Petterson Molina Vale, Marcelo C. C. Stabile

GIS without GPS: new opportunities in technology and survey research to link people and place

Article (Accepted version) (Refereed)

GIS without GPS: new opportunities in technology and survey design for linking people and place

Petterson Molina Vale (p.m.vale@lse.ac.uk)

London School of Economics (LSE)

Marcelo C. C. Stabile (marcelo.stabile@ipam.org.br)

Amazon Environmental Research Institute (IPAM)

Abstract

This paper reports on innovative ways to relate survey data to GIS maps, thereby making the connection of people and place more accessible for the research community. Based on data from rural areas in the Brazilian Amazon, we describe a successful effort to sample households while linking farm-level data to property boundaries, these boundaries generated from subjects' interpretations of satellite images on a computer screen. The sampling framework is based on legislation requiring farmers to report to a government agency in a four-week period and the farmers' input allows for a more efficient means of identifying property boundaries as compared to GPS. We show that the resulting sampling is statistically representative. We discuss the potential of this association of institutional design and low-cost methods of data collection to allow for more cost-effective generation of spatial data and conduction of geospatial analysis.

Keywords: GIS, GPS, survey, probabilistic sampling, research design, satellite imagery, Brazil, Amazon.

Resumo (Portuguese)

Este artigo apresenta uma forma inovadora de utilização de características de desenho institucional e tecnologias digitais para lidar com conhecidos desafios para o georeferenciamento de dados de campo. O estudo se baseia em uma experiência de coleta de dados em áreas rurais na Amazônia. Trata-se da primeira *survey* (até onde sabemos) com amostragem aproximadamente aleatória a relacionar dados de propriedades rurais a mapas *offline* dessas propriedades usando informações visuais fornecidas por pecuaristas aos quais foram apresentadas imagens de satélite em uma tela de computador. A amostragem se baseia em uma legislação que obriga todos os pecuaristas a comparacerem a uma agência governamental em um período de quatro semanas. Mostra-se que o resultado é estatisticamente representativo. Discute-se o potencial dessa associação de desenho institucional a métodos de baixo custo de coleta de dados para gerar dados espacias e conduzir análises geoespaciais de forma mais custo-efetiva.

Palavras-chave: GIS, GPS, survey, amostragem aleatória, desenho de pesquisa, Amazônia.

Where mules were concerned, I had no choice: within a radius of thirty miles around Cuiaba there were not more than fifteen for sale.

— (Claude Lévi-Strauss, Tristes Tropiques, 1961, p. 253)

1. Introduction

What challenges are involved in collecting primary data in a harsh environment such as the Amazon? When Claude Lévi-Strauss collected the first systematic ethnographic evidence on the Nambikwara tribe, back in 1938, he assembled no less than "*fifteen men, fifteen mules, and thirty oxen*" (ibidem). Difficulties were close to insurmountable. As soon as the journey started, his transportation animals "*began to suffer great pain from the fact that saddles bit into their skins* (…) *These skeletal, festering beasts were my first casualties*" (ibidem). The trip was eventually cut down to six months due to a generalized lack of resources.

Fieldwork is different today. The toolkit available to researchers has evolved to include 4x4 vehicles, GPS (Global Positioning System) devices, portable computers, and myriad digital technologies that make many things easier. In the era of internet and technology-based data collection, researchers are spared the operational problems faced by Lévi-Strauss back in the 1930s, even as novel and equally challenging difficulties arise.

Can survey data be made statistically representative and spatially explicit at a cost accessible to a wider body of researchers (especially those in developing countries)? The use of spatial data is increasingly recognized as imperative for the advancement of many of today's most relevant research problems. Yet despite the advances in technology and methods, linking people (socio-economic data) and place (spatial data) still requires considerable human and financial resources, restricting the use of a much demanded analytical kit to a small group of scholars. We report on an experience in which well-known challenges to GIS survey data

collection were overcome by making use of new technologies and opportunities in institutional design[1].

This paper is part of a research project that tries to answer the following question: "Does rising land productivity of cattle increase deforestation? If so, how?". Based on a comparative case-study approach, the research assesses the micro-level foundations of the proposition that land use intensification leads to migration to and deforestation in forest margins. It employs an innovative procedure to collect georeferenced survey data that is used to provide an initial test of the proposed model of intensification and migration. The research further uses secondary data and spatial econometrics to look for evidence of a positive relation between cattle intensification and deforestation ('rebound effect'). The results suggest a substantial land-sparing effect of intensification on deforestation in frontier municipalities. This paper focuses on the methodological innovations that underpin the land use study. Readers interested in the empirical results are referred to Vale (2015).

The information a surveyor can collect on the ground is a grain of sand when compared to the layers of spatially explicit data freely available in GIS data repositories. The most common spatially referenced data that can be readily linked to a grid map are satellite images with information on land cover, water availability, carbon content and others, but there are countless other raster and vector data layers such as household censuses, road networks, property boundaries, etc. This paper explores applications to land use problems, especially the interplay between technological trajectories, land use decisions, and the environment. However, spatially explicit household and farm-level data can be useful for applications ranging from marketing to psychology to ecology, to cite just a few.

[1] We use this term to refer to institutional processes (particular features of the functioning of institutions and organizations, such as regulations, laws, and informal procedures) that can be instrumented to improve research design.

Our method links farm-level survey data to spatial information about properties (in the form of vector polygons) using farmers' visual input on a computer screen. We use a form of participant GIS that includes the direct interaction between informants and the computer. This is made possible under probabilistic sampling due to the knowledge of institutional processes that have enabled the drawing of respondents according to arrivals at a government agency. The key insight was that since very close to the full population of farmers appears at a handful of government agencies in a relatively small time window, a representative sample could in principle be obtained by sampling from those arrivals (this is discussed in section 3).

A central feature of participatory approaches is that they make the best possible use of the detailed knowledge that farmers have of their own farms. This has important consequences for data quality. By drawing maps based on the interaction between farmers and satellite images we let the farmer provide detailed input on the shape and location of the boundaries. Since many plots have gone through successive waves of cuts and redraws, this approach minimizes errors by capturing fine-grained detail on the boundaries. The same result could, in principle, be achieved using a GPS device, but it would take the farmer to walk the boundaries of the plot with the surveyor.

Participatory GIS is a way of addressing issues of empowerment and legitimacy. In the policy arena, it is often used in land-use planning to allow residents to take part in the decision process (McCall and Minang, 2005). In academia, participatory GIS has been used to assess qualitative features of complex geographical phenomena such as the value of ecosystem services (Balram and Dragicevic, 2005; Raymond et al., 2009; Sherrouse et al., 2011) or the different uses of and perceptions about particular features of land such as rivers, trails, and forests (Bernard et al., 2011).

In our case, the use of a participatory technique addresses a more prosaic issue: that official grid maps of rural properties are often outdated and imprecise. We allow farmers to update

the cadastral maps by drawing the contours of their plots directly on a computer screen. This of course raises issues of representation and legitimacy such as discussed by Elwood (2006) and Dunn (2007). For example, the precise shape and locations of boundaries can be a matter of dispute both within the community and with official sources. Whose position are to be favoured? While such questions are central to the field of human geography, here we take a positive approach to science and focus on the operational challenge of linking space and people. The participatory approach is thus not aimed at gaining legitimacy but more simply at providing an updated and as realistic as possible representation of the plot's boundaries.

