"Economic Determinants of Land Invasions" Dataset

F. Daniel Hidalgo*
Suresh Naidu
Simeon Nichter
Neal Richardson

August 3, 2010

*Authors' email addresses: fdhidalgo@berkeley.edu; sn2430@columbia.edu; nichter@berkeley.edu; npr@berkeley.edu.

1 Data

Tables 1 and 2 provide names and brief descriptions of variables used in the article's cross-sectional and fixed-effects specifications. More detailed information about each variable is provided below, including sources.

1.1 Land Invasions

We employ municipal-level data on land invasions provided by Brazil's Pastoral Land Commission (*Comissão Pastoral da Terra* – CPT) and Dataluta (*Banco de Dados da Luta pela Terra*). This dataset provides one of the largest samples on redistributive conflict in the literature, covering 5,299 land invasions. The variable labeled *occs* is the number of distinct land invasions per year in each Brazilian municipality between 1988 and 2004.[1] Additionally, the variable labeled $logfam$ is the logged number of families participating in each land invasion. The CPT, a church-based NGO active since 1976, collects these data from various sources.[2]

There may be concerns with the coverage of the reports: for example, they may be systematically biased towards underreporting conflict in remote areas. All of our results are qualitatively similar when we restrict our sample to only municipalities that have reported invasion activity. There may also be systematic measurement error due to journalistic bias towards overreporting the size of a land invasion. For this reason, we focus on the number of invasions, which is relatively less vulnerable to such bias than reported size of invasions. However, for comparison we also provide results for the number of families involved in land invasions. In addition, we examine a binary measure of whether conflict occurred in a municipality-year (labeled *occs_0_1*), as this measure may be less prone to measurement error.

1.2 Agricultural Income

Our primary independent variable of interest (labeled *ag_income*) is agricultural income, which is measured by crop revenue. Data on crop production are from statistics on municipal agriculture production (*Produção Agrícola Municipal*) from Brazil's national census bureau (*Instituto Brasileiro de Geografia e Estatística* – IBGE). For each muncipality-year, we take a revenue-weighted sum of the log crop yields (tons per hectare) as a measure of

[1] The CPT defines land occupations as "collective actions by landless families that, by entering rural properties, claim lands that do not fulfill the social function" [CPT 2004, p. 215, authors' translation].

[2] The CPT compiles information on land invasions from a range of data sources, including local, national and international news articles; state and federal government reports; reports from various organizations such as churches, rural unions, political parties and NGOs; reports by regional CPT offices; and citizen depositions [CPT 2004, pp. 214-26]. When data sources conflict, reports from regional CPT offices are used. In 2004, the CPT collected land invasion data from 171 sources. Repeated invasions of the same property in a given year are only counted once.

log agricultural income [Jayachandran 2006; Kruger 2007]. This calculation includes the eight most important crops in Brazil, which are cotton, rice, sugar, beans, corn, soy, wheat and coffee [Helfand and Resende 2001, p. 36].[3] We assume that municipalities take crop prices as given in domestic and international markets. Given the difficulties of ongoing data collection in rural areas, crop data are subject to numerous sources of measurement error. Macroeconomic conditions, including periods of hyperinflation, may also contribute to measurement error. Because measurement error in our data may be nonclassical, the direction of the resulting bias is ambiguous, highlighting the need for an instrument.

1.3 Rainfall

We use rainfall as a source of exogenous variation in agricultural income. Daily rainfall data from 2,605 weather stations across Brazil for the 1985-2005 period are from the Brazilian National Water Agency (*Agência Nacional de Águas* – ANA). In order to derive a municipal-level measure of rainfall, we match municipalities to the nearest weather station within 50 kilometers. Municipalities without a weather station within this matching radius were excluded. All of our results are robust to changing the acceptable matching radius to values between 20 and 100 kilometers.

Raw rainfall totals would be inappropriate to use in this context given the wide range of climatic variation in a country the size of Brazil. A 10-centimeter drop in annual rainfall may be expected to have different economic effects in arid regions of the Northeast than in the temperate South. In order to measure comparably the effect of rainfall on agricultural productivity, we transform the rain data as follows. Monthly rain totals are first standardized by station-month, then summed across the twelve months of the year. These annual totals are then standardized by station-year. Standardizing by month accounts for seasonal patterns and may more closely identify aberrant rainfall years [Mitchell 2003]. Standardization also makes rainfall measurements comparable across municipalities since agricultural production is likely to be adapted to the average level and variance of rainfall in each municipality.

The primary measure of rainfall (labeled *rain_monthly*) is the absolute value of standardized rainfall because the relationship between rainfall and income changes is non-monotonic. As shown in the article, both drought and flooding are negatively correlated with agricultural income. Formally, the primary measure is given by:

$$z = \left| \frac{x_{it} - \overline{x}_i}{s_i} \right| \tag{1}$$

where rain observations x for every rain measuring station i and year t pair are standardized by the mean $\overline{x}_i$ and standard deviation s_i of the rain data from each rain station's 21-year

[3] Helfand and Resende also include oranges, which are missing from our data.

