to capture the effect of local economic conditions.[1] Besides controlling for broad geographical differences, with regional dummy variables, locational variables that have been used include: distance to regional capital city and local population density; distance to nearest health center; number of population centers within one hour's commuting distance; distance to nearest market and local market size; and neighborhood average household income, local urbanization and electricity.

A number of observations on the previous literature are pertinent. First, it is not always possible to separate proxies for demand side effects from proxies for participation costs. For example, does distance to a state capital proxy for the potential size of the local market or for the transactions costs of accessing the market? Infrastructural quality such as a paved road, in contrast, clearly reduces the costs of participating in the market. Whenever possible, unambiguous proxies are preferred. Second, variables that relate to location in space--such as longitude, latitude, and altitude--can provide an attractive alternative to geographical dummies but, like dummies, they do not always have a natural interpretation. Variables that measure the distance to markets are likely to be preferable. Third, when measuring the size or distance to a relevant market, researchers should strive to be comprehensive and precise. Distances to the nearest school, health clinic, or state capital all carry some information about remoteness, but the information is fuzzy. Certainly, it should matter if the nearest urban location has five thousand or five hundred thousand people. Forth, while it is clear that participation costs should play an important role in influencing the probability of RNAE, proxies for these costs should be

[1] See, for example, the articles in the special edition of *World Development* edited by Reardon et al. (2001), and Isgut (2004).

interpreted with caution. The magnitude of a coefficient can vary considerably depending on if it is used to represent the entire group of transactions cost variables or is included as only of these variables.

More often than not, the decisions about which geographical variables to use are driven by data availability. Due to the abundance of data contained in the Demographic Census, we seek to shed light on the extent to which alternative choices that are common in the literature are adequate, or not, for capturing the effects of the local economic geography.

3. Empirical approach

The empirical study consists of two parts. First, we estimated a binomial probit in which the dependent variable indicates whether the individual was engaged in RNAE as opposed to agriculture. Second, we estimated an income model for workers engaged in RNAE. The purpose was to assess the degree to which local economic factors affect employment outcomes and earnings opportunities in the RNA sector.

The binomial model assumes that a set of exogenous variables determines an endogenous, but unobserved (latent) variable V_i. If V_i exceeds a certain threshold value, V_i*, the individual is engaged in RNAE; otherwise, she is engaged in agriculture. The latent variable V can be thought of as the rural worker's expected earnings in the rural non-agricultural sector. The threshold V* could be the shadow wage for agricultural work on the own farm or the wage rate on the agricultural labor market. Thus, the probability P_i of individual i being engaged in RNAE was modeled as:

$$P_i = PROB_i\left(RNAE_i = 1 \middle| X_{ijk}^P, H_{jk}^P, M_k^P\right) = PROB(V_i \geq V_i^*) \qquad (1)$$

in which X_{ijk}, H_{jk}, and M_k denote vectors of variables that characterize individual i, household j to which the individual belongs, and municipality k in which the household is situated. Superscripts distinguish variables in the probability (p) and income (y) models.

About 30% of rural workers reported RNAE as their principal occupation. As suggested by model (1), people engaged in non-agricultural employment are assumed to differ systematically from people engaged in agriculture. Failure to control for this selection mechanism, and the possibility that unobserved factors influence both selection and income, would provide inconsistent coefficient estimates in an OLS regression. To adjust for the effects of censoring, we applied the Heckman (1979) sample selection model. Our approach assumes that selection into RNAE is determined by a model analogous to model (1), with the slight difference that we focus only on *paid* RNAE.[2] Accounting for the selection process, we assume that income can be modeled as a log-linear function of individual, household, and locational characteristics:

$$y_i = X_{ijk}^{y}\beta_1 + H_{jk}^{y}\beta_2 + M_{k}^{y}\beta_3 + \gamma\lambda_{ijk} + \eta_{ijk} \qquad (2)$$

in which y_i is the logarithm of non-agricultural income of individual i, λ is the inverse Mills ratio, η is the error term, and β and γ are coefficients to be estimated.

4. Data and variable construction

The data come mainly from the year 2000 Demographic Census long form . The sample includes 12 percent of the population, constructed to be representative at the municipal level. There were 5,507 municipalities, with an average population of approximately 30,000 people. The analysis used the rural adult labor force as the base sample, which included around 1.7 million observations. Adults were defined as everyone 15 years or older. Anyone reporting an occupation was considered as a participant in the labor force, including unpaid workers.