The full cost of our survey was in the order of US$ 19,500 (all figures in 2013 dollars) or, approximately, US$ 50 per surveyed plot. For comparison, the collection of similar data in Thailand using a much more complex approach cost US$ 44 per surveyed plot only for laminating and digitalizing the maps (Rindfuss et al., 2004). Our survey was fully implemented by twelve paid enumerators working for one month plus one researcher working for six months: approximately 1.4 person-days per plot surveyed. This compares to 3 person-days per plot surveyed in the cited study. In the other end of the cost spectrum, Bernard et al. (2011) used participatory GIS to collect spatially explicit survey data for 415 families in a sustainable use reserve in the Brazilian Amazon for a cost of approximately US$ 25 per family surveyed. This reflects an approach to collecting spatial data that is comparable to ours, but with the crucial difference that they did not aim for a statistically representative sample, which significantly reduces the costs.

The figures above are in no way systematic, yet they provide some grounding to the assertion that the method presented here allows for more cost-effective generation of spatially explicit survey data. We put our method to the test by studying eight municipalities in Northern Brazil (State of Rondônia). We collected data from cattle-oriented rural properties on topics including land use patterns, adoption of technologies, migration history, quality of pastures,

availability of capital, land values, and attitudes towards pasture management and environmental preservation[2]. With the assistance of 12 paid enumerators, we interviewed a total of 384 farmers in April / May 2013, generating spatial information (a property grid containing the boundaries of properties) for 95.5% of the surveyed farms:

[Table 1 about here]

To the best of our knowledge, this is the first scientific survey to employ a probabilistic sampling framework to generate an offline property grid, without in-situ visits, using farmers' visual input directly on a computer screen. We start the next section by surveying the literature for the standard data collection procedures and their limitations. We then present and discuss our approach, including how we obtained GIS information without using a GPS and how we interviewed farmers (section 3). We subsequently discuss the results of the survey by assessing the degree to which the data collected approximates a random sample (section 4). We initially assess the level of non-response bias, then use spatial and non-spatial statistical approaches to check for randomness. For this we take advantage of a dataset of the full population of cattle ranchers in our study area. We conclude by arguing that the new approach to collecting spatially explicit data in a more time and money-efficient manner was successful in many ways, is replicable in different contexts, but cannot be used in multiple waves of a longitudinal survey.

2. An overview of the existing literature

We start by discussing challenges to data collection in land use studies. Land use studies can follow two broad methodological approaches. The first is to look at landscapes from an

[2] The full questionnaire and interview protocols are available in the Appendix to Vale (2015). Examples of questions are: "In the coming 3 years, would you like to sell out (or rent out) your land here and buy (or rent) somewhere else? Yes / no / maybe"; "If you had enough money to double the size of your herd in the next 3 years, in what pastures would you put the extra cattle? Why? I would: clear new areas / buy / rent pastures / use the existing pastureland / other"; "Soil management technologies: Liming: since when (year)? Hectares?; Fertilizing: since when (year)? Hectares?".

aggregative perspective, using census tracts, municipalities or states as spatial units. This allows the researcher to use secondary data and avoids the problem of fieldwork altogether, but many research questions require data at a smaller scale. In the second approach, scholars collect socioeconomic information at the household level by surveying a sample of the population of interest, typically using on-site interviews and GPS readings to geoference the data.

A probabilistic sampling protocol that takes space into account presumes knowledge of the geographical distribution of the population; therefore, the lack of a publicly available spatial database of properties is a crucial challenge for sampling. The displacement of people in rural areas is normally such that in few occasions is an updated database of farmers and their locations available. If sites are remote and difficult to access, the problem gets worse. A second challenge, specific to the Brazilian Amazon, is that due to stronger enforcement of the environmental legislation farmers have grown suspicious of strangers. This may lead to biased answers if trust cannot be created. Finally, infrastructure limitations pose difficulties to data synchronization during fieldwork and to accessing online materials.

Land use scholars have responded to the issues above in four ways.

I) One option is to rely on an existing, often outdated, cadastral base. For example, Caviglia-Harris and Harris (2008) used Brazilian government's official maps from INCRA (the Federal government's agency for colonization and agrarian reform) as property grid both for sampling and spatial analysis. Yet without correction for changing boundaries, lot aggregation / disaggregation and farmers' relocation, this is far from ideal. The best solution for an outdated property grid is to apply some correction procedure to account for unrecorded changes. This will normally consist of visits to the sampled farms and recording of GPS readings of the property's boundaries, such as in Lorena and Lambin (2009) and McCracken et al. (1999). Visits do invariably require a lot of financial and human resources though; each team of researchers

is normally able to make only between one and two successful in-situ interviews per day (Moran, Siqueira and Brondizio, 2004).

II) Alternatively, a cadastral base can be created from scratch. This is of course ideal, but implies a large operational effort and even more funds than the previous strategy, as it amounts to a mini-census of the population. A derivation of this strategy is to do the sampling first, based on an existing non-spatial cadastral base, then to draw the maps. For example, Turner II and Geoghegan (2003) describe a survey where the property grid was drawn by walking each property with the household members and taking GPS readings of the borders of each subpart of the plot. A less extreme example is that of Futemma and Brondizio (2003), who also used GPS readings to draw the property grid.

III) The third route is to estimate a property grid based on previous knowledge of the spatial pattern of settlement. Here the property maps are derived from image classification (supervised or unsupervised) and algorithms that estimate boundaries based on satellite imagery, then validated with farmers who are shown the resulting grid either on printed or digital form (Walsh and Welsh, 2003). Walsh et al. (2004) used algorithm-generated property grids for sampling, validated boundaries with farmers by visiting their plots and showing them paper sketch-maps of their farms, and took GPS readings for final validation. They used the farmers' visual input on the printed sketch, but with limited room for interaction as they could not zoom in/out or show/hide different data layers as can be done in a computer.

IV) The fourth group of sampling strategies is that of selecting subjects directly at some physical location where they normally gather, such as a government agency or a church. For example, Bell (2011) interviews farmers who show up at a government agency to request free technical assistance or to register for government programmes. This is a cheap and straightforward approach, yet it carries one major drawback. Provided that the subjects who show up at the chosen location self-select to do so, the resulting sampling is fundamentally

selection-biased as the probability of being sampled depends on characteristics of the subjects.

There is no obviously superior approach, and the choice very much depends on the individual requirements of each survey and the availability of resources. For the work presented here, the fact that the first three solutions above rely on GPS readings is a major constraint. The advantage of spatial information generated by a GPS is precision, but there are downsides. First, visiting each single plot is highly time and money-consuming (distances are not trivial), especially with the elevated incidence of absentee farmers that worsens as rural areas become more connected with towns. Second, collecting GPS readings implies a high level of commitment by the farmer, which is difficult to obtain, especially in situations where there is suspicion towards strangers as already discussed. Third, when plots have an irregular shape, GPS coordinates can be an inadequate solution.

