Impact Evaluation of Feeder Roads:

Phase 1, Baseline Findings Report

NORC
at the UNIVERSITY OF CHICAGO

Report submitted to:
Government of Ghana
Millennium Development Authority
4th Floor Heritage Tower, 6th Avenue
Ridge West, Accra, Ghana

Report submitted by:
Raymond J. Struyk, Joseph G. Caldwell,
John Felkner, Kareem Kysia, and Shova KC
National Opinion Research Center
4350 East-West Highway
Bethesda, MD 20814
Tel. (001)(301)634-9300

September 20, 2010

Executive Summary

Under the Agricultural Project being implemented by Ghana's Millennium Development Authority (MiDA) some feeder roads are to be rehabilitated or reconstructed to promote development in the sector. In the first phase, about 336 km of feeder roads in eight (8) districts in two intervention zones are to be rehabilitated to reduce transportation costs and time, and increase access to major domestic and international markets. The feeder roads activity will also facilitate transportation linkages from rural areas to social service networks (including hospitals, clinics and schools).

The purpose of this project is to conduct an impact evaluation of the MiDA's Feeder Roads Activity. As stated in the Terms of Reference of the request for proposals, "the primary data for the impact evaluation will be a series of surveys similar in scope to the Consumer Price Index (CPI) survey, examining changes in prices over time Findings from the market surveys will contribute to the overall impact evaluation conducted by the Institute of Statistical, Social and Economic Research (ISSER). The Ghana Living Standards Survey (GLSS) 5+ is the primary instrument used in the overall evaluation, and 'Difference in Difference' is the proposed method of evaluation of data."

Thus, this study focuses on how prices of goods sold at local markets (that are transported on improved roads) change over time. It is also to document the changes in goods transport tariffs and passenger fares to market places served by the feeder roads.

This report presents a description of the evaluation design, the method employed in drawing the sample and sample description, the implementation of baseline survey, adjustments made to the data initially collected, and spatial patterns in the prices and transportation tariffs.

The sample design uses a carefully tailored algorithm employed to match 154 localities that will benefit from the road improvements with an identical number of control localities that are comparatively far from the improvements. The sample size is sufficient to provide robust estimates of price effects associated with the road improvements. The minimum population for a locality to be included in the sample is 1,000, a condition imposed to help ensure that most designated items could be found in most localities.

Beginning in August 2009 interviewers began visiting the sample localities to obtain three price observations for each item in the defined "basket" of goods and transportation services. The final "basket" contains 39 fresh food items, 24 packaged food items, 19 non food items and 6 transportation tariffs—3 for the locality's residents' most frequent passenger destinations and 3 for the most frequent freight destinations. About half of the price observations were not obtained because the good was not available in a locality or it was only sold by one or two sellers. There were a significant number of observations where the description of the item priced differed from that specified on the pricing sheet (1 cigarette was priced rather than a pack that had been specified). In cases where only invalid data were gathered in a locality for an item, imputation procedures were employed to generate one valid observation for the place.

Chapter 4 presents item-by-item price and tariff information for control and treatment localities and summary data on the extent of price variance across localities. Prices demonstrate varying degrees of variation. Because this is the baseline report it is not possible to compute price changes associated with the road improvements.

To what degree can price variations be associated with simple measures such as the size of the locality, measured by population, or the distance from the next large community? Simple regression models are estimated for aggregate prices of goods' groups: fresh foods, packaged foods, and other goods. (They cannot be estimated for transport tariffs because the destinations differ for each locality.) The models show that accessibility matters least for fresh goods, since a good share of these are often available locally. But accessibility does have a statistically significant effect for packaged foods and non food goods. The quantitative effects of greater distance and travel times are, however, quite modest.

Overall, the models substantially confirm the kind of systematic price variation one would anticipate. This in turn lends confidence to the quality of the data collected.

Acknowledgements

The work reported here has benefitted substantially from our interaction with MiDA's Office of Monitoring and Evaluation. In particular, Mrs. Abigail Abandoh-Sam and Patrick Fosu-Siaw provided consistent advice and guidance during the execution of this project phase. They were instrumental in the team acquiring data needed for matching treatment and control localities and the Office's review of field operations brought important matters to our attention.

Dr. Emmanuel Amamoo-Otchere worked tirelessly in helping the NORC team acquire certain data used in matching treatment and control localities and was generally a valuable source on Ghanaian data sets.

Pentax Management Consultancy Services carried out the survey whose results are reported here. Dr. Mohammed Muslim was particularly helpful in managing the relations between Pentax and NORC.