With the exception of income, we consider the data in the Census to be of high quality. Thus, data quality is not a significant concern for the econometric analysis of the probability of RNAE. The income data in the Census are based on a single question about earnings that does

[2] The share of unpaid workers in the RNA sector is below 10%, compared to over 30% in agriculture.

not distinguish clearly between gross and net income for the self-employed, take proper account of seasonal earnings, or include own consumption of agricultural production. These limitations are most problematic for small farmers and the self-employed. For this reason, our econometric analysis of earnings is restricted to people employed in RNAE, and contains a robustness check limited to the sub-sample of wage earners.

Dependent and supply-side variables

The dummy variable indicating that the individual was engaged in RNAE is based on reported occupation. By RNAE we mean that a person resides in a rural domicile, yet has a principal occupation in a non-agricultural activity. This person could work at home producing handicrafts, in a rural home as a maid, in a rural area with tourism, or in an urban area in any number of possible non-agricultural occupations. The individual characteristics included in the probability model are age, gender, race, education, and migrant status. Age, age squared, and years of schooling serve as proxies for human capital. People who have moved could have a lower opportunity cost of staying on the farm. Migration could also be an indicator of unobserved ability and risk-taking.

The household characteristics included in the probability model are number of adult household members, average education in the household (excluding individual *i*), and an index of household wealth. The number of adults controls for opportunities for employment diversification: the larger the labor supply in the household, the more the opportunities to devote some labor to non-agricultural activities. Average years of schooling is a proxy for the household stock of human capital. Wealthier households are better able to finance search and participation costs associated with RNAE. Wealth can also serve as a proxy for social capital which can facilitate access to non-agricultural jobs. Finally, two variables were included to indicate whether the household lived in a rural town or urban extension as opposed to a rural exclusive area.

Earned income in July 2000, the dependent variable in the income analysis, refers to monthly wage earnings for employees and monthly gross returns for the self-employed. In addition to the variables described above, we included number of hours worked and dummies for employment status: formal-sector employee, self-employed, and three groups of employers based on the number of people they hired. We interacted the self-employment dummy with the household wealth index to control for productive assets among the self-employed.

Demand-side and transactions-costs variables

Municipal-level characteristics were included to assess the importance of local demand and market participation costs. To capture local market size, we used a distance-weighted measure of local aggregate income in the spirit of Harris's (1954) market potential analysis. It includes the income of people in municipality (k) plus income in surrounding municipalities weighted by distance. The weight is linearly declining and takes into account only municipalities (l^*) within 100 kilometers of a typical rural household:

$$Inc100d_k = \sum_{l \in l^*} Income_l \left(1 - D_{kl}/100\right) \qquad (3)$$

$Income_l$ refers to the sum of all income received by households in each municipality l ($l = 1, 2\ldots, l^*$) as reported in the Census, and D_{kl} denotes the distance from a typical rural household in the municipality of origin k to the seat of municipality l.[3] We included measures of distance to population centers to estimate the effect of being situated away from markets of different sizes. Using D_{kl}, we defined distances to the nearest municipality with 50–100, 100–250, 250–500, and more than 500 thousand people, respectively. The size of the local market and distance to markets might be considered as alternative proxies for demand. In contrast to the local demand variable *Inc100d*, which emphasizes the total size of the local market, we primarily used the distance measures to assess the importance of transactions costs associated with access to

[3] Other specifications provided very similar results.

markets. The distance variables also permit capturing non-linearity in the relationship between RNAE and distance to markets of different sizes.

Three additional municipal variables were included: the shares of rural households with access to a telephone line and to electric lighting were used to capture municipal rural infrastructure; the share of urban households in the municipality was used to capture broader infrastructural development which should lower participation costs in input and output markets.

5. Empirical results

Probability model

The results from the probit model are provided in Table 1. The marginal effects give the estimated change in the probability of employment in the RNA sector, as opposed to agriculture, given a small change in the explanatory variable or a change from 0 to 1 for the dichotomous variables. Due to the sample size, nearly all coefficients are statistically significant at least at the one percent level and the standard errors are quite small.

Model (i) includes only individual variables. When household characteristics were controlled for in model (ii), the coefficient estimates on some individual characteristics changed significantly. Omitted-variable bias is also evident when model (ii) is compared to models (iii) and (iv), which include geographical variables. Thus, failure to adequately control for the local economic geography can generate significant bias. The model specifications also provide insight into the extent to which local conditions matter for employment outcomes. Comparing the pseudo R^2 across specifications shows that the locational variables explained more of the variance in the probability of RNAE than did the household variables. When the locational variables were added to model (ii), the explained variance increased by almost 50 percent.