Given the drawbacks of solutions I to III, we develop a data collection method that does not rely on field visits or GPS readings and that takes advantage of an institutional design opportunity to employ strategy IV while minimizing sampling bias.

3. Methodological approach - new roads to linking space and people

We now present and discuss the method employed. We start by briefly describing the background research project, the sampling strategy, and the case study area. We then go through the details of the data collection strategy, discussing the options considered, why, and how the survey was implemented.

Aim of the larger research project, population and sampling

Deforestation and cattle are the key variables in the broader research of which this methodological paper is part. The aim of the project is to uncover land use dynamics in areas of agricultural frontier and how those relate to an intensification process that takes place mostly in consolidated areas, where migration and deforestation have stabilized. Here we succinctly describe the key features of the research project in order to contextualize the paper.

The underling theoretical approach distinguishes four categories of municipalities: pre-frontier, frontier, transition and consolidated. These are not purely temporal nor purely spatial abstractions: they are spatiotemporal units. In a stylized world, a given municipality is expected to be a pre-frontier until the moment when a settlement process starts, turning the area into an agricultural frontier. The open access situation attracts flows of migrants in search of cheap and fertile lands, but at some point the process is checked by both economic and biophysical factors. A crisis then arises that forces farmers to choose between land use intensification and out-migration: this is the transition phase. Eventually, the area evolves to a consolidated situation where the private property regime takes over and land use intensification is more prevalent.

To fully capture within-variation in each of the categories, we adopt stratified sampling with four groups of municipalities: pre-frontier, frontier, transition and consolidated, defined on the basis of deforestation rates and extent[3]. This permits the testing of both within and between-group implications of the theory. The data collected is summarized in table 2 below:

[Table 2 about here]

[3] For the demarcation of the categories we adapt the method by Rodrigues et al. (2009). First, we calculate deforestation outside of protected areas. Second, we estimate a k-means clustering model with 4 clusters and two variables: municipality cumulative deforestation in 2000 and deforestation growth from 2000 to 2010. K-means is a method of clustering that partitions points into k pre-defined groups, randomly assigns k centroids to the data and calculates the distance from each point to the nearest centroid. The algorithm keeps switching the centroids until the sum of squares from points to the centres of the groups is minimized.

Pre-frontier areas are where no colonization process took place, and social dynamics are only weakly influenced by the induced settlement logic observed elsewhere. In our study area there is one such location in the municipality of Guajará-Mirim. Because this location has only 0.5% of the State's cattle farms, it was oversampled to assure within-case variation.

For the other strata, we started by picking two municipalities (Ouro Preto and Machadinho do Oeste) that can be used for data validation by relating the results to longitudinal data available from other studies. To minimize transportation costs, we excluded 6 municipalities with three or more IDARON agencies (where the interviews were conducted). We created four geographical clusters along and across the main road of the State, and sampled 5 municipalities from those. The resulting sample includes 8 out of 52 municipalities (1 pre-frontier, 3 frontier, 2 transition and 2 consolidated), accounting for 19.2% of the total population (84,594 cattle farmers). The final sample includes 0.45% of the studied population, and can be said to be roughly representative of the State and of each one of the four categories (as discussed below). Figure 1 shows the sampled municipalities and farms.

[Figure 1 about here]

Methodological innovations and procedures

We took advantage of an institutional opportunity provided by the law. Cattle farmers are legally bound to go to a government agency (IDARON) every year to report that they have vaccinated their herds. They may come anytime in a one-month period, and there is strong enforcement in place, so it is one of the few pieces of State legislation with almost universal compliance. The share of farmers who failed to comply with the reporting obligation in 2010 was negligible—about 1.2% of the overall population (according to IDARON). Moreover, non-compliers tend to be subsistence farmers with very small herds who trade little or no cattle—and are thus not affected by trading restrictions. Even if the sample may be slightly biased

against subsistence farmers, these account for a small share of deforestation so their absence should not significantly impact results.

A statistically representative sample of farmers could thus be obtained by sitting at the IDARON agency and interviewing subjects as they arrived. This way, we avoided the problem of absentee farmers that makes sampling procedures tricky and that can be severe with in-situ data collection.

We (and a team of surveyors) obtained information both about the property and the household. We approached farmers who came to IDARON to report their herd's vaccination, excluding those who came for other reasons. Every IDARON report form is linked to one 'property', defined as a closed boundary spatial unit that is managed (not necessarily owned) by the respondent. As farmers were prepared to answer the agency's questions, we took advantage of that cognitive link and structured the survey around the so defined 'properties'. Approximately 22% of respondents were in charge of multiple properties, so we chose to record the size of all properties but only conduct the full survey on the oldest one.

To reduce selection bias, the sampling protocol consisted of interviewing the first farmer who stepped in immediately after the preceding interview was over. Probability of arrival of potential subjects depends on farm distance and farmers' characteristics such as management skills (planning the reporting ahead to avoid congestion). Arrivals not being completely random, there was a risk of bias if the probability of a farmer being surveyed is different from his probability of arrival. Thus, interview dates were allocated according to prior estimates of weekly and weekday arrival frequencies, and we made sure to use all available hours of the day. Farmers not willing to be surveyed (41.1% of total) where asked three auxiliary questions that we use to check for non-response bias (more on this below).

Another important advantage of this approach is that being supported by IDARON helped to build trust. Relying on non-local surveyors and asking for a high level of commitment from

farmers in terms of time, physically showing their farm, or answering detailed questions about sensitive issues, tends to compromise answer quality. One way of building trust is to link the research team to people or institutions that farmers recognize as trustful. Because IDARON is seen as an institution that supports the cattle sector and that is not related to the environmental agency, it is recognized by farmers as trustful. To reinforce the link between the survey and IDARON we gave farmers a particular type of non-monetary incentive.

Studies on the effect of monetary and non-monetary incentives on survey response rates have consistently shown a positive effect (Mizes, Fleece and Roos, 1984; Davern et al., 2003). When the incentives are in the form of a lottery, however, it is not clear that response rates increase (Singer, 2002; Porter and Whitcomb, 2003). Studies have also pointed out that incentives may have a positive effect on response quality, even if the evidence is mixed (Hansen, 1980; Willimack et al., 1995). In all cases, incentives do not seem to cause response bias, especially those of the non-monetary type. Given the many pros and few cons, and given the problem of farmers being highly suspicious of strangers, we adopted two strategies to motivate individuals to take part in the survey.

First, we provided refreshments and cookies during the interview, which should have had an effect on response completeness. Second, farmers were given the possibility of taking part in a raffle in which two vaccination guns were drawn among the respondents. Vaccination guns are essential tools for cattle ranchers, but there is also an important symbolism attached to them. They are at once seen as emblems of manhood and as a symbol of responsible cattle ranching. More importantly, vaccination guns can be said to convey a (subtle) message that goes counter to environmental and conservationist narratives.

The drawing of vaccination guns is thus expected to have increased response quality by creating an implicit resonance between the survey, the sanitation agency (IDARON), and the farmers, and away from the environmental agency. Also, being relatively expensive items, the

guns may have had a positive effect on the response rate. This, however, could also have biased the sample in favour of poorer farmers, but that would have counterbalanced the bias against subsistence farmers mentioned before.