(1985-2005) time-series.[4] For this rainfall measure, x_{it} is the annual sum of standardized monthly rainfall, given by:

$$x_{it} = \sum_{m=1}^{12} \frac{x_{imt} - \overline{x}_{im}}{s_{im}} \tag{2}$$

For producing Figure 2 in the paper, we utilize a version of this measure without first taking the absolute value (labeled *rain_monthly_noabs*).

For robustness, we also examine additional measures of rainfall. A second rainfall measure (labeled *rain_monthly_sq*) is squared rain deviation, given by:

$$z = \left(\frac{x_{it} - \overline{x}_i}{s_i} \right)^2 \tag{3}$$

As a third rainfall measure (labeled *rain_yearly*), we use the absolute value of standardized annual rainfall, without first standardizing by month. This more coarse measure is given by Equation 1, where x_{it} is now the measured rainfall for each station-year.

We also examined various alternative rainfall measures, including non-standardized measures, higher-order polynomials, and dummies for various rain thresholds. These alternative measures of rainfall are not employed as instruments because they are not as highly correlated with agricultural income.

1.4 Land Inequality and Land Tenure

Municipal-level data on the distribution of land (labeled *land_gini*) in 1992 and 1998 were calculated from INCRA's land registry by Rodolfo Hoffmann [Hoffmann 1998]. In addition to Gini coefficients of the land distribution, these data contain the number of properties within given size brackets in each municipality. The top 10% landowners' share of land area (labeled *land_top*10*pct*), as well as bottom 50% landowners' share of land area (labeled *land_bottom*50*pct*), were calculated from Hoffmann's data. Because the land Ginis were calculated based only on landowners, we adjusted them using the share of population that is landless in each municipality (labeled $frac_landless$), taken from the 1995/96 IBGE Agricultural Census. To reduce measurement error and increase the sample size, we average the 1992 and 1998 observations where both are available, and take the available observation from 1992 or 1998 otherwise. Endogeneity of inequality to land invasions is not a concern given how persistent land inequality is over time, with correlation coefficient 0.86 between Hoffmann's 1992 and 1998 observations.

Because our dependent variable is a measure of redistributive conflict, we also construct a measure of economic polarization (labeled *land_polar*) [Esteban and Ray 1994;

[4] While our agricultural income and land invasions data are limited to a shorter panel, we use this 21-year rainfall data series for standardization in order to attain a better measure of the local rainfall conditions.

Duclos, Esteban and Ray 2004]. This measure, while closely linked to the Gini coefficient, captures the degree of bimodality in the distribution of land. Esteban and Ray [1999] argue that polarization is a better predictor of conflict than the Gini coefficient.

Polarization is calculated using discrete distribution data by the formula

$$\sum_i \sum_j \pi_i^{(1+\alpha)} \pi_j |\mu_i - \mu_j|$$

where π is the fraction of landowners in group i or j, and μ is the share of land owned by the corresponding landowners, for all pairs of i and j [Esteban, Gradín and Ray 2005]. In this study, we let $\alpha = 0.5$.

Data on the fraction of land under three types of land tenure—rental (labeled $frac_rented$), ownership (labeled $frac_owned$), and sharecropping (labeled $frac_shared$)—also come from the 1995/96 IBGE Agricultural Census. Landholding is classified as rental if the tenant pays the owner a fixed amount of money in rent, or if the tenant must meet a production quota. Sharecropping, on the other hand, refers to properties in which tenants pay owners a certain share (a half, quarter, etc.) of the harvest. When a landowner engages in production, land is classified in the "ownership" category.[5] We measure these variables as the fraction of a municipality's arable land under each land contract system.

1.5 Other Variables

Data on rural population (logged variable labeled *logruralpop*), poverty rates (labeled $frac_ext_pov$), income Gini coefficients (labeled *gini* for the 1991 measure, used in table 5, and *mean_gini*, the average of the 1991 and 2000 measures, used as an interaction in table 10), rural unemployment (labeled *unemp*), and per capita income (logged variable labeled *logy*) are from the 1991 and 2000 national censuses.[6] The variable $frac_rural$ is the fraction of the total population in a municipality that is rural, based on census data. Data on agricultural workers and the quality of land (used for the share of land that remains unused, labeled *unusedland*) come from the 1995/96 Agricultural Census. IBGE administers both the national and agricultural censuses, and also provides land area data (logged variable labeled *logarea*). Annual population data (logged variable labeled *logpop*), rural population (logged variable labeled *logruralpop*), number of banks (labeled *banks*), and Human Development Indices for Education (labeled *hdi_education*) are from the Institute for Applied Economic Research (*Instituto de Pesquisa Econômica Aplicada* – IPEA), a Brazilian government agency. This source also provides municipal budget data, from

[5] These three categories, as shares of arable land, do not sum to one because the Agricultural Census has a fourth category of land, land that is "occupied," or invaded. We exclude this category due to obvious endogeneity concerns with the dependent variable, not to mention the fact that it is probably more time-variant than the other three types of land tenure.

