# USING THE GLOBAL POSITIONING SYSTEM (GPS) IN HOUSEHOLD SURVEYS FOR BETTER ECONOMICS AND BETTER POLICY*

John Gibson, *University of Waikato*
David McKenzie, *Development Research Group, World Bank*

**Revised June 2007**

## Abstract

Distance and location are important determinants of many choices that economists study. While these variables can sometimes be obtained from secondary data, economists often rely on information that is self-reported by respondents in surveys. These self-reports are used especially for the distance from households or community centers to various features such as roads, markets, schools, clinics and other public services. There is growing evidence that self-reported distance is measured with error and that these errors are correlated with outcomes of interest. In contrast to self-reports, the Global Positioning System (GPS) can determine location within 15 meters (95 percent of the time). The falling cost of GPS receivers (typically below US$100) makes it increasingly feasible for field surveys to use GPS as a better method of measuring location and distance. In this paper we review four ways that GPS can lead to better economics and better policy: (i) by helping to understand policy externalities and spillovers, (ii) through better understanding of the access to services, (iii) by improving the collection of household survey data, and (iv) through use in econometric modeling to understand the causal impact of policies. We also discuss several pitfalls and unresolved problems with using GPS in household surveys.

*JEL codes:* C81, O12, R20
*Keywords:* Global Positioning System; Distance; Location; Survey Measurement; Networks; Externalities.

* We are grateful to three anonymous referees, Kathleen Beegle, Chris Bennett, Piet Buys, Geua Boe-Gibson, Alan de Brauw, Uwe Deichmann, John Hoddinott, Ben Olken, Duncan Thomas, John Thyne and conference participants at SITE 2007 for helpful comments and advice. Financial support from Marsden Fund grant UOW0503 is gratefully acknowledged. This paper is dedicated to the memory of Piet Buys, who served as a champion for GIS use at the World Bank through his knowledge, enthusiasm and eagerness to help others use this emerging technology.

## Introduction

Distance and location are important determinants of many choices that economists study. For example, in the von Thünen model, distance to market determines land owners decisions about what crop is most profitable to produce. In studies of child labor market activity, distance from urban areas is shown to be an important determinant of both schooling and work decisions (Fafchamps and Wahba, 2006). In migration models, greater distance between origin and destination implies larger migration costs and reduces migration flows (Borjas, 2004).

While location and distance can sometimes be obtained from secondary data, economists often rely on information that is self-reported by respondents in surveys. These self-reports are especially used for the distance from households or community centers to various features such as roads, markets, schools, clinics and other public services. We provide evidence here that self-reported distances and areas are measured with considerable error and are often correlated with outcomes of interest. In contrast to self-reports, the Global Positioning System (GPS) can determine location within 15 meters (95 percent of the time).

These GPS locations are determined from satellites (currently 30) with precise atomic clocks that orbit about 20,000 kilometers above the surface of the earth and send out a unique radio signal with a time-stamp. A GPS receiver uses the time delay between transmission and reception to calculate the distance to each satellite, and the latitude and longitude of the location by triangulation. More precise calculations, including elevation, can be made if four satellites are in view (El-Rabbany, 2006). Accuracy depends partly on whether anything obscures the GPS receiver's view of the sky and the quality of the receiver used to process the satellite signal. Consumer-grade GPS receivers are accurate to within 15 meters, 95 percent of the time,[1] with further improvements in accuracy to about three meters achieved by using *differential GPS* where information from a local reference station augments that from the satellites.[2]

Two principal factors have dramatically increased the feasibility and usefulness of collecting GPS information in household surveys. First, on May 1st, 2000, the U.S. military turned off selective availability (SA), which had introduced random errors of up to 100 meters in the civilian signal. The removal of SA allowed more accurate measurement, increasing the range

[1] http://www.garmin.com/support/faqs/faq.jsp?faq=582&webPage=Main%20web%20page

[2] Specifically, a 'base station' GPS receiver is set up on a precisely known location and used to compare position based on the satellite signals with this known location. The difference is then applied to other GPS receivers in the area to correct their calculations of their unknown locations.

of possible applications. Secondly, the cost of a basic GPS receiver has fallen to under US$100, making it within the budget of most household surveys.[3] Coverage and precision will improve even further in the next few years with the launch of the European GALILEO system, expected to be operational by 2008.[4]

Surveys that are well-known to economists that have used GPS to geo-reference the locations of community centers (and hence clusters of households, given the sample design) include the Demographic and Health Surveys (DHS) (since 1997) and the Indonesia Family Life Survey (IFLS). GPS has been used to provide locations for individual households (and enterprises) in a few recent World Bank surveys including the Rural Investment Climate Surveys (RICS) in Indonesia and Sri Lanka, and the Living Standard Measurement Surveys (LSMS) in Albania and Tanzania. Nevertheless, the majority of household surveys collected in developing countries still do not geo-reference communities or households, in part through lack of information about the benefits of doing so.

An alternative to using GPS is to use secondary data on locations. In developed countries postal addresses are widely used for this purpose. For example, the United Kingdom has 2.1 million postcodes for 26 million addresses. Thus postcodes are a very accurate proxy for household location since it is possible to get map grid references to the nearest 100 meters for most postcodes.[5] These very detailed location data have been used to examine the location patterns of manufacturing by Duranton and Overman (2005). However very few developing countries have detailed post codes so this method is typically infeasible. It is also the case that in developing countries, face-to-face interviewing predominates, compared with telephone interviewing in developed countries, so it is quite feasible for field teams to gather GPS data as a part of their usual survey workload.

Another source of secondary data is from remote sensing, which is the gathering of data from a sensor mounted on either an aircraft or satellite. These data are especially used in studies of land cover, such as (de)forestation (Deininger and Minten, 2002) and urban sprawl (Burchfield, Overman, Puga and Turner, 2006). The unit of analysis is the *pixel* or picture element, which determines the size of the smallest landscape feature that can be distinguished

[3] The Garmin eTrex GPS unit has been used in a number of household surveys. On February 21, 2007 it could be purchased for $88 at Walmart.com and $93 at Amazon.com.
[4] See http://ec.europa.eu/dgs/energy_transport/galileo/index_en.htm for more details. [accessed March 12, 2007].
[5] http://www.xyzmaps.com/NewPostcode.htm

and mapped. Typical sizes are 30 meter × 30 meter or 1 kilometer × 1 kilometer grids. However, these grid cells are not individual agents and so using data at this level may involve aggregating across decision makers and a possible 'ecological fallacy' of drawing inferences about the behavior of individuals from analyses based on grouped or area-level data (Freedman, 2004). A better matching of the spatial scale of the decision process and the scale at which measurement is carried out (Anselin 2002) may come from surveying individual decision-making agents and using GPS to link them to other spatial data.

This linking of different layers of data takes place in a Geographic Information System (GIS). While GIS can be seen simply as a tool for combining, manipulating and displaying spatial information that may have been captured in a variety of ways, including GPS, a broader view sees an emerging geographic information science (Goodchild, 1992). This science may enable researchers to produce more measurements than just the distance between features and to discover new relationships for geographically referenced information. Some of the literature that we review here relies more heavily on GIS than on GPS but still serves as an example of the types of analyses that could be facilitated by a greater use of GPS in household surveys.[6]

In this paper we review four ways that GPS can help lead to better economics and better policy: (i) by helping to understand policy externalities and spillovers, (ii) through better understanding of the access to services, (iii) by improving the collection of household survey data, and (iv) through use in econometric modeling to understand the causal impact of policies. We then discuss some pitfalls, unresolved problems, and ongoing research issues.

## FOUR WAYS USING GPS CAN LEAD TO BETTER ECONOMICS AND BETTER POLICY

### *1. Using GPS can help understand policy externalities and spillovers*

The spatial proximity of one household to another may be directly of interest, particularly for understanding the interactions between actions taken by different households, the role of social networks, and the potential spillovers from policies which treat some households and not others.

[6] Recent reviews of the use of GIS in economics that are based largely on developed countries are Overman (2006) and Bateman et al. (2002).

One example of interactions between households is the possibility that they learn from one another's actions. Conley and Udry (2005) study learning in the context of the decision to adopt pineapple in Ghana, and of how much fertilizer to apply to it. They note that the classic identification problem here is that the fact that a farmer is more likely to adopt a new technology soon after his neighbors have done so might just be a consequence of some unobserved variable that is spatially correlated - such as soil types, pests or topographic features – rather than the result of genuine learning. They therefore use GPS to define the geographic neighbors of a given plot to be those within 1 kilometer of the center of the plot, and also collect data on who farmers talk to (informational neighbors). Controlling for the deviation of a farmer's input from his geographic neighbors, they can then identify learning through the impact of informational neighbors' choices.[7] Furthermore, they do find evidence of positive spatial correlation in unobserved shocks to the productivity of fertilizer, highlighting the importance of controlling for geographic effects when examining learning.

Another example of using GPS to study learning from neighbors is provided by McKenzie, Gibson and Stillman (2007) who study how negative employment experiences for emigrants affect the expectations of would-be emigrants. These would-be emigrants were all unsuccessful in a random ballot in Tonga that offers an opportunity for ballot winners who obtain employment to move to New Zealand. When interviewed subsequently about their employment (and income) expectations had they moved to New Zealand, the would-be emigrants greatly understated employment rates and incomes compared with the actual outcomes for the emigrants. One factor explaining this understatement is that many ballot winners who moved found that their initial job opening in New Zealand was no longer available, and news of this negative outcome appears to flow back to the would-be emigrants in Tonga. Specifically, if all ballot winning emigrants within a six kilometer circle (based on the GPS measurements) did not take up their initial job in New Zealand, the employment expectations of the ballot losers were lower by 19.6 percentage points.

The standard approach to evaluating the impact of a policy is to compare outcomes for those subject to the policy to outcomes for a comparable group not subject to that policy. However, as Miguel and Kremer (2004) point out, this can give misleading estimates of the

[7] They also allow for the error term to be spatially correlated across plots as a general function of their physical distance, using the spatial GMM estimator of Conley (1999).

effect of a policy when there are externalities. They investigate the impact of a deworming treatment in schools in Kenya. Using GPS distances at the level of the school, they control for the number of primary school pupils within a certain distance of the school, and then use the number of treated pupils within this distance to measure health spillovers. They find that naïve estimates which fail to take externalities into account would underestimate the program treatment effects, leading to the mistaken conclusion that deworming is not cost-effective. Such an approach could be extended by obtaining GPS locations of the residences of each individual child,[8] which could then be used to construct a child-specific measure of exposure to treated and non-treated children.[9] This would provide more variation in the extent of spillover, which could be used to examine the heterogeneity in treatment effects.

### *2. Information on the spatial distribution of population and services is essential to understanding access to services*

One of the most common uses of GPS information to date in developing countries has been to measure access to infrastructure and social services, particularly health care. For example, Perry and Gessler (2000) use GPS to measure access from communities to primary health care facilities in Andean Bolivia and use this to propose an alternative model of health distribution in the study area.

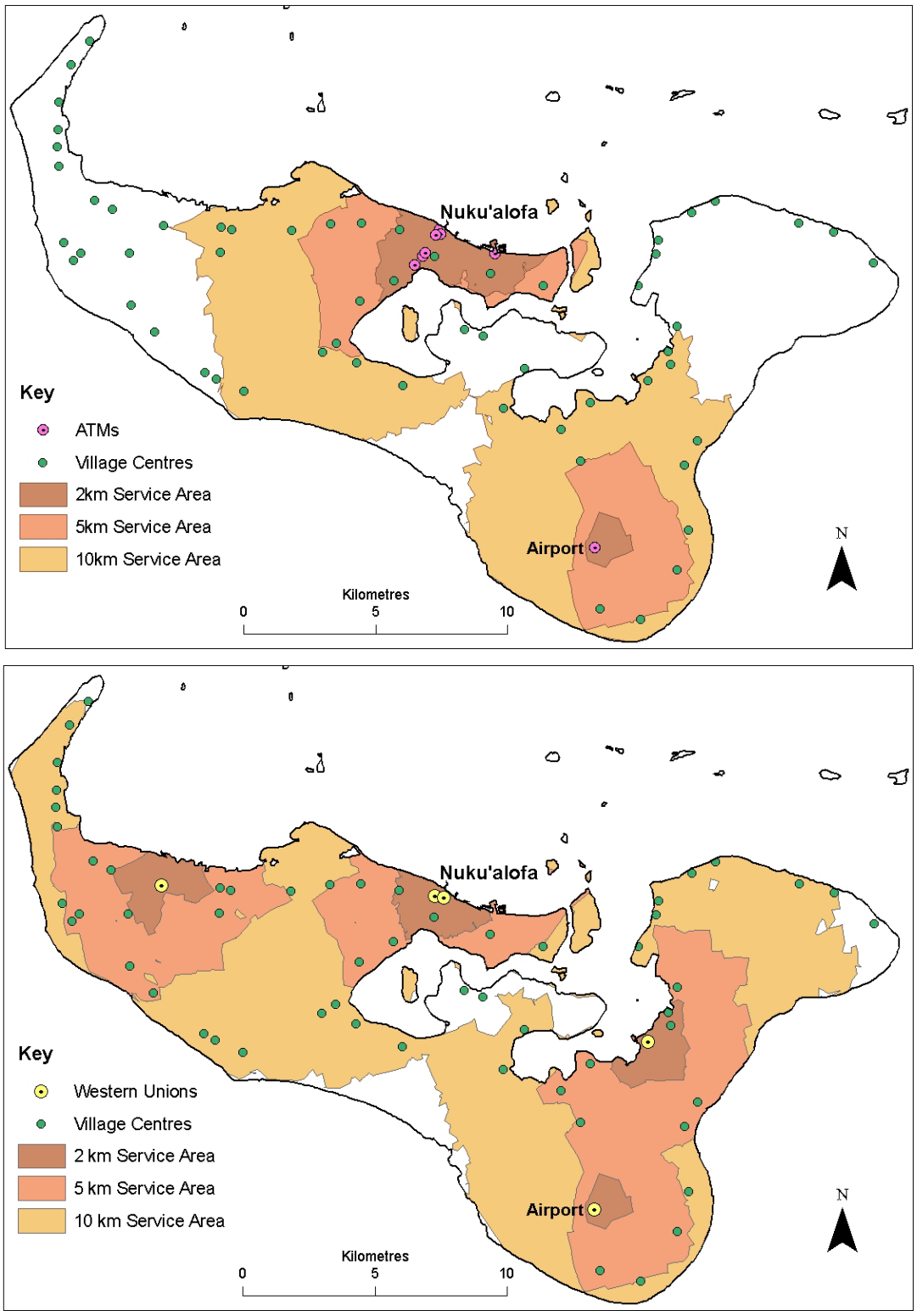
In addition to providing purely descriptive measures of access, GPS data on distance and travel times can be used to understand barriers to the use of particular services. Entwisle et al. (1997) examine the importance of accessibility to family planning on choice of contraceptive device, and in doing so, demonstrate two advantages of GPS over survey-based measures of access. They note first that data on family planning accessibility is often collected in surveys only for certain political or administrative boundaries, such as whether there is a facility in the village. However, facilities in neighboring administrative units may be closer. Using geo-referenced data allows more flexible specification of boundaries, which are not constrained by administrative definitions. Secondly, they note that reported travel times to health facilities in

[8] This is also more useful since the removal of Selective Availability allows for more accurate location measures.
[9] A recent example of looking for micro-level spillovers in a randomized experiment is found in de Mel, McKenzie and Woodruff (2007), who consider the effect of giving firms a capital shock on firms located within 500 metres, 1 kilometer and 5 kilometers of the treated firm.

their survey are often heaped in terms of 30 minute multiples, whereas using GIS gives no clumping, allowing better specification of functional form.

Gibson et al. (2006) examine the use of different financial channels for receiving remittances in Tonga. Transactions costs on money transfers are much higher using Western Union than when the recipient withdraws funds from an Automatic Teller Machine (ATM). There are eight ATMs on the main island of Tongatapu compared to five Western Union branches, so a branches per capita measure of access would suggest that ATMs are more accessible. However, they collect GPS coordinates of the ATMs and Western Union branches, and combine this with village-level population information from the Census and a digitalized road network to measure the share of the population within different travel distances of the two competing financial channels. Figure 1 shows their results. Although Western Union has fewer branches, they are more dispersed, and cover 97 percent of the population within a 10km travel distance, compared to only 77 percent of the population covered by ATMs within this distance. Figure 1 also illustrates how effective the combination of GPS data collection and mapping software can be at illustrating access in a form accessible for policymakers.

**Figure 1: Service Areas for ATMs (top) and Western Union branches (bottom) for Tongatapu, Tonga.**




Source: Figure 4 in Gibson et al. (2006)

More recent health applications combine distance with measures of infrastructure quality. Hong, Montana and Mishra (2006) use the 2003 DHS in Egypt to look at the relationship between IUD contraceptive use and the quality of family planning services available. They link each household to the nearest family planning clinic within 10km, and then use detailed DHS survey data to measure the quality of the facility. Rosero-Bixby (2004) uses GPS data on census tracts and locations of health facilities in Costa Rica to assess the extent to which health reforms led to improvements in access, measuring access with a combination of distance and services provided by the facility. He notes that households may not necessarily use the nearest facility, particularly if it is low quality, and by using GIS one can calculate measures such as the density of services that meet a standard quality within a specified radius.

A limitation with the above set of health studies is that they only measure distance at the level of a community, whereas households on opposite sides of a village or town may be closer to different facilities from each other. A second limitation is that distance to health facilities could be correlated with a host of other unmeasured factors, such as poverty, disease environment, and other infrastructure, which could also affect health decisions.

### *3. Using GPS can improve the collection of household survey data*

GPS is also starting to be used to improve the quality and cost-effectiveness of collecting household survey data. These uses occur at several phases of data collection, from the development of a sample frame, to quality control, and use for follow-up surveys. More accurate and cost-effective surveying enables researchers to carry out better analysis and provide better evidence-based advice to policymakers.

Household surveys require an accurate sample frame. The most common approach involves using a recent Census to select enumeration areas. However, censuses may become outdated during periods of rapid urbanization, and will be of little use in drawing samples in post-conflict countries that haven't had a census for decades. For example, Afghanistan is planning on completing a census in 2007, its first since 1979[10], while Lalasz (2006) reports 15 countries have not taken a Census since 1990. The traditional solution to this problem is to do area sampling, in which enumerators list all households in a well-defined block, such as a village

[10] http://afghanistan.unfpa.org/projects.html [accessed December 28, 2006]

or an area bounded by certain city streets. Such blocks are largely determined by the convenience in defining them and locating them, and can be expensive to enumerate.

Landry and Shen (2005) show how GPS can be used to do area-based sampling quickly and cheaply, since enumeration areas can be defined in terms of spatial coordinates, and made arbitrarily small. They apply this to the problem of surveying in China, where household registration lists are widely used as sample frames. Widespread migration from rural areas however means that many households are unlikely to be found on these registration lists. They use GPS to survey randomly chosen 54 × 54 meter squares (approximately one square second), and find that 45 percent of the households reached were not on household registration lists.

However, one potential problem with this approach is that the sample size is not known until data collection has occurred, since the number of households within a spatial block is not known ex ante. Landry and Shen use existing population data to create a rough population model of Beijing, but since the number of dwellings within their spatial units was four times as large as they had budgeted for they only administered their questionnaire to one-fourth of the units. It appears likely that aerial photography will alleviate such problems in the future. For example, Cowen and Jensen (1998) extracted individual dwelling unit information in a 32 census block area in South Carolina from aircraft multispectral data. They found the dwelling unit data derived from remote sensing had a correlation of 0.91 with similar data derived from the census. As the resolution of satellite imagery continues to improve and fall in price, it appears likely that the combination of remote sensing and spatial sampling will become the standard for constructing sample frames in situations where reliable census or registrar data are not available.

Another example of combining remote sensing and GPS for drawing samples is provided by Kumar (2007) in a survey of 1600 households spread across different air pollution zones in Delhi, India. The study area was partitioned into different strata characterized by air pollution levels (obtained by remote sensing) and proximity to main point sources of air pollution. Random points were then simulated using GIS techniques (weighting by the size of the residential area in each strata) and GPS was then used to navigate to the households located at each selected point. These households were then asked to participate in the survey. This method of creating a frame and drawing a sample should be more efficient than simply imposing a regular grid across the study area, since air pollution is irregularly distributed over space.

Visualization of the locations at which sampling has occurred can provide a useful form of quality control to ensure that interviewers conduct surveys where they are supposed to, and to check whether any dwellings are inadvertently missed. In 2004, Timor-Leste became the first country to use GPS units to record the locations of all households in their Census. USAID Timor-Leste (2004) reports that survey managers checked the GPS points visited by the Census teams against detailed aerial photograph maps, and used this to detect areas missed in the enumeration, sending enumerators back to complete the surveys. Census undercounting matters for a variety of policy purposes, including the division of federal money and for defining political representation. Undercount can be particularly high in developing countries – Lalasz (2006) reports that the 1991 Census is thought to have undercounted Nigeria's population (officially put at 89 million) by perhaps 20 million people. The use of GPS can help show where such undercount has occurred, and help survey managers reduce it.

Another potential use of GPS is through reducing the cost and time taken to re-locate the same dwelling for follow-up surveys. Two reasons for follow-ups are to allow field managers to check errors made by enumerators, and for the collection of panel data. A recurrent problem in many developing countries is the lack of street addresses, making re-locating the same dwelling time consuming, especially in densely populated urban areas. Furthermore, changes in administrative boundaries between waves of a survey can plague attempts to re-identify the same households. A pilot study conducted by Dwolatsky et al. (2006) with the aim of tracing patients who left a tuberculosis control program in South Africa shows the potential for using GPS to re-locate dwellings. They compared the time taken to re-find a home given residential addresses with the time with a customized personal digital assistant (PDA) linked to GPS. The time taken to find the dwelling was found to be 20-50 percent less using the PDA/GPS device. The main limitation of this study was that it was a small pilot of only 20 houses, so further experiments are needed to confirm the promising results found here.

When panel surveys attempt the more difficult task of tracking individuals rather than dwellings, GPS can be very useful for tracking people who had previously been co-residents in the same household. For example, the Kagera Health and Development Survey 2004 (KHDS 2004) in Tanzania used GPS to record the locations of 2700 households that contain members who had been in the baseline sample of 900 households first interviewed in 1991-94 (Beegle, de Weerdt and Dercon, 2006). Measures such as how far people have moved from either their

baseline village center or from households with members who had been co-residents in the baseline surveys can be related to various socioeconomic characteristics.

Finally, collecting GPS data for households allows the possibility of linking the household data set to other surveys and other datasets. There is considerable option value in doing this, since many potential uses of the data will not be known at the time of collecting the survey.

***4. GPS can be used in econometric models to help identify causal impacts***

The majority of empirical work in development economics aims to identify the effect of a particular variable of interest, *X*, on a particular outcome, *Y*. A standard concern is that there are other variables which are correlated with *X* and which also affect *Y*. Failure to control for these variables then gives biased results. One of the most basic uses of GPS is to allow researchers to better control for geographic and locational characteristics in their regressions. Such characteristics are increasingly found to be relevant to outcomes of interest for development economists and practitioners. For example, Deininger and Minten (2002) obtain data from a GIS on soil quality, rainfall, elevation, slope and other geographic features, and find that higher levels of poverty are statistically associated with *greater* likelihoods of deforestation. However, when they re-estimate the model without using the GIS data they find poverty to be associated with *lower* levels of deforestation. The problem here is that the poor live on worse quality land which limits the benefits of deforestation, so failing to control for land quality gives an opposite result.

Propensity-score matching has become a popular tool for investigating policy impacts (see Ravallion, 2006 for a recent review). The basic idea is to compare individuals subject to a policy to similar individuals not subject to the policy. Typical variables used for matching are household socioeconomic characteristics, and an often crude set of community-level variables. Brady and Hui (2006) argue that GIS can be used to more explicitly include geography in matching. They present three arguments for doing so:

1) lots of individual data that we would like to match on is unmeasured, and so place can serve as a proxy for unmeasured individual characteristics;
2) near-by places are more likely to share community characteristics, such as culture, trust, and government ability; and

3) geographic matching can be visually persuasive, if you see sudden changes in outcomes across administrative borders when a program is in one community and not its neighbor.

Nevertheless, they acknowledge that in some cases places most comparable in terms of cultural or socioeconomic characteristics may not be geographically close. Therefore it is important that matching not only be done on geography. Although the U.S. labor literature has emphasized the importance of comparing participants in training programs and non-participants from the same local labor markets (see Heckman et al. 1997), the matching literature to date has generally not explicitly included geographic proximity as a criteria when matching individuals in different communities. As more surveys include GPS coordinates, this will become increasingly possible.

The above two examples highlight the ability of GPS to help researchers to better control for (potentially) observable characteristics. A more controversial use of GPS is the use of distance or other geographical variables as instruments in instrumental variables estimation.[11] Examples include Oster (2006) who uses distance to the Democratic Republic of Congo as an instrument for HIV prevalence when examining the response of sexual behavior to HIV prevalence rates in Africa; McKenzie, Gibson and Stillman (2006) who use the GPS-measured distance from a household in Tonga to the location of the New Zealand immigration office in Tonga where application forms must be deposited as an instrument for migration, when looking at the effect of migration to New Zealand on income; and Olken (2006) who uses GIS data on community locations and geography to look at the impact of television and radio on social capital in Indonesian villages. He argues that geography leads to differences in the over-the-air signal strength in different villages due to mountains located between some villages and the transmission towers, but after including various controls, that this geography has no independent effect on social capital.

However, the use of distance as an instrument is subject to several potential problems. The first is that distance to borders and major cities is likely to also determine access to markets, schools, health facilities, and other infrastructure (see section above), which in turn can have important impacts on economic behavior. Secondly, people, villages and cities are not randomly

[11] Of course researchers can use geography to create instruments without using GPS, through manual map work. Recent examples include Woodruff and Zenteno (2007) who use distance from the capital of the state an individual was born in to the nearest station on the north/south railway lines as they existed in the early 1900s as an instrument for migration in Mexico; and Hoxby (2000) who uses the number of stream mouths in a metropolitian area in the U.S. as an instrument for the number of school districts in examining the impact of school choice. GPS can make such applications more accurate and less time-consuming.

allocated in space. As a result, distances usually incorporate the results of some behavioral choices, which may impact outcomes. The standard reply to such concerns is to try and include as many other geographic controls as possible. For example, Olken (2006) controls for district fixed effects, distance and travel time to major cities, and elevation and uses a physical model of radio transmission that predicts how signal strength should theoretically vary with topography. Nevertheless, even after including such controls, as with all instrumental variables, a case needs to be made as to why the exclusion restriction should hold – in these cases, why one should believe that unobserved geographic features are not also influencing the outcomes of interest.

A second potential concern with the use of distance as an instrument arises when the response of interest varies across individuals. Even if the exclusion restriction holds, the instrumental variables estimator will only identify the local average treatment effect (LATE) in this case. As Heckman (1997, p. 451) notes in the context of distance to the nearest school being used as an instrument for schooling, "LATE estimates the effect of variation in distance on the earnings gain of persons who are induced to change their schooling status as a consequence of commuting costs that vary within a specified range". Whether or not estimation of such a parameter is of interest to policymakers is a matter of some doubt. There will be less concern about this issue when most individuals respond in a similar manner to distance. McKenzie, Gibson and Stillman (2006) find that 98 percent of individuals who did not apply for a migration lottery gave lack of information as the main reason for not applying. A closer distance to the consulate office will increase information for most individuals, so that distance might be expected to change migration status for most individuals in their sample. Indeed, they find that using distance as an instrument gives an estimated income gain from migration within two percent of that obtained from the experimental estimator provided by a migration lottery. This provides one example where distance has been shown to provide a reasonable instrument.[12]

Similarly, it may be that shocks to local environments, as may be captured by remote sensed data in two time periods, provides a more defensible identification strategy. For example, households can be linked to areas of flooding, earthquakes, tsunamis and other such shocks. One practical constraint to implementing this at present is that although the satellite images are

[12] However, as a further note of caution, note that this application involves using distance within Tonga to instrument for migration to another location (New Zealand). As a result, one is less concerned in this application about other geographical features in Tonga affecting the outcome of interest, since this outcome is in New Zealand.

usually available, the process of converting them to usable data is costly and time-consuming with the current manual techniques.

A final use of GPS is in spatial econometric models. Many unobserved variables, such as climate and soil in agricultural settings, are spatially correlated, leading to spatial autocorrelation in the error term of regression equations.[13] Failure to account for this structure in the error terms will lead to incorrect standard errors being used for inference, possibly lead one to conclude that a policy has a significant effect when it does not, or vice versa. Distances between observations obtained through GPS can be used to account for spatial autocorrelation in the error term of the regression equation. Case (1991) and Conley (1999) provide procedures for doing this.

## HOW MUCH IMPROVEMENT DOES GPS GIVE OVER SELF-REPORTS, AND IS A STRAIGHT LINE GOOD ENOUGH?

A subset of the uses of GPS detailed above involves measuring distances from households and communities to other households, communities, or infrastructure. The natural question which then arises is whether we need to use GPS to measure these distances, or whether distances can be obtained directly through self-reports in household surveys. A follow-up question is then whether a simple straight-line (crows-fly) distance is sufficient, or whether the GPS coordinates should be integrated with GIS information on transport routes and topography to measure travel distances and travel times.

The consequences of mis-measuring distance depend on how distance is going to be used, how badly it is mis-measured, and on the nature of the mis-measurement. If measurement errors are classical, then when distance is used as a regressor, as in the studies of access, the effect will be an attenuation bias which understates the impact of distance. Using distance as an instrument with classical measurement error will lower the power of the instrument, potentially giving rise to weak instrument concerns, but will still result in consistent estimates.

However, there are strong reasons to believe that measurement errors are not random. Entwisle et al. (1984) note as an example that if people are asked to report travel times to a health provider, those who currently use the source will have more accurate knowledge than

[13] See Anselin (2002) for an accessible review of spatial econometrics. Note also that in the agricultural example given here, the omitted climate and soil variables are likely to be correlated with the regressors of interest, and so one will wish to also include detailed spatial variables as controls in the regression.

those who do not. Thus the measurement error is likely to be correlated with usage patterns, a problem if one wishes to investigate the impact of distance on usage. Indeed, Andrabi et al. (2007) report that in their survey in Punjab, Pakistan, many households do not even know the name of their nearest school, let alone its location. If the measurement error is correlated with socioeconomic variables which also affect the outcome of interest, then the mis-measured distance will also give inconsistent instrumental variable estimates.

At present there are very few studies which systematically compare self-reports of distance and travel times to GPS measurements, particularly in developing countries.[14] We use a recent World Bank survey of owners of micro- and small enterprises in Bolivia to provide the first known comparison of self-reports of physical distance to GPS-measured straight line distances.[15] Firm owners are required to register at the local branch of the National Tax system in order to receive a tax identification number: only 30 percent of the firms in the sample have registered. Firm owners were asked the distance in kilometers to the nearest tax office, which can be compared to the straight-line distance in heavily urban areas taken from the GPS coordinates of the firm and tax office. Over half the firms say they don't know the distance, with lack of a response strongly correlated with whether or not an individual is registered: 68 percent of unregistered firms say they don't know, compared to 25 percent of registered firms.

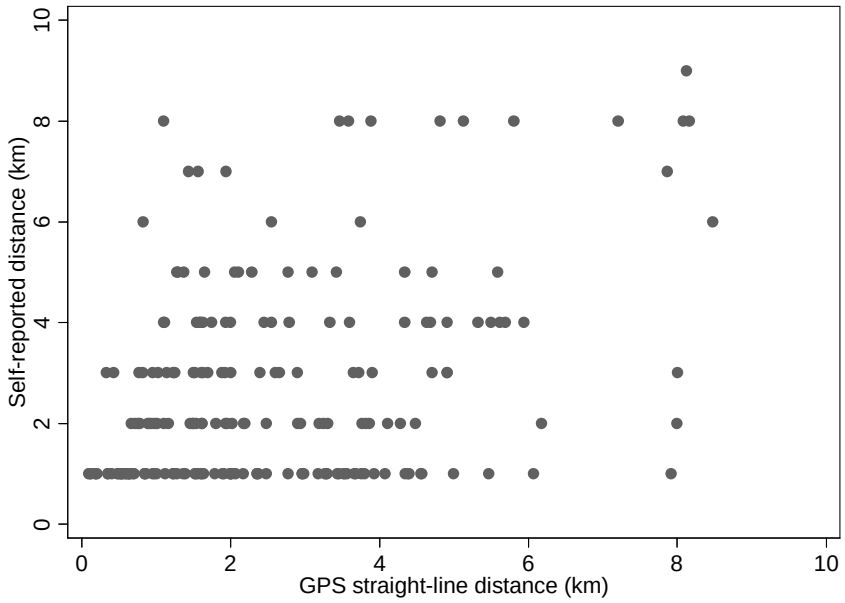
Figure 2 shows a scatterplot between the reported and measured distances for firms within 10 kilometers of the tax office, conditional on them also reporting the distance as 10 kilometers or less. The Pearson (Spearman) correlation between reported and actual distance is still only 0.39 (0.31). The degree of measurement error conditional on giving a self-report is not

[14] A related question is how self-reports of plot sizes compare with GPS-measured plot sizes. Alan de Brauw and John Hoddinott at IFPRI report (from unpublished work) that surveys in rural Ethiopia and China have generally found quite good agreement between self-reported area and measured area, whereas in Mozambique self-reports of area were very inaccurate. One hypothesis is that the accuracy of land area measurement is related to the scarcity of land: in places like China and Ethiopia, where land is scare and allocated by the Government, people have a better idea of the size of their land, than in places where land is relatively abundant. Goldstein and Udry (1999) report a correlation of only 0.15 between reported plot size and mapped plot size in Ghana, but since self-reports are in terms of "ropes", a unit of measurement which itself can vary (Hunter, 1963; Benneh, 1973), it is difficult to interpret how much of the low correlation in their work is due to measurement error in self-reporting versus the use of ropes for self-reporting. Even with errors of up to 15 meters in consumer grade units, GPS measures of plot sizes should be superior since as long as the plot corners (or perimeter) are measured at a similar time (rather than several hours or days later) with an unchanged view of the sky and number of GPS satellites, all corners will be offset by the same amount and direction, and the calculated plot area should have very little error.

[15] Roberts et al (2006) report on a survey in Bukoba, Tanzania where self-reported distance was compared with distances calculated by using pedometers and an estimate of average step length. They find that over 60 percent of self-reported distances were more than twice the calculated distances.

significantly related to whether or not the firm is registered, or to the age, gender, marital status or education of the firm owner. However, men, more educated individuals, and registered firm owners are more likely to report a distance.

**Figure 2: The low correlation between reported distance to the tax office in urban Bolivia and the GPS-measured straight-line distance.**




Source: World Bank Enterprise Survey in Bolivia.

Self-reports of time can also contain systematic measurement error. Escobal and Laszlo (2005) compare self-reports of the time in minutes it takes agricultural producers in Peru to get to the nearest populated center with the true travel time. The latter is measured by having surveyors walk with a random sample of respondents, and time their journeys, following the same route and pace as the respondent, and using GPS to measure latitude, longitude, altitude and distance. GIS is then used to compute travel time accounting for terrain for those in the survey who weren't accompanied by the surveyor. They find respondents consistently under-report the time taken to reach the center. For example, among coffee producers in the Selva, mean self-reported time was 6.7 minutes, compared to a mean true time of 13.0 minutes. The correlation between self-reported time and true travel time is only 0.28 for the coffee producers, 0.29 for the potato farmers, and -0.08 for the rice farmers in their sample. Furthermore, Escobal and Laszlo do find

that measurement errors are correlated with socioeconomic variables. Not surprisingly, individuals who own a watch give more accurate reports of travel times. They also find a negative correlation between the measurement error and education, so that more educated people have less measurement error.

These results therefore strongly suggest that self-reported distances will be misleading, with measurement errors correlated with outcomes of interest. GPS coordinates can then be used to give more accurate measurement. The simplest approach is to calculate the straight-line distance between points. This has the advantage of computational simplicity, and does not require the user to have additional geo-coded information on transport networks or topography. Alternatively, users can combine GPS point coordinates with information on the location of transportation routes, and perhaps information on road quality and topography, to measure exact travel distances and predict travel times.

The correlation between GPS straight-line travel distances and exact travel distances is likely to be much higher than the correlation between self-reported distance and GPS-distances. For example, in McKenzie, Gibson and Stillman (2006) we computed the distance in meters from each household in our sample in Tonga to the New Zealand immigration office in Nuku'alofa. The Pearson (Spearman) correlation between the straight-line and exact travel distance based on road networks is 0.82 (0.78). We do find the absolute percentage measurement error to be correlated with whether or not an individual migrates, and with their income from work in Tonga. The measurement error is greater for more remote individuals (on the other side of a lagoon), and this remoteness in turn is correlated with economic behavior. Nevertheless, the size of this error is small enough in our application that we find no difference in the IV estimates when we use straight-line distances compared to road distances. Using the straight-line distance as an instrument for migration we estimate the income gain from migration to be $280 (s.e. $122) compared to $281 (s.e. $101) when we use the road-distance.

More generally, the difference between straight-line and road distance measures will be larger when geographical features such as mountains, lakes, lagoons, and rivers lie between a household or village and the location of interest. Hence the difference between straight-line and road distances will be correlated with how remote a location is, which in turn is likely to affect many variables of interest. As a result, road distances are preferred to straight-line distances where possible.

Furthermore, since a curve between two points is always longer than a line, travel distances will be longer than straight-line distances. As a result, measures of access based on straight-line distances will over-estimate the proportion of the population which is covered by a given service. A good example of this is provided by Noor et al. (2006) who examine health coverage in Kenya, where the government has set a target of ensuring that the whole population lives within 1 hour of effective health services by 2010. Using straight-line distances which is the standard approach to coverage, they would estimate that 82% of the population is currently within 1 hour of government health services, however when they adjust this for the travel network, the proportion currently covered drops to 63-68% of the population being covered. Extrapolating this to all of Kenya, this would mean that 19 million rather than 25 million are currently covered.

## PITFALLS AND UNRESOLVED PROBLEMS

### Interviewer error

Although taking GPS readings is quite straightforward, it requires good training, otherwise surveys end up with another source of measurement error. One solution is to have interviewers take multiple readings for the same location and then use their average. Some GPS receivers have an inbuilt function for doing this. For example the guidelines for collecting GPS data in DHS surveys (Montana and Spencer, 2004) recommend taking multiple readings within a five minute period and then averaging them.[16]

### Datum and Coordinate Projection Problems

A spheroid approximation to the shape of the earth is used to solve geodetic problems for point location, and this surface, its origin and the orientation of its latitude and longitude lines make up a "geodetic datum". GPS receivers typically use the World Geodetic System 1984 (WGS84) datum, which is a geocentric datum (its origin coincides with the center of the Earth) designed for making measurements world wide. However there are hundreds of other datums which may use a different center, spheroid, or reference point on the Earth's surface in order to be locally more accurate. Coordinate values resulting from interpreting latitude, longitude, and

[16] The fieldguides of Spencer et al. (2003) and Montana and Spencer (2004) provide a good starting point for researchers planning on using GPS in a household survey. They recommend that hands-on training of interviewers with the GPS units take place outside and take at least 60 minutes.

height values based on one datum, as though they were based on another datum, can cause position errors of up to one kilometer (Ramachandran, 2000) although the discrepancy will typically be less. Bennett (2006) gives the example of walking around Tiananmen Square in Beijing with a GPS receiver and then importing the measurements into Google Earth which showed a path offset by approximately 14 meters due to the difference between the WGS84 datum and the datum used by Google Earth.

Another common source of error comes from mixing geographic and projected coordinate systems. Projected coordinates overcome the problem that latitude and longitude are not constant units so Cartesian geometry cannot be used to measure either distances or areas when working with latitude and longitude coordinates.[17] Projected coordinate systems convert latitude and longitude coordinates from the earth's three dimensional surface onto a two dimensional map.[18] Consequently if location data from a GPS (which are for a three dimensional surface and so use a geographic coordinate system) are combined with two dimensional map data (in a projected coordinate system), a conversion has to be made to line up the various data layers. This is a common and easily done task in a GIS, and building up different layers of data for households, villages, and features of interest like roads, coastlines, rivers, or public services adds value to the information in each layer. However, the data layers will not line up if different coordinate systems are used for each layer. It is surprisingly easy to incorrectly have unmatched coordinate systems because metadata, which should tell users about the coordinate system and datum used, are not always included with existing geographical data. For example, one of the authors digitized a road around the edge of the island of Tongatapu in Tonga by driving on it with a GPS receiver turned on. An existing base map of the coastline was obtained but no metadata were available to show the projection used. Even after choosing the most likely coordinate system for the base map, the road and the coastline were misaligned since it appeared that on one side of the island the road that the author had driven on was in the ocean.

[17] A degree of latitude is 110.6 kilometers at the equator and 111.7 kilometers at the poles. A degree of longitude is 111.3 kilometers at the equator, 55.8 kilometers at 60 degrees latitude and only 16.9 kilometers at 80 degrees latitude. Instead the great circle distance should be used (http://mathforum.org/library/drmath/view/51711.html) and Stata code that implements this formula is available in the "globdist" ado written by Ken Simons.

[18] For example, Universal Transverse Mercator (UTM) divides the earth into 60 zones, each spanning six degrees of longitude, ranging from latitude 80 degrees South to 84 degrees North. Locations within a zone are measured in meters eastward from its central meridian (given a value of 500,000 meters so that the furthest west point in the zone does not get a negative coordinate) and meters northward from the equator (in the Southern Hemisphere, 10,000,000 minus the meters south of the equator).

A more general issue is that in many developing countries there is not a lot of off-the-shelf geographical information available at the resolution level needed to merge it with village or community level data. Information is more often available at a coarser scale, making it difficult to link household locations to local level geographic features. Even when information is available at high resolutions, it may not always match up with the household survey due to these differences in coordinate systems. It is therefore important for countries to have a spatial data infrastructure (SDI) which coordinates different collection activities so that different geographic datasets can be matched together.[19]

**Road Network Problems**

Practical difficulties are increased when constructing a road network for measuring either distances or traveling times. The algorithm used in a GIS for calculating shortest distance requires good alignment between the lines and junctions of the digitized road network. For example, if digitized road segments at a junction do not line up, the algorithm will back-track and seek a path elsewhere. These problems may be especially apparent when roads have been digitized for another reason, such as cartographic display, so once again metadata about the origin and purpose of geographic layers used in conjunction with GPS data is very useful. It is also sensible to budget considerable research assistance time to clean a roads dataset since misalignment problems may be common. For example, the road network underlying the service areas for money transfer facilities in Figure 1 took more than one week to clean. The digitized roads had been obtained from an earlier cartographic project rather so even though it looked like a digital road network it was more like a picture of a network and a large effort was required to convert it into the continuous lines and junctions needed for calculating travel distance and time.

One way of reducing the effort required to obtain a usable roads network dataset is to digitize only the main roads and then assume a network of feeder roads, which will automatically have nicely aligned junctions. This approach is used by Staal et al. (2000, 2002) who study market access and its effect on market participation and technology adoption for smallholder Kenyan dairy farms. To build a road network linking their sample of farms to Nairobi and other urban areas they use topographic maps to digitize three classes of roads: 1) all weather, bound

[19] The Global Spatial Data Infrastructure Association provides a codebook of how an SDI can be built. See www.gsdi.org [accessed February 21, 2007].

surface, 2) all weather, loose surface, and 3) dry weather only roads. Since this left many of the surveyed farms still off the actual road network they add a 4-kilometer grid of assumed feeder roads to fill in the areas between existing roads. It is not clear how much error is introduced by using this combination of actual and assumed roads.

**Confidentiality Issues**

The accuracy of GPS in measuring either household or community locations also poses a challenge for maintaining the privacy of survey respondents. VanWey et al. (2005) discuss how the ethical need to ensure confidentiality of information collected about human research subjects may conflict with a desire to link the characteristics and actions of individuals or households to a particular geographic location. Uncertainty about how best to proceed in these circumstances may mean either that spatially explicit data are under-utilized, undermining the role of data sharing and data preservation in advancing science, or that researchers inadvertently disclose information that can identify survey respondents. These conflicts affect not only the original producers of data but also any data archivists charged with maintaining the database and providing it to other researchers while continuing to honor the commitments to confidentiality made when the data were first collected.

Moreover, this conflict between the confidentiality and usability of GPS data is not limited just to the sharing of data. It also affects the reporting and display of results based on geo-referenced data. For example, to show the confidentiality challenges posed by mapping point data, Curtis, Mills and Leitner (2006) re-engineer (i.e., reverse address match back to an individual residence) a newspaper map showing the locations in New Orleans where deaths occurred during Hurricane Katrina. The location marks in the newspaper each covered approximately one and a half city blocks and there were no roads and few other reference points illustrated. Nevertheless, over 30 percent of the re-engineered locations fell within 25 meters of the actual residence where a death occurred.[20] In several cases both the re-engineered location and the field verified residence where death occurred were for the same house. The authors scatter a series of random coordinates throughout the study area to show that chance alone would not give the same level of discovery. They conclude that: "[t]he fact that many of the re-

[20] The validation for the actual location where deaths occurred came from the search and rescue markings on dwellings, which were recorded during a field survey.

engineered coordinates could be used to identify an actual address, or an address within the immediate vicinity, should sound a note of caution for academics publishing maps displaying human cases as points" (p.53).

Typical approaches for maintaining confidentiality with GPS data are to limit access to approved researchers who promise not to identify respondents, to convert point data to either surfaces or distances so that individual locations are not revealed, to aggregate and report data only for larger areas, and to use some geographical masking procedure. These masking procedures add either stochastic or deterministic noise to the geographic coordinates for sampled households and communities. Many surveys use a combination of these four approaches. Human subject panels can play an important role in protecting confidentiality, but need to be aware of the costs to research of the different approaches.

Stricter approval procedures for obtaining geo-referenced data than for ordinary household survey data are common. For example, researchers have to provide additional justifications and commitments before they can obtain (masked) GPS data on community locations in DHS surveys. An even more stringent approach is to use a data enclave, which is typically located within a survey organization and accredited researchers come to the enclave to run their analysis. All output is checked for disclosure risk before release and even in this monitored environment there are typically restrictions on the linking of datasets and on the identity or location of individual respondents. Researchers often have to pay entry fees to these data enclaves and the limited number of enclave locations may act as a barrier to research. Remote (or virtual) data enclaves are being explored in some countries to overcome these problems. With a remote enclave a researcher can access the data server where the confidential data are stored, carry out the desired analysis and obtain aggregated results (e.g. choropleth maps or regression coefficients). Various rules on minimum cell sizes or the size of spatial units to be mapped can be imposed so that no individual-identifiable details are provided to the researcher. An example is provided by Cromley, Cromley and Ye (2004) where user queries yield results only if the cell contains at least six records (and the minimum population in the smallest mapping unit is about 1,000).

Point data on household locations can be converted to a continuous surface to represent the spatial distribution of either characteristics or outcomes without identifying respondents. For example, geographers have a variety of spatial interpolation methods, such as spatial variants of

the kernel density estimators increasingly used by economists (Bithell, 1990). Alternatively, point location coordinates can be replaced with distances to various features of interest in any public release dataset. However these methods are not very flexible and are likely to limit future research use of the data. Surfaces do not provide the micro data needed for studies that seek to measure causal impacts. The features of interest that distances are reported for are likely to vary from one study to another and as distances to more features are included the possibility of using triangulation to identify household locations increases. Moreover, distances are often needed for more than just features of interest. For example, knowing the position of households relative to other households helps study learning from neighbors (Conley and Udry, 2005) and helps improve the modeling of spatial autocorrelation (Gibson and Olivia, 2007).

Aggregating groups of observations into larger reporting units is a widely used method for maintaining confidentiality of both survey and census records. With GPS coordinates, the locations of individual households could be aggregated into larger areal units like Census Blocks or Census Tracts so that all that is reported is either an administrative code for the larger area or a polygon showing the boundaries of the area where the household is located. These techniques can also be applied to the visual display of data by using dot points on a map that are sufficiently large to prevent disclosure risk. For example, VanWey et al. (2005) show how the size of the required buffer around the locations of sampled schools in a U.S. survey would need to vary from only six kilometers in a city to over 50 kilometers in the countryside in order to hold disclosure risk to only five percent when mapping the sample points. Although aggregation can reduce disclosure risk it comes at the cost of seriously degrading the analyses that can be conducted since so much detailed spatial information is lost. For example, Fefferman, O'Neil and Naumova (2005) providing an example where areal aggregation yields little benefit of additional privacy and large costs in terms of impaired ability of statistical tools to analyze patterns of disease prevalence.

Geographical masking methods work by modifying the geographic coordinates linked to each household or community. Either random perturbations or affine transformations can be used. If the relationship of sample points to one and other is important while the relationship to another data layer (say, a base map or a road network) is not, then simply moving all points by some given distance and direction or rotating them about a chosen point may preserve

confidentiality and not greatly degrade usability of the data.[21] Normally however, point locations obtained with a GPS are merged onto other data layers so these affine transformations will introduce error that reduces the usability of the GPS data. Another option is to introduce random perturbation around the original point with the radius of the perturbation circle chosen by the data custodian, possibly weighting the size of the circle by population density at each point to take into account the effect of population density on the risk of disclosure (Kwan, Casa and Schmitz, 2004). For example, many recent DHS surveys include tests for HIV infection and because of confidentiality issues related to HIV status, up to two kilometers of random error in any direction is added to cluster locations in urban areas, and up to five kilometers of random error is added to cluster locations in rural areas. Additionally, one point in each survey with HIV testing is displaced up to 12 km in any direction.[22] Only limited research has been conducted on the effect of random perturbation on either disclosure risk or the accuracy of results. Kwan et al (2004) show that as the size of the perturbation circle increases the accuracy of results diminishes. Zimmerman and Pavlik (2006) show that releasing metadata about the perturbation methods and having different masked versions of the same dataset (e.g., a spatially aggregated dataset and a randomly perturbed dataset) can considerably increase disclosure risk.

## CONCLUSIONS

The removal of selective availability and falling costs of GPS receivers have made the collection of GPS information increasingly feasible in household surveys – yet many household surveys still do not use this technology. This paper argues that the collection of GPS coordinates should become a routine part of household survey collection, since doing so can lead to better economics and better policy advice. In particular, we have shown how GPS is being used to help better measure and understand the causal impacts of policies, policy externalities, and access to services. In addition, using GPS can improve the quality of the household survey data collected.

Moreover, one of the greatest arguments for collecting GPS information now is the option value it gives for unforeseen future applications.[23] As the stock of geo-referenced data increases within developing countries, we are likely to see a number of innovative applications

[21] A variant of this approach, rotating points in each sample cluster, is used by the Rural Investment Climate Survey in Indonesia.
[22] For more details see: http://www.measuredhs.com/topics/gis/methodology.cfm.
[23] See Turner (2006) for some intriguing ideas with regard to visual representation of geo-coordinates.

which combine household survey data with other geographical information. There are also a number of interesting new research questions in econometrics and sampling methodology which arise out of the use of GPS. For example, typical household surveys often involve population-based clusters, which are not randomly spread across geographic areas (Montana and Spencer, 2004). As a consequence, more research is needed to determine how to best estimate spatial models using a sample which is not spatially representative at the local level, and to determine how to best sample within communities in order to best allow both spatial and non-spatial uses of the data.

The increasing prevalence of GPS data will create greater research value if current practices are improved somewhat. For example, the provision of accurate, clear and timely metadata about the various data layers in existing GIS collections would allow more seamless and reliable merging of such data with GPS data. In our opinion, the best way to ensure that respondent privacy is maintained whilst maximizing the research value of GPS data that are collected is to have surveys approved through human subjects panels and the release of data to valid researchers who sign confidentiality provisions.

**References**


Andrabi, Tahir, Jishnu Das, Asim Khwaja, Tara Vishwanath, and Tristan Zajonc (2007) "The Learning and Educational Achievement in Punjab Schools (LEAPS) Report" *Mimeo* The World Bank.

Anselin, Luc (2002) "Under the hood: Issues in the specification and interpretation of spatial regression models", *Agricultural Economics* 27: 247-67.

Bateman, I, A. Jones, A. Lovett, I. Lake and B. Day (2002) "Applying Geographical Information Systems (GIS) to Environmental and Resource Economics", *Environmental and Resource Economics* 22:219-69

Beegle, Kathleen, Joachim De Weerdt and Stefan Dercon (2006) "Kagera Health and Development Survey 2004 Basic Information Document" The World Bank www.worldbank.com/lsms/country/kagera2/docs/KHDS2004%20BID%20feb06.pdf [accessed March 13, 2007].

Benneh, George (1973) "Small-scale Farming Systems in Ghana", *Africa: Journal of the International African Institute* 43(2): 134-146.

Bithell, John (1990) "An application of density estimation to geographical epidemiology", *Statistics in Medicine* 9(5): 691-701.

Borjas, George (2004) "Economics of migration" *International Encyclopedia of the Social and Behavioral Sciences* pp. 9803-9809.

Brady, Henry and Iris Hui (2006) "Is it worth going the extra mile to improve causal inference? Understanding Voting in Los Angeles County", Mimeo. Department of Political Science, UC Berkeley.

Burchfield, Marcy, Henry Overman, Diego Puga and Matthew Turner (2006) "The determinants of sprawl: A portrait from space", *Quarterly Journal of Economics* 121(2): 587-633.

Case, Anne (1991) "Spatial patterns in household demand", *Econometrica* 59(4): 953-65

Conley, Timothy (1999) "GMM estimation with cross-sectional dependence", *Journal of Econometrics* 92(1): 1-45.

Conley, Timothy and Christopher Udry (2005) "Learning about a new technology: Pineapple in Ghana", Mimeo. Yale University.

Cowen, David and John Jensen (1998) "Extraction and Modeling of Urban Attributes using Remote Sensing Technology", pp. 164-88 in D. Liverman, E. Moran, R. Rindfuss and P. Stern (eds.) *People and Pixels: Linking Remote Sensing and Social Science*, National Academy Press, Washington D.C.

Cromley, Ellen, Robert Cromley and Yanlin Ye (2004) "On-line reporting an mapping of spatially aggregated individual records selected by user queries" *Cartographica* 39(2): 5-13.

Curtis, Andrew, Jacqueline Mills and Michael Leitner (2006) "Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina" *International Journal of Health Geographics* 5(1): 44-56.

Deininger, Klaus and Bart Minten (2002) "Determinants of Deforestation and the Economics of Protection: An Application to Mexico", *American Journal of Agricultural Economics* 84(4): 943-960.

De Mel, Suresh, David McKenzie and Christopher Woodruff (2007) "Returns to Capital in Microenterprises: Evidence from a Field Experiment", *World Bank Policy Research Working Paper No. 4230.*

Duranton, Gilles and Henry Overman (2005) "Testing for Localisation Using Micro-Geographic Data", *Review of Economic Studies* 72(4): 1077-1106.

Dwolatzky, Barry, Estelle Trengove, Helen Struthers, James McIntyre and Neil Martinson (2006) "Linking the global positioning system (GPS) to a personal digital assistant (PDA) to support tuberculosis control in South Africa: a pilot study", *International Journal of Health Geographics*, August 16, 5:34.

El-Rabbany, Ahmed (2006) *Introduction to GPS: The Global Positioning System* Artech, Boston, MA.

Entwisle; Barbara, Albert Hermalin, Peerasit Kamnuansilpa, and Apichat Chamratrithirong (1984) "A Multilevel Model of Family Planning Availability and Contraceptive Use in Rural Thailand" *Demography*, 21(4): 559-574.

Entwisle, Barbara, Ronald R. Rindfuss, Stephen J. Walsh, Tom P. Evans, and Sara R. Curran (1997) "Geographic Information Systems, Spatial Network Analysis, and Contraceptive Choice" *Demography*, 34(2): 171-187.

Escobal, Javier and Sonia Laszlo (2005) "Measurement Error in Access to Markets", Mimeo. McGill University.

Fafchamps, Marcel and Jackline Wahba (2006) "Child labor, urban proximity and household composition" *Journal of Development Economics* 79(2): 374-397.

Fefferman, Nina, Eileen O'Neil and Elena Naumova (2005) "Confidentiality and confidence: Is data aggregation a means to achieve both? *Journal of Public Health Policy* 26(4): 430-449.

Freedman, David (2004) "Ecological inference and the ecological fallacy" *International Encyclopedia of the Social and Behavioral Sciences* pp. 4027-4030.

Gibson, John, Geua Boe-Gibson, Halahingano Rohorua and David McKenzie (2006) "Efficient Financial Services for Development in the Pacific", Mimeo. University of Waikato and World Bank.

Gibson, John and Susan Olivia (2007) "Spatial autocorrelation and non-farm rural enterprises in Indonesia", *Paper presented at the 51st Conference of the Australian Agricultural and Resource Economics Society*, Queenstown, February, 2007.

Goldstein, Markus and Christopher Udry (1999) "Agricultural Innovation and Resource Management in Ghana", Final Report to IFPRI under MP17, Mimeo. Yale University.

Goodchild, Michael (1992), "Geographical Information Science", *International Journal of Geographical Information Scienc*e 6(1): 31-45.

Heckman, James (1997) "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations", *Journal of Human Resources* 32(3): 441-62.

Heckman, James, Hidehiko Ichimura, and Petra Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64(4), 605-654.

Hong, Rathavuth, Livia Montana and Vinod Mishra (2006) "Family planning services quality as a determinant of use of IUD in Egypt", *BMC Health Services Research*, June 22, 6:79.

Hoxby, Caroline (2000), "Does Competition Among Public Schools Benefit Students and Taxpayers?", *American Economic Review*, 90(5):1209-38.

Hunter, John (1963) "Cocoa Migration and Patterns of Land Ownership in the Densu Valley near Suhum, Ghana", *Transactions and Papers* 33:61-87.

Kumar, Naresh (2007) "Spatial sampling for collecting demographic data" *Paper presented at the Annual Meeting of the Population Association of America*, New York, March, 2007.

Kwan, Mei-Po, Irene Casas and Ben Schmitz (2004) "Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica* 39(2): 15-28.

Landry, Pierre F. and Mingming Shen (2005) "Reaching Migrants in Survey Research: The Use of Global Positioning System to Reduce Coverage Bias in China", *Political Analysis* 13:1-22

Lalasz, Robert (2006) "In the news: The Nigerian Census", Population Reference Bureau, http://www.prb.org/Template.cfm?Section=PRB&template=/ContentManagement/ContentDisplay.cfm&ContentID=13767 [accessed December 28, 2006].

McKenzie, David, John Gibson and Steven Stillman (2006) "How Important is Selection? Experimental Vs Non-experimental Measures of the Income Gains from Migration", *World Bank Policy Research Working Paper No. 3906*

McKenzie, David, John Gibson and Steven Stillman (2007) "A land of milk and honey with streets paved with gold: Do emigrants have over-optimistic expectations about incomes abroad?" *Mimeo* The World Bank.

Miguel, Edward and Michael Kremer (2004) "Worms: Identifying Effects on Education and Health in the Presence of Treatment Externalities", *Econometrica* 72(1): 159-217.

Montana, Livia and John Spencer (2004) "Incorporating Geographic Information into MEASURE surveys: A Field Guide to GPS Data Collection", *MeasureDHS*, *http://www.measuredhs.com/basicdoc/gps/DHS_GPS_Manual.pdf* [accessed February 21, 2007].

Noor, Abdisalan, Abdinasir Amin, Peter Gething, Peter Atkinson, Simon Hay and Robert Snow (2006) "Modelling distances travelled to government health services in Kenya", *Tropical Medicine and International Health* 11(2): 188-96.

Olken, Benjamin (2006) "Do Television and Radio Destroy Social Capital? Evidence from Indonesian Villages", BREAD Working Paper No. 130

Oster, Emily (2006) "HIV and Sexual Behavior Change: Why not Africa?", Mimeo. University of Chicago.

Overman, Henry G. (2006) "Geographical Information Systems (GIS) and Economics", forthcoming in S. Durlauf and L. Blume (eds.) *The New Palgrave Dictionary of Economics*, Palgrave Macmillan.

Perry, Baker and Wil Gessler (2000) "Physical access to primary health care in Andean Bolivia", *Social Science and Medicine* 50(9): 1177-88.

Ramachandran, R (2000) "Public access to Indian geographical data", *Current Science* 79(4): 450-467.

Ravallion, Martin (2006) "Evaluating Anti-Poverty Programs", forthcoming in R.E. Evenson and T.P.Schultz (eds.) *Handbook of Development Economics, Volume 4*, Amsterdam, North-Holland.

Roberts, Peter, KC Shyam and Cordula Rastogi (2006) "Rural Access Index: A Key Development Indicator" *Transport Papers* No. 10, Transport Sector Board, The World Bank.

Rosero-Bixby, Luis (2004) "Spatial Access to Health Care in Costa Rica and its Equity: A GIS-Based Study", *Social Science and Medicine* 58(7): 1271-84.

Spencer, John, Brian Frizzelle, Philip Page and John Vogler (2003) *Global Positioning System: A Field Guide for the Social Sciences*, Blackwell Publishing: Oxford.

Staal, S., Delgado, C., Baltenweck, I., and Kruska, R. (2003) "Spatial aspects of producer milk price formation in Kenya: a joint household-GIS approach", *Paper presented at the 24th Conference of the International Association of Agricultural Economists*, Berlin, August, 2000.

Staal S., Baltenweck I., Waithaka M., de Wolff T. and Njoroge L. (2002) "Location and uptake: Integrated household and GIS analysis of technology adoption and land use, with application to smallholder dairy farms in Kenya", *Agricultural Economics* 27(2): 295-315.

Turner, Andrew (2006) "Introduction to Neogeography", *O'Reilly Short Cuts*, December.

USAID Timor-Leste (2004) "East Timor Completes the World's First GPS-Based Census", USAID Timor-Leste Small Grants Program Highlights, http://timor-leste.usaid.gov/PrintVersion/SGArchive51Print.htm [Accessed December 28, 2006]

VanWey, Leah, Ronald Rindfuss, Myron Gutmann, Barbara Entwisle, and Deborah Balk (2005) "Confidentiality and spatially explicit data: concerns and challenges", *Proceedings of the National Academy of Sciences* 102, pp.15337-15342.

Woodruff, Christopher, and Rene Zenteno (2007) "Migration networks and microenterprises in Mexico", *Journal of Development Economics*, 82(2): 509-28.

Zimmerman, Dale and Claire Pavlik (2006) "Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data", *Mimeo* Department of Biostatistics, University of Iowa.