To generate the property grid, we drew the maps right after the interview at the government agency. We loaded the following data on in individual projects (using QGis Lisboa[4]): 2,5m pan-sharpened 2008 SPOT satellite images, a detailed vector of the road network, an outdated cadastral map provided by the government (INCRA), and maps of protected areas (indigenous and conservation units). We would start by explaining that the research included spatial information of properties and how the data would be used, then would ask for the respondent's consent to provide that information. If approved, we would ask for the plot number, as this could easily direct us to the right location through the cadastral map, and otherwise use the address and visual information to locate the farm. The zoom was then used to show the pre-selected farm to the farmer, ensuring the location was correct. Next, the farmer's input was used either to draw out or correct the property boundaries. Last, the final shape was validated with the farmer (figure 2).

[Figure 2 about here]

Two types of sensitive information were collected. One is data on how much cattle ranchers have cleared, how much land they own, how much cattle they own, etc. On their own, these data would not have raised privacy concerns as we did not identify the farmers, so there could be no way to link them back to the data once the interview was over. The second type of sensitive information are the georeferenced property maps. This is more problematic as knowing the location of the farm can enable us to get back to the respondent's plot. For this reason, the GIS part of the dataset will remain confidential and only for the use of our team,

[4] http://www.qgis.org/en/site/

which is strictly for academic aims. The parts of the dataset that instead cannot be linked back to respondents may be made available to other researchers upon request.

Our paid enumerators were all locals. Undergraduate students were hired whenever possible, and otherwise young people with knowledge of the cattle business. One day's theoretical and one day's practical training were provided. In the first day, we went through the questionnaire, the sampling procedure, the ethics protocol, and the GIS software. In the second day each surveyor interviewed real subjects under supervision and received feedback.

We also considered whether to use paper or computers to record the data. Paper might have made subjects feel more at ease, but that had to be weighed against the difficulty of synchronizing data and the consequent delay in verification and correction. Efficient synchronization would allow for the correction of errors at early stages of the survey, so we opted for netbooks rather than paper. To store questionnaires we used a survey software ('survey gizmo'[5]) that allows for offline recording of data and subsequent synchronization with an online server. Questionnaires from all surveyors were uploaded to the online servers each evening, allowing for a daily routine of verification and correction. This was instrumental in spotting common mistakes and discussing them with the team.

To match polygons (saved on a shapefile) and survey IDs (saved on a spreadsheet) we used a combination of identifiers (surveyor name, municipality, date, time, and lot size) rather than a dedicated code, thus reducing the risk of typing mistakes compromising the matching. We still could not match nine polygons to any questionnaire, as well as eight polygons whose questionnaires were lost due to synchronization problems.

[5] http://www.surveygizmo.com/

The above are pragmatic solutions to the challenges presented by the standard methodological procedures. The solutions we adopted were specifically leaned towards reducing operational costs and making collection of spatial survey data more accessible. But cost reduction often entails losses in at least some relevant aspects. In the next section we evaluate the results of the survey, assessing data quality and presenting the challenges and potential trade-offs that we faced.

4. Assessing the success of the new method

Two common concerns in the evaluation of surveys are representativeness and non-response bias. Non-probabilistic sampling can invalidate generalizations and in some cases make results meaningless, while a high incidence of non-response may lead to important biases that can likewise have a negative impact on analytical outcomes. More importantly, lack of independence between drawn observations may invalidate inference, so it is important to test for random sampling. Here we provide a detailed analysis of the data collected with respect to these two aspects, with the aim of assessing the quality of the sampling.

Non-reponse

Non-response bias is an increasingly important topic in survey design as drop-out rates have risen substantially in the last decades (Särndal and Lündstrom, 2005). This survey's response rate was of 58.5%, falling well within the normal range for surveys where individuals are interviewed in person. To estimate the degree of bias that may be associated with drop-outs in the survey, three pieces of auxiliary information from non-respondents were collected: pasture area, cattle herd size and time in the plot.

We follow Särndal and Lündstrom's (2005) procedure to estimate the impact of non-response. We compute a binary variable for the response / non-response outcome and model it as a dependent variable in a logistic regression model. If the auxiliary vector is a good predictor of

the probability of response, then there is evidence that some degree of bias stemmed from the absence of non-respondents. If the auxiliary vector is instead a poor predictor of the response outcome, then there is no evidence of bias. Table 3 shows the results of fitting a logistic model of the binary response / non-response outcome on the auxiliary vector.

[Table 3 about here]

The results show no correlation between the auxiliary variables and the probability of response. In column 1 the auxiliary variables show no individual statistical significance to predict the probability of a farmer responding to the survey. In column 2 municipality dummies are added as independent variables, and the coefficient for the variable time in the plot becomes significant at the 10% level, but a Likelihood-Ratio test for all auxiliary variables rejects the hypothesis that their coefficients are jointly different from zero. In columns 3 and 4 dummies for surveyors as well as interactions between surveyors and municipalities are added, and the auxiliary variables remain individually as well as jointly non-significant. Therefore, even if the non-response rate was high (around 42%), non-respondents were not systematically similar to each other, so non-response is unlikely to have biased the sampling.

Sampling

We use IDARON data on the population of cattle ranchers and farm sizes in the State of Rondônia from the year 2010 to test for selection bias. A test for spatial randomness using the survey's property grid as well as INCRA's property grid is also conducted. We start by confronting the cattle herd population data with the sample data, then do the same for farm size and location. We calculate double-sided t-tests for the equality between the sampled cattle herd means and the true population means from three years earlier. This is done for both the full sample and each of the four strata. We further generate random samples from the population with equal size to the survey's samples and calculate correlations between the two. The results are in table 4.

[Table 4 about here]

The first thing to note in table 4 is that correlations between the sample and a random sample from the population are high, indicating that the sample's distribution resembles that of a random sample. Looking at averages, it can be seen for the State as a whole (upper line of the table) that the sample's mean herd size is 12% higher than the population's mean, but a t-test fails to reject the null hypothesis of sample randomness. For pre-frontier and frontier areas too it can be said that sample means are not statistically different from the population's. For transition and consolidated areas, the sample means are statistically different from the population means. For the latter, this is most likely due to an important growth of the herd from 2010 to 2013, not to sampling. For transition areas, however, the result suggests that the sample is not representative.

In terms of distributions, a Kolmogorov-Smirnov test for the equality of cumulative distribution functions only rejects the null hypothesis—of random sampling—for consolidated areas. For the state as a whole, the test suggests that the sampled distribution is indistinguishable from the population distribution.

We further compare the sampled farm size distribution to the population's distribution. In this case, however, the population data is very noisy as the government agency that collects it (IDARON) is focused on herd sizes, not farm sizes, so farm size intervals are used to reduce error. Figure 3 shows that the distribution of sampled observations is again very close to the population distribution.

[Figure 3 about here]

Finally, we test for spatial randomness. Taking the INCRA property grid as a proxy for the population of farm locations, we calculate the average number of neighbours for every plot in the population and in the sample, compute municipality averages and run tests of equality

between sample and population. A spatially random sample is expected to be clustered (high number of neighbours) where the population is clustered, and disperse (low number of neighbours) where the population is spread, so we also run non-parametric tests of equality between distributions. Table 5 presents the results.