[6] Our primary measure of agricultural income is discussed above (Section 1.2). Census data on per capita income are used as a robustness check.

which we examine logged public security expenditures (labeled *log_security_budget*) and logged social expenditures (labeled *log_social_budget*). The political competition variable (labeled *close*) is defined as the difference in vote share received by the top two mayoral candidates in the last mayoral election (mayors were elected in 1996, 2000, and 2004). Data on municipal elections come from the Supreme Electoral Court (*Tribunal Superior Eleitoral* – TSE). All analysis is restricted to municipalities with more than 10 percent rural population (averaged for the 1991 and 2000 census years); results are robust to different thresholds.

Table 1: Variable Names and Descriptions for Cross-Sectional Specifications

Variable Name	Description
occs	Land Invasions (Count), 1988-2004
occs_0_1	Land Invasions (Dichotomous), 1988-2004
$logfam$	Log (Families), 1988-2004
land_gini	Land Gini
logy	Log (GDP per capita), 1991
unusedland	Unused Arable Land (Proportion)
gini	Income Gini, 1991
$frac_ext_pov$	Extreme Poverty (Percent), 1991
logpop	Log (Population), 1991
logruralpop	Log (Rural Population), 1991
logarea	Log (Land Area)
hdi_education	Education HDI, 1991

References

Duclos, Jean-Yves, Joan Esteban and Debraj Ray. 2004. "Polarization: Concepts, Measurement, Estimation." *Econometrica* 72(6):1737–1772.

Esteban, Joan, Carlos Gradín and Debraj Ray. 2005. "Extensions of a Measure of Polarization, with an Application to the Income Distribution of Five OECD Countries." Working Paper.

Esteban, Joan and Debraj Ray. 1994. "On the Measurement of Polarization." *Econometrica* 62(4):819–851.

Esteban, Joan and Debraj Ray. 1999. "Conflict and Distribution." *Journal of Economic Theory* 87:379–415.

Helfand, Steven and Gervásio Castro de Resende. 2001. "The Impact of Sector-Specific and Economy-Wide Policy Reforms on Agriculture: The Case of Brazil, 1980-98." Working Paper.

Hoffmann, Rodolfo. 1998. A Estrutura Fundiária no Brasil de Acordo com o Cadastro do INCRA: 1967 a 1998. Technical report UNICAMP Working Paper.

Jayachandran, Seema. 2006. "Selling Labor Low: Wage Responses to Productivity Shocks in Developing Countries." *Journal of Political Economy* 114(3).

Kruger, Diana. 2007. "Coffee Production Effects on Child Labor and Schooling in Rural Brazil." *Journal of Development Economics* 82:448–63.

Mitchell, Todd. 2003. "Northeast Brazil Rainfall Anomaly Index, 1849-2002." Joint Institute for the Study of the Atmosphere and Ocean, http://jisao.washington.edu/data_sets/brazil/.

Table 2: Variable Names and Descriptions for the Fixed-Effects Specifications

Variable Name	Description
code	IBGE Municipality Code
year	Year
occs_0_1	Land Invasions, Dichotomous
occs	Land Invasions, Count
$logfam$	Log (Families)
sum_occs	Land Invasions, Total 1988-2004
rain_monthly	Rain Deviation (Monthly)
rain_monthly_noabs	Standardized Rainfall (Monthly)
rain_monthly_sq	Rain Deviation (Squared)
rain_yearly	Rain Deviation (Annual)
meanrain	Average Rainfall (mm/day)
sdrain	SD of Average Rainfall
$coefvarrain$	Coefficient of Variation (CV) of Rainfall
rain_shock	Coded 1 if $rain_monthly > 2$, 0 otherwise
$future_rain_monthly$	Rain Deviation (Monthly), $t + 1$
ag_income	Agricultural Income
logpop	Log (Population)
land_gini	Land Gini
land_polar	Polarization
*land_top*10*pct*	Top 10% Landowners' Share
*land_bottom*50*pct*	Bottom 50% Landowners' Share
$frac_landless$	Landless Population (Proportion)
$frac_rented$	Land with Fixed-Rent Tenure (Proportion)
$frac_owned$	Land with Ownership Tenure (Proportion)
$frac_shared$	Land with Sharecropping Tenure (Proportion)
gini	Income Gini, 1991
mean_gini	Income Gini, averaged 1991 and 2000
close	Political Competition
unusedland	Unused Arable Land (Proportion)
banks	Banks
log_security_budget	Log (Security Budget)
log_social_budget	Log (Social Spending)
logy	Log (GDP per capita)
unemp	Rural Unemployment
logruralpop	Log (Rural Population)
hdi_education	Education HDI