Model (iii) and (iv) differ in that the former includes the local aggregate income measure *Inc100d* and the latter includes the distance measures. When included together, the coefficient on local income became insignificant, suggesting that distance and local income have some overlapping effects but that distance appears to be the dominant factor. Therefore, while recognizing the importance of income, we used model (iv) as the reference model.

Model (iv) shows that human capital is positively associated with the probability of engagement in RNAE: Age had a positive and decreasing effect on the probability of non-agricultural employment, and the probability increased non-linearly with the level of educational attainment. Women had a substantially higher probability of engaging in RNAE.[4] People who moved from one municipality to another--migrants--were slightly more likely to have RNAE.

All but one of the proxies for demand-side effects and participation costs were statistically significant of the expected sign. Living in a rural area that was an urban extension, as opposed to living in the rural exclusive category, was associated with a 50 percentage point increase in the probability of RNAE. Residence in a rural town was associated with more than 20 additional percentage points. The results also suggested that distance to population centers mattered. The greater the distance to large municipalities of all four size categories, the lower was the probability of engagement in RNAE. One measure of remoteness would be to move an additional standard deviation of distance away from each of the four classes of large municipalities. The combined effect would be a reduction of 10.4 percentage points in the probability of RNAE. Municipalities of different sizes, however, had quite different impacts. Moving 100 km away from the largest class of municipalities was associated with a change in the probability of RNAE that was five times larger than the change for municipalities in the 50-100 thousand class, and three times larger than those in the 100-250 thousand class. We suspect that it is because of these non-linearities that the distance model fit the data better than the local income model. This also suggests that proxies that only measure distance to an urban area or state capital, without accounting for its size, miss an important part of this relationship.

[4] Estimation of a multinomial probability model showed that women's higher probability of involvement in RNAE was driven mainly by their participation in low-productivity RNAE. Men, and people with more education, had a higher probability of participating in high-productivity RNAE.

The one case where we found mixed evidence for participation costs relates to the proxies for rural infrastructure. The share of rural households with telephones was associated with a higher probability of RNAE, whereas the share with electricity was associated with a lower probability. The unexpected result for electricity persisted even after the exclusion of municipal outliers.

We performed a host of robustness checks on the reference model to detect potential bias in the results. First, the estimated effects of the individual and household characteristics could be influenced by unobserved local factors that we were unable to control. In order to explore this issue, the model was estimated with municipal fixed effects instead of municipal level variables. The coefficients on all non-municipal level variables were quite similar to those in model (iv). We concluded that the geographical controls in the model were adequate.

Another concern relates to the possible endogeneity of several of the regressors. Unobserved individual characteristics that have higher returns in RNAE could induce people with these characteristics to move to locations where they have a higher probability of finding RNAE. If true, the coefficients on urban extensions, rural towns, and even the distance variables, would be biased upwards because people have chosen to reside closer to where RNA jobs exist. We therefore re-estimated model (iv) without migrants and individuals who lived in urban extensions or rural towns. Model (v) shows that there was some evidence in favor of the hypothesis of endogenous sorting of the rural population, but that this did not alter any of the qualitative results, nor did it alter the fundamental conclusions about the importance of the local economic context. Education and household wealth could also be endogenous. We restricted the sample to be much more homogenous along these two dimensions, and explored whether any important conclusions were altered. No qualitative results changed, and most quantitative results remained stable.

Income model

The results from the income model are also reported in Table 1. Specification (vi) included only individual and household characteristics. The distance specification (vii) added the locational variables. Specification (viii) restricted the sample to wage laborers. In all specifications the coefficient on the Mills ratio γ was statistically significant, suggesting that correcting for sample selection was important. The sign indicated that the error terms in the selection and income equations were negatively correlated. Thus, unobserved factors correlated positively with the probability of RNAE tended to lower earnings prospects in the RNA sector.

The results suggested that local characteristics tend to affect employment outcomes and income prospects in different ways. Whereas nearly all locational variables had the expected effect on employment, the results were more mixed when the dependent variable was earnings. For example, earnings appeared to fall slightly with residence in an urban extension or rural town, and with urbanization. An important result is that the magnitude of the impacts was substantially smaller for earnings than for employment. Residence in an urban extension or rural town was associated with a 20 to 50 percentage point increase in the probability of RNAE. The impact on earnings was only in the range of three to four percent. A possible explanation for the lack of any strong positive relationship between earnings and location relates to an excess supply of labor for RNA jobs which prevents wages from rising. Thus, while non-agricultural employment prospects improve for rural residents who live closer to large urban centers, competition with the urban residents--and unemployment--implies that there is no clear earnings premium associated with residence in these locations.