[Table 5 about here]

The average number of neighbours is sensible to the distance band used, so we used two different methods—same band for all municipalities, and a different band for each (the smallest band that allocates at least one neighbour to every plot), obtaining similar results. The upper lines in the table above show municipalities where the sampling was successful in obtaining randomness. In Cacoal, Cujubim and Machadinho, both the t-test the k-s test indicate random sampling. As for the other municipalities, Ouro Preto and Guajará show ambiguous results. Ouro Preto has a thin shape which increases the problem of counting neighbours on edges, but from plotting the sample on the map it is clear that the sampling was spatially random. In Guajará-Mirim the overall sample is small (25), and a few farmers did not provide spatial data, so results are probably biased indeed.

As for the three lower lines in the table, the results indicate one downside to the sampling strategy: that the spatial distribution of the IDARON agencies affects the sampling. Farmers can do their paperwork at any of the agencies across the State, no matter where their farm is located, and very often agencies are placed near municipal borders, so farmers from one municipality will visit the agency at a neighbouring municipality. This was exactly the case for Buritis, Campo Novo and São Miguel, where the sampling missed parts of the municipalities.

In all cases, the results clearly suggest that the sampling procedure was reasonable for the State as a whole. This assures that inference can be made for the State with no presumption of selection bias. Our research method thus shows that it is possible to generate approximately representative, non-biased samples at a relatively low cost. Had we visited properties

randomly to achieve a similar result, the human and financial resources required would have been orders of magnitude higher.

5. Discussion and conclusions

Data collection in a vast area such as the Amazon where subjects are dispersed and transport infrastructure is poor tends to be a strenuous enterprise whatever the epoch. While modern technologies have made the task easier, spatially referenced household surveys do still demand a considerable amount of resources, making it challenging for researchers with limited funds to collect spatially explicit survey data. In this paper we present a method that takes advantage of the internet and ever smaller computers to generate data on cattle ranching in the Amazon in a more time and money-efficient manner.

No sampling protocol can be perfect; populational structures are seldom known in advance, and when known it is in an imperfect way as cadastral bases get rapidly outdated. In this context, the strategy of sampling by approaching farmers in a central place was a way of circumventing some of the key challenges faced by land use surveys. As a result, despite potential room for sampling bias in individual strata and municipalities, the sample passed all tests in what regards the overall population, so generalizations can in principle be made based on the survey data.

Two standard data collection procedures make georeferenced surveys especially expensive and time-consuming: in-situ interviews and farmer-assisted GPS data collection. Interviewing subjects in-situ implies the use of a property grid for the sampling framework, which creates the challenge of obtaining (or generating from scratch) a property grid of the whole population under study. Once the grid is retrieved, surveyors must travel long hours to reach the sampled plots, with no guarantee that a respondent will be present. If the farmer is absent, surveyors will try once or twice more before sampling another farm in a different location. The procedure is tedious and resource-consuming.

Secondly, obtaining a boundary map of properties by using a GPS requires farmers to be willing to give not only their time, but also to give out sensible information on their plots. Besides being highly time consuming, this requires a non-trivial level of commitment.

Can survey data be made representative and spatially explicit at a cost accessible to a wider body of researchers (especially those in developing countries)? Based on the method advanced here, the answer should be affirmative. The anecdotal evidence on costs that we provide in the introduction suggests that our method is accessible to relatively small research budgets, and that it can be much cheaper to implement than a traditional set up.

Two major methodological innovations are adopted to conduct a standard household survey while also generating a detailed property grid of the surveyed farms. First, the problem of sampling from a previously existing property grid and travelling to sampled plots is bypassed by taking advantage of a simple institutional opportunity. Furthermore, we explore the fact that all farmers raising cattle in the State of Rondônia (as in other States) are legally obliged to appear in person at a government agency (IDARON) to report the vaccination of their cattle herds in April / May every year. Since 98.75% of farmers do comply with this particular law (according to official figures), it can be safely assumed that approximately the full population of farmers appears at IDARON sometime in a 30-day window.

It follows that a sampling procedure that draws respondents based on arrivals at the IDARON agency is approximately free of the gravest of all biases: self-selection. It is of course to be expected that the probability of arrival is not totally random. To account for the various factors that can influence the likelihood of a farmer showing up at IDARON at a given time, the sampling is made based on previously known frequency of arrivals to the IDARON agency. For example, a large share of farmers appear in the last few days of the vaccination reporting period, so a proportional frequency of interviews was allocated to that period. The result is a

sample that is stratified over time, with five strata representing weeks, and randomized within weeks.

More generally, using the institutional opportunity allowed us to avoid relying on an existing property grid, as well as avoiding travel costs. Moreover, the fact that the survey was being supported by the government's livestock sanitation agency (IDARON) was instrumental in creating trust vis-à-vis farmers, who tend to have a positive view of that agency while being suspicious of strangers.

The second innovation consisted in drawing property maps by using the visual input given by the subjects on a computer screen. This not only avoids the use of a GPS, but is likely more accurate inasmuch as the borders of the plot can be drawn in whatever irregular shape is necessary, sufficing that the farmer is able to correctly visualize the satellite image and recognize its constitutive features (which was true in most cases). Additionally, farmers responded to this procedure in an unexpectedly positive way. Most had never visualized their land from above and were keen to learn from the experience. More than an amusement effect, they seemed to have become more conscious of the extent to which their activities can be monitored: many were shocked to learn that anyone with access to the internet can see every change they make to their land cover.

In addition to operational advantages, our approach also seems to have passed the test of sample representativeness. In the results section we run randomness tests and find robust evidence in support of the random sampling hypothesis. We first look at the possibility of non-response bias. We analyse the effect of 3 key characteristics of farmers on the probability of a subject having dropped out of the survey, and find that despite the dropout rate of 42%, there is no evidence of non-response bias. We then compare the sample estimates of cattle herd and farm size to data on the full population, and find that the sample is representative of the population at the State level, even though it may be non-representative in one stratum, that of

transition municipalities. Finally, we use the sampled and the official government's property grids to test for spatial randomness, and find consistent evidence of random sampling in 5 out of 8 sampled municipalities.

The approach described here cannot be employed in different waves of a balanced longitudinal survey. If a fixed set of individuals is to be followed over time, then in the second and subsequent waves farmers must be interviewed in-situ, as it is impossible to foresee when and where they will appear to do their yearly vaccination reporting. For the first wave, however, the proposed method is an efficient way of building a representative sample and creating a cadastral base that may be used and updated in subsequent waves. Moreover, the use of a GPS can in principle be totally avoided in all waves of the survey.

The crucial advantage of the above innovations for this research, therefore, is to reduce financial and time costs to a manageable level. This should be true also for the work of others, as the research strategy can be easily replicated in different contexts and the institutional opportunity we used is available in other States as well as domains. Any situation in which an institutional constraint forces all members of the population of interest to visit a given location in a pre-determined period may in principle be used in a similar way to our approach.