Table of Contents

List of Tables

1. Introduction and Summary

Agriculture is a very significant component of Ghana's economy accounting for about 40 percent of the country's gross domestic product (GDP), employing 60-70 percent of the labor force and generating more than 55 percent of the foreign exchange earnings. Overall, poverty rates in the target areas are generally above 40 percent (income of under US$ 1 per day). In the north, as well as in parts of the Central Afram Basin area, poverty among the rural population is as high as 90 per cent. With the poverty incidence so high in rural areas, any improvement in the agricultural sector will work to improve the economy of the poor.

MiDA's Agriculture Project within the Government of Ghana's Compact with the Millennium Challenge Corporation is design to improve farming in a number of areas including:

- Increasing farmer and enterprise training in commercial agriculture
- Irrigation development
- Land tenure facilitation
- Reducing post-harvest losses
- Improving access to credit, and
- Improving linkages to farmlands and markets by rehabilitating and expanding the transportation network.

Under the Agricultural Project some feeder roads are to be rehabilitated or reconstructed. In the first phase, about 336 km of feeder roads in eight (8) districts in two intervention zones are to be rehabilitated to reduce transportation costs and time, and increase access to major domestic and international markets. The feeder roads activity will also facilitate transportation linkages from rural areas to social service networks (including hospitals, clinics and schools).

The purpose of this project is to conduct an impact evaluation of the MiDA's Feeder Roads Activity. As stated in the Terms of Reference of the request for proposals, "the primary data for the impact evaluation will be a series of surveys similar in scope to the Consumer Price Index (CPI) survey, examining changes in prices over time. Findings from the market surveys will contribute to the overall impact evaluation conducted by the Institute of Statistical, Social and Economic Research (ISSER). The Ghana Living Standards Survey (GLSS) 5+ is the primary instrument used in the overall evaluation, and 'Difference in Difference' is the proposed method of evaluation of data."

Thus, this study focuses on how prices of goods sold at local markets (that are transported on improved roads) change over time. It is also to document the changes in goods transport tariffs and passenger fares to market places served by the feeder roads.

This report presents a description of the evaluation design, the method employed in drawing the sample and sample description, the implementation of baseline survey, adjustments made to the data initially collected, and spatial patterns in the prices and transportation tariffs.

2. Approach, Sample Design, and Sample Selection

This section describes the evaluation design and sample selection process for the survey conducted to collect data for the Ghana Millennium Development Authority (MiDA) feeder roads program evaluation.

2.1 Method

The primary objective of this evaluation project is to obtain an estimate of the impact of MiDA's feeder-roads improvement program on prices of commodities purchased in local markets in Ghana. An "ideal" approach to an impact evaluation project is to measure the difference, between a randomly selected "treatment" sample (selected from a population of "treatment" units) and a randomly selected "control" sample (selected from a population of "control" units), of the difference in quantities of interest at a time before the project implementation and at a time after. This kind of research design is a "true" experimental design, and the primary estimate of impact is called a "double-difference" or "difference-in-difference" estimate. For roads-improvement projects, this ideal approach is not feasible. Roads are not randomly selected for improvement, but selected based on a variety of economic, political and technical reasons.

In this situation, when the program roads are not selected by randomization, an alternative approach is to use a "quasi-experimental" design – a design that reflects the important features of a true experimental design to the extent possible. The design is constructed in a way that achieves a high level of accuracy (high precision, low bias) for estimates of interest, or a high level of power for hypothesis tests of interest, subject to practical constraints. In the present application, the approach that is being used, in lieu of randomization, to select a control sample is statistical matching. With this procedure, a random sample of localities is selected from the treatment population, and each treatment sample unit is individually matched to a locality in the control population, using suitable match variables.

More will be said later on how the "treatment" and "control" populations are defined. For now, it is sufficient to say that the "treatment" population is localities that are likely to realize significant changes in travel time because of the program intervention (road improvements), and the "control" population is all other localities (or a large portion of them).

With roads, a key problem that arises is that the effect of the road improvement depends very much on how close a locality is to an improved road. In the treatment area, some localities are near program roads, but many are far away. If a random sample of all localities is selected from a treatment area (i.e., from a MiDA program district), many of them will exhibit little change in prices because of the program. If the mean price change is calculated for the entire area and compared to the mean price change for a non-program area, little change will be observed. Moreover, since the average change between the groups is small, a very large sample would be required to detect such a change.

To address this problem, it is important that the evaluation design take into account the distance (or travel time) of a location to a program road. This can be done by developing an analytical model that expresses program impact (price changes) at a location as a function of distance of the location from an improved road. With such a model, it is possible to compare program impact for the program and non-program areas *as a function of distance (of localities) from improved roads.* Such a measure of impact is much more sensitive to the program intervention than an aggregate measure such as the