We also performed multiple robustness checks on the income model. Column (viii) reports results of a specification that only included employees. This eliminated the problem of income mis-measurement of the self-employed. This narrowing of the sample did not generate any important changes in the coefficient estimates compared to the reference model.

6. Conclusions

The prospects for RNAE depend jointly on supply-side factors, demand-side factors, and the magnitude of market participation costs. The empirical analysis showed that, when holding individual and household characteristics constant, demand side factors such as local market size have a strong impact on an individual's probability of having RNAE. Proxies for transactions costs, such as distance to markets, correlate negatively with RNAE. This does not mean that supply-side effects are unimportant. Even when controlling for local factors, the effects of education, gender, and other individual characteristics are statistically and economically significant. Individual characteristics also play a key role in sorting people across low- and high-productivity RNAE. In contrast to the probability of employment, our results suggest that the local economic context is considerably less important for shaping earnings.

Implications for the poverty alleviation potential of the RNA sector are mixed. While RNAE may help to diversify income risk, with low levels of education access to RNAE is to a large extent limited to low-productivity jobs. Equally problematic is the geographical mismatch between the location of the rural poor and the location of RNAE opportunities. We conclude that RNAE should be an important component of a rural development strategy, but that it is unlikely to provide a feasible pathway out of poverty for the majority of the rural poor.

Policies that support the RNA sector should be designed with the role of location in mind. The sector's potential is conditioned by distance to larger markets, infrastructure, and the level of local aggregate demand. The benefits of geographical concentration of economic activities become increasingly important as agriculture absorbs less and less of the rural labor force, and to the extent that farm households are unable to escape poverty solely with agricultural income.

7. References

Dirven, M. (2004), “Rural Non-farm Employment and Rural Diversity in Latin America”, *CEPAL Review*, vol. 83, August.

Echeverría, R. G. (2000), “Options for Rural Poverty Reduction in Latin America and the Caribbean”, *CEPAL Review,* vol. 70, April.

Graziano da Silva, J. and M. E. del Grossi (2001), “Rural Nonfarm Employment and Incomes in Brazil: Patterns and Evolution”, *World Development*, 29(3), pp. 443–453.

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

Heckman, J. J. (1979), “Sample Selection Bias as a Specification Error”, *Econometrica*, 47(1), pp. 153–161.

Isgut, A. E. (2004), “Non-farm Income and Employment in Rural Honduras: Assessing the Role of Locational Factors”, *Journal of Development Studies*, 40(3), pp. 59–86.

Lanjouw, J. O. and P. Lanjouw (2001), “The Rural Non-farm Sector: Issues and Evidence from Developing Countries”, *Agricultural Economics*, vol. 26, pp. 1–23.

Mellor, J. W. (1976), *The New Economics of Growth. A Strategy for India and the Developing World.* Cornell University Press, Ithaca.

Reardon, T., J. Berdegue and G. Escobar (2001), “Rural Nonfarm Employment and Incomes in Latin America: Overview and Policy Implications”, *World Development*, 29(3), pp. 395–409.

World Bank (2008), *World Development Report 2008: Agriculture for Development*. World Bank, Washington D.C.