Acknowledgements

The authors are thankful to Diana Weinhold, Daniel Caixeta, Henrique Neder, Andrei Cechin, Sandra Sequeira, Tony Hall, Ademar Romeiro, Ricardo Gomes, Zander Navarro, Alberto Lourenço, and Geraldo Martha for helpful comments and suggestions on research design. Leonardo Araújo and Derquiane Sabaini have provided great assistance in managing fieldwork. This research would not have been possible without the support of the State of Rondônia's Livestock Sanitation Control Agency (IDARON) and Secretary of Environment (Sedam). Funding was provided by the CAPES Foundation (Brazilian Ministry of Education), the Brazilian National Council for Scientific and Technological Development (CNPq), and the Gordon and Betty Moore Foundation.

Declaration

The data collection methods reported here were elaborated in accordance with the London School of Economic's ethical guidelines. Subjects where provided a summary of the research's aims and scope, and asked for consent for the interview and for the mapping of their rural property. The procedures employed comply with the current Brazilian laws.

References

Balram, Shivanand, and Suzana Dragićević. 2005. "Attitudes toward urban green spaces: integrating questionnaire survey and collaborative GIS techniques to improve attitude measurements. " *Landscape and Urban Planning* no. 71 (2): 147-162.

Bell, Andrew. 2011. "Highly Optimized Tolerant (HOT) Farms in Rondônia: Productivity and Farm Size, and Implications for Environmental Licensing." *Ecology and Society* no. 16 (2): 7p.

Bernard, Enrico, Luis Barbosa, and Raquel Carvalho. 2011. "Participatory GIS in a sustainable use reserve in Brazilian Amazonia: implications for management and conservation. " *Applied Geography* no. 31 (2): 564-572.

Caviglia-Harris, Jill, and Daniel Harris. 2008. "Integrating survey and remote sensing data to analyze land use at a fine scale: insights from agricultural households in the Brazilian Amazon." *International regional science review* no. 31 (2): 115-137.

Davern, Michael, Todd Rockwood, Randy Sherrod, and Stephen Campbell. 2003. "Prepaid monetary incentives and data quality in face-to-face interviews: Data from the 1996 survey of income and program participation incentive experiment." *Public Opinion Quarterly* no. 67: 139-147.

Dunn, Christine. 2007. "Participatory GIS—a people's GIS?" *Progress in human geography* no. 31 (5): 616-637.

Elwood, Sarah. 2006. "Critical issues in participatory GIS: deconstructions, reconstructions, and new research directions." *Transactions in GIS* no. 10 (5): 693-708.

Futemma, Célia, and Eduardo Brondizio. 2003. "Land reform and land-use changes in the lower Amazon: Implications for agricultural intensification." *Human Ecology* no. 31 (3): 369-402.

Hansen, Robert. 1980. "A self-perception interpretation of the effect of monetary and nonmonetary incentives on mail survey respondent behavior." *Journal of Marketing Research* no. 17 (1): 77-83.

James, Preston. 1938. "The changing patterns of population in São Paulo State, Brazil." *Geographical Review* no. 28 (3): 353-362.

Lorena, Rodrigo, and Eric Lambin. 2009. "The spatial dynamics of deforestation and agent use in the Amazon." *Applied Geography* no. 29 (2): 171-181.

McCall, Michael, and Peter Minang. 2005. "Assessing participatory GIS for community-based natural resource management: claiming community forests in Cameroon." *The Geographical Journal* no. 171 (4): 340-356.

McCracken, Stephen, Eduardo Brondizio, Donald Nelson, Emilio Moran, Andrea Siqueira, and Carlos Rodriguez-Pedraza. 1999. "Remote sensing and GIS at farm property level: Demography and deforestation in the Brazilian Amazon." *Photogrammetric Engineering and Remote Sensing* no. 65: 1311-1320.

Mizes, Scott, Louis Fleece, and Cindy Roos. 1984. "Incentives for Increasing Return Rates: Magnitude Levels, Response Bias, amd Format." *Public Opinion Quarterly* no. 48 (4): 794-800.

Moran, Emilio, Andréa Siqueira, and Eduardo Brondizio. 2004. "Household demographic structure and its relationship to deforestation in the Amazon Basin." In *People and the environment: approaches for linking household and community surveys to remote*

sensing and GIS, edited by Jefferson Fox, Ronald Rindfuss, Stephen Walsh and Vinod Mishra, 61-89. New York: Kluwer Academic Publishers.

Porter, Stephen, and Michael Whitcomb. 2003. "The impact of lottery incentives on student survey response rates." *Research in Higher Education* no. 44 (4): 389-407.

Raymond, Christopher, Brett Bryan, Darla MacDonald, Andrea Cast, Sarah Strathearn, Agnes Grandgirard, and Tina Kalivas. 2009. "Mapping community values for natural capital and ecosystem services." *Ecological economics* no. 68 (5): 1301-1315.

Rindfuss, Ronald, Pramote Prasartkul, Stephen Walsh, Barbara Entwisle, Yothin Sawangdee, and John Vogler. 2004. "Household-parcel linkages in Nang Rong, Thailand: challenges or large samples." In *People and the environment: approaches for linking household and community surveys to remote sensing and GIS*, edited by Jefferson Fox, Ronald Rindfuss, Stephen Walsh and Vinod Mishra, 131-172. New York: Kluwer Academic Publishers.

Särndal, Carl-Erik, and Sixten Lundström. 2005. *Estimation in surveys with nonresponse*. Chichester: John Wiley & Sons.

Sherrouse, Benson, Jessica Clement, and Darius Semmens. 2011. "A GIS application for assessing, mapping, and quantifying the social values of ecosystem services." *Applied Geography* no. 31 (2): 748-760.

Singer, Eleanor. 2002. The use of incentives to reduce nonresponse in household surveys. In *Survey Methodology Program*, edited by The University of Michigan.

Turner, Billie Lee, and Jacqueline Geoghegan. 2003. "Land-cover and land-use change (LCLUC) in the southern Yucatán peninsular region (SYPR): an integrated approach." In *People and the environment: approaches for linking household and community surveys to*

remote sensing and GIS, edited by Jefferson Fox, Ronald Rindfuss, Stephen Walsh and Vinod Mishra, 31-60. Boston: Kluwer Acad Press.

Vale, Petterson. 2015. Land use intensification in the Amazon. Revisiting theories of cattle, deforestation and development in frontier settlements. PhD thesis, Department of International Development, LSE.

Walsh, Stephen, Richard Bilsborrow, Stephen McGregor, B Frizzelle, Joseph Messina, W Pan, Kelley Crews-Meyer, Gregory Taff, and Francis Baquero. 2004. "Integration of longitudinal surveys, remote sensing time series, and spatial analyses." In *People and the environment: approaches for linking household and community surveys to remote sensing and GIS*, edited by Jefferson Fox, Ronald Rindfuss, Stephen Walsh and Vinod Mishra, 91-130. New York: Kluwer Academic Publishers.

Walsh, Stephen, and William Welsh. 2003. "Approaches for linking people, place, and environment for human dimensions research." *GeoCarto International* no. 18 (3): 51-61.

Willimack, Diane, Howard Schuman, Beth-Ellen Pennell, and James Lepkowski. 1995. "Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey." *Public Opinion Quarterly* no. 59 (1): 78-92.

Tables and Figures

Table 1. Responding and non-responding farmers

Farmers approached during survey	Total (%)
Total approached farmers	666 (100%)
Interviewed, of which:	392 (58.86%)
Agreed to provide boundaries	368 (55.26%)
Did not agree to provide boundaries	16 (2.4%)
Questionnaires lost in synchronization	8 (1.2%)
Not interviewed, but asked auxiliary questions	274 (41.14%)

Table 2. Summary of the key variables collected in the survey

Variable	Source	Observation unit	Years
Stock of cleared area: pasture, degraded pasture, fallow, crops, reforestation	Survey	Property	2000; 2005; 2010; 2013
Perceived quality of pastures	Survey	Property	2010-2013
Cattle sales (AU[1] / year)	Survey	Property	2012-2013
Cattle stocking ratio (AU[1] / ha)	Survey	Property	2013
Technology use: limestone, fertilizer, number of paddocks, artificial insemination, tractors	Survey	Property	2013
Total assets (herd, land, capital)	Survey	Farmer	2013
Perceived enforcement of forest law	Survey	Property	2013
Tenancy contracts (area of rented land, time of tenure)	Survey	Farmer	2013
Land price (R$ / ha)	Survey	Property	2013; 2016 (expected)
Migration history, intention to migrate	Survey	Farmer	2000-2013
Farm's boundaries	GIS	Property	2013

[1]Animal Units (1 AU = 450 Kg of live weight).

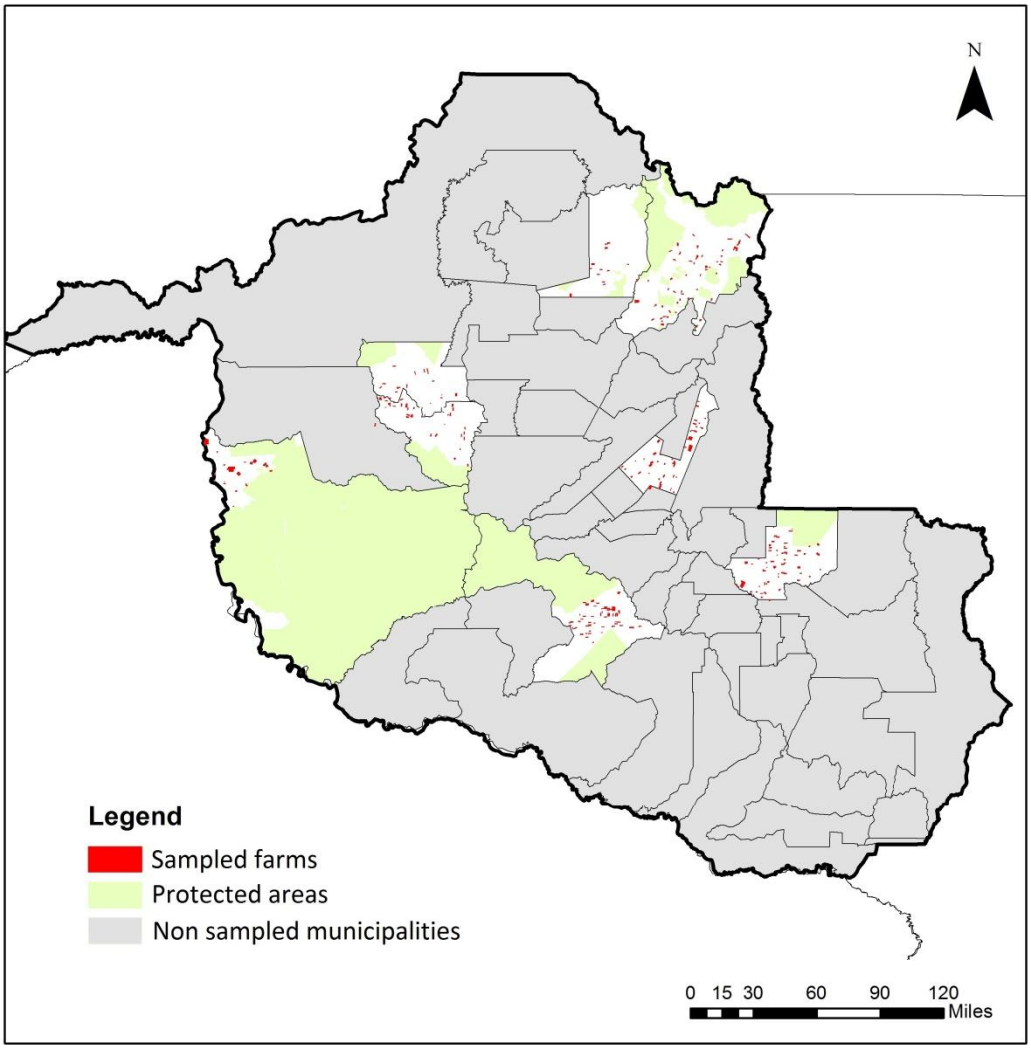

Figure 1. Sampled municipalities and farms. Colours: white are sampled municipalities, light green are protected areas, and red are sampled farms.

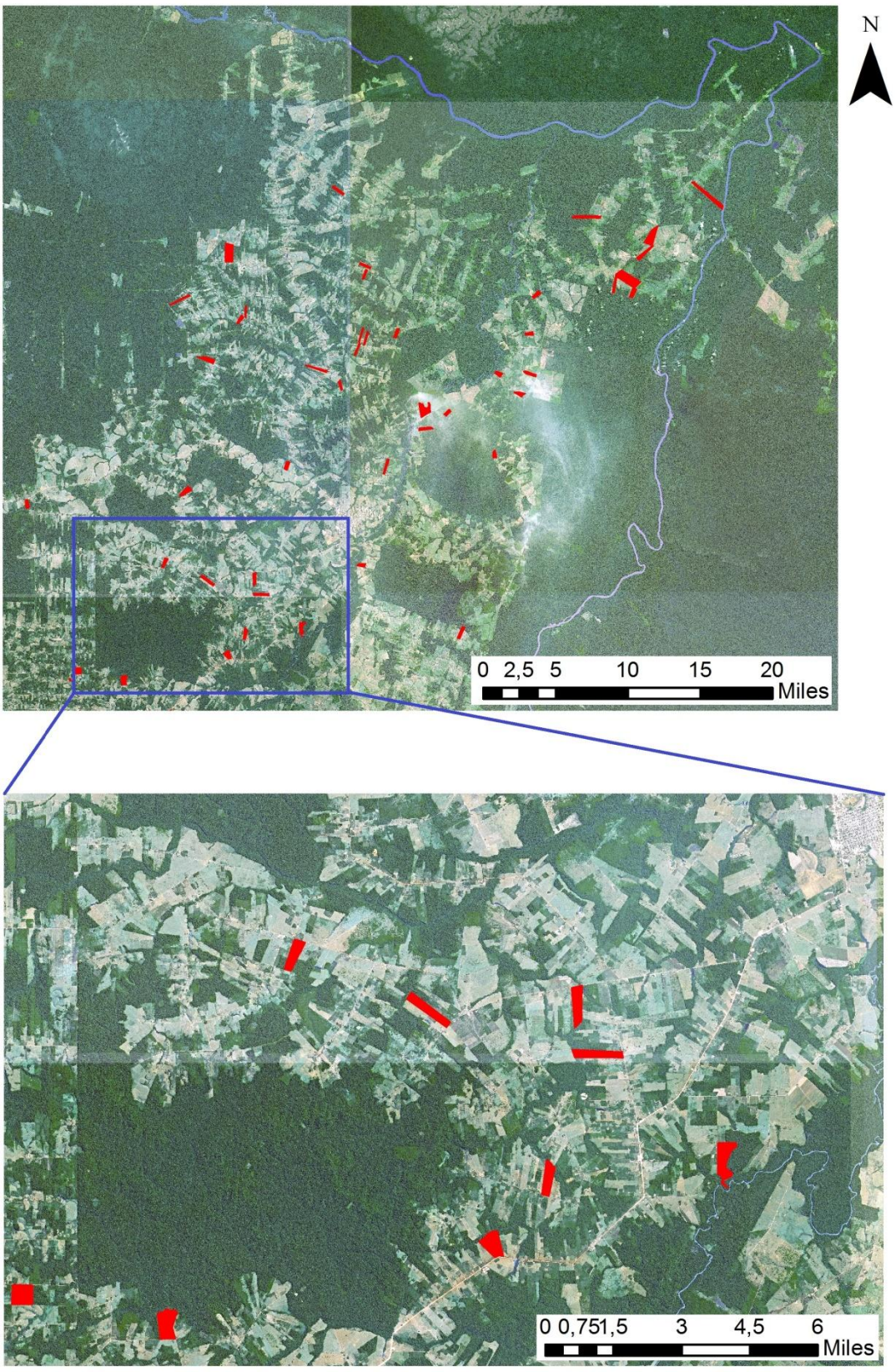

Figure 2. Mapped farms in one municipality in the State of Rondônia.

Source: Spot (2008) satellite images provided by Rondônia's Secretary of the Environment.

Table 3. Logit regression of response to survey on auxiliary variables
Binary dependent variable: response to survey (0 / 1)

	(1)	(2)	(3)	(4)
Cattle herd	-0.000409	5.10E-05	3.54E-05	4.76E-05
Pasture area	5.76E-05	-0.00122	-0.00116	-0.00112
Time in plot	0.00595	0.0138*	0.0133	0.0136
LR test (p-value)[1]	—	0.1473	0.173	0.195
Single municipalities	no	yes	yes	yes
Single surveyors	no	no	yes	yes
surveyors*municipalities	no	no	no	yes
Constant	0.235	0.23	-12.38***	-0.0463
Observations	620	620	620	615
Adj. Pseudo R-squared	0.00267	0.0387	0.0458	0.0513

[1]Likelihood-Ratio test: tests the joint significance of the auxiliary variables by comparing the fit of two models, one being nested within the other.

Robust z-statistics *** p<0.01, ** p<0.05, * p<0.1

Table 4. Comparing population and sample cattle herd data, Rondônia

	Population (2010)		Survey sample (2013)			Tests		
						Bi-caudal equality test p-values[1]		
Location	Size (*N*)	Mean	Size (*n*)	Mean	Standard-error	t-test	K-S test of distributions[2]	Correlation[3]
Rondônia	84,594	117.98	384	145.03[a]	21.41	0.19	0.459	0.92
Pre-frontier	550	183.36	21	204.76	70.62	0.76	0.120	0.94
Frontier	32,523	128.56	99	126.37	21.02	0.91	0.077	0.89
Transition	24,161	104.36	144	68.79	9.57	0.00	0.246	0.88
Consolidated	27,035	114.39	120	196.55	32.16	0.01	<0.001	0.98

[1]If lower than the significance level (normally 5%) the null hypothesis—of random sampling—can be rejected.

[2]The Kolmogorov-Smirnov test is a non-parametric test of equality of continuous, one-dimensional probability distributions. The null hypothesis is that the sample distribution is a random draw from the population distribution. The test is run by calculating a distance between the sample and population cumulative probability distribution functions.

[3]Correlation between sampled observations and a random sample of the same size taken from the population.

[a] Weighted for sample selection: due to oversampling of the pre-frontier stratum, in the absence of proper weighting the overall State statistics would be biased against the other strata. When taking the State mean I account for that by multiplying observations by the following weight: —— ——, where *n* is the sample size, *N* is the population size, and *m* is municipality. The non-weighted mean is 131 and yields the same t-test result.

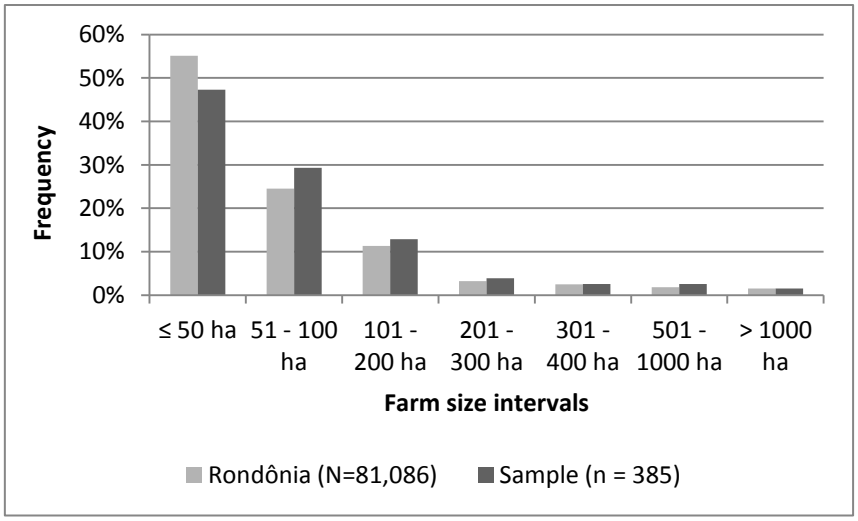

Figure 3. Comparing farm size data between population (year=2010) and sample (year=2013), Rondônia.

Table 5. Comparing population and sampled locations. Average number of neighbours (spatial clustering), Rondônia

	Population		Sample		Bi-caudal equality test p-values[2]	
Municipality	Size (*N*)	Mean[1]	Size (*n*)	Mean[1]	t-test	Kolmogorov-Smirnov test of distributions[3]
Cacoal	2,947	33.92	66	33.40	0.994	0.844
Cujubim	3,796	70.88	13	79.38	0.315	0.356
Machadinho	7,098	71.37	75	75.73	0.126	0.178
Guajará	1,770	48.20	18	37.11	0.054	0.160
Ouro Preto	2,015	12.90	41	12.02	0.111	0.054
Buritis	4,245	100.41	17	72.41	0.016	0.048
Campo Novo	1,799	51.37	35	41.17	0.029	0.028
São Miguel	3,115	54.03	57	60.54	0.027	0.009

[1]Calculated using an Euclidean distance band.

[2]If lower than the significance level (normally 5%) the null hypothesis—of random sampling—can be rejected.

[3]See note 2 in Table 4.