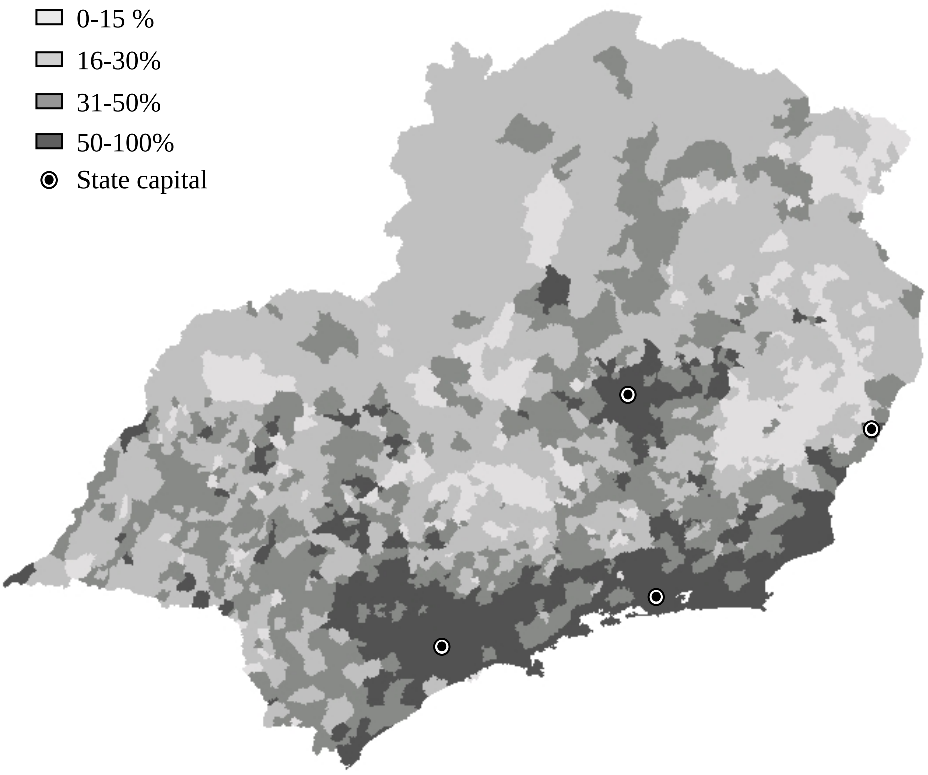

Figure 1. *Rural Non-Agricultural Employment in the Brazilian Southeast*

Table 1. *Results of the binomial probit and rural non-agricultural income models*

	Marginal effects on probability					*Marginal effect on RNA income*		
	(i) ind.	*(ii) hh*	*(iii) income*	*(iv) distance*	*(v) excl migr.*	*(vi) hh*	*(vii) distance*	*(viii) workers*
Supply-side factors								
Age	0.014	0.010	0.010	0.010	0.010	0.048	0.051	0.054
Age squared	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.001	-0.001
Male [d]	-0.139	-0.145	-0.150	-0.151	-0.138	0.476	0.457	0.446
Education, 1-4 years [d]	0.091	0.056	0.057	0.057	0.050	0.088	0.105	0.101
Education, 5-8 years [d]	0.273	0.190	0.175	0.177	0.158	0.200	0.231	0.208
Education, 9-11 years [d]	0.486	0.361	0.359	0.363	0.360	0.410	0.461	0.433
Education, >11 years [d]	0.602	0.429	0.469	0.467	0.541	0.961	1.020	1.046
Migrant [d]	0.058	0.047	0.012	0.022		0.058	0.055	0.034
Household adults		-0.010	-0.005	-0.005	-0.002			
Household education		0.013	0.010	0.010	0.009	0.034	0.034	0.035
Household wealth		0.102	0.055	0.058	0.053			
Hhd wealth × self empl.						0.331	0.339	
Demand-side factors and participation costs								
Inc100d			0.051					
Dist500				-0.073	-0.048		-0.010	-0.011
Dist250500				-0.040	-0.034		-0.013	-0.019
Dist100250				-0.011	-0.012		-0.004	-0.006
Dist50100				-0.004	(-0.001)		0.006	0.006
Urban extension [d]			0.519	0.500			-0.029	-0.015[†]
Rural town [d]			0.238	0.236			-0.038	-0.038
Share urban households			0.118	0.099	0.100		-0.040	-0.052
Share hhds with tele.			0.294	0.246	0.109		0.599	0.576
Share hhds with elec.			-0.118	-0.097	-0.072		-0.126	-0.180
Employment controls	No	No	No	No	No	Yes	Yes	Yes
Predicted RNAE	0.288	0.286	0.287	0.286	0.228	-0.200	-0.120	-0.180
Mills ratio						-0.200	-0.120	-0.180
McFadden R^2	0.112	0.130	0.190	0.198	0.126			
Observations	1,724,822	1,724,822	1,724,822	1,724,822	1,005,911	1,724,822	1,724,822	1,724,822
Uncensored observations						469,667	469,667	340,931

Note: All coefficients are statistically significant at the 1-percent level other than in the following cases: [††] denotes significance at 5-percent level; [†] at 10-percent level. Coefficients within parentheses are not significant at the 10-percent level. The dependent variable in the probit model (columns i–v) is a dummy equal to 1 if the individual is engaged in non-agricultural employment and zero if in agriculture. In the income model (column vi-viii) the dependent variable is log of earned non-agricultural income. Dummy variables are indicated by d. All specifications include control variables for race and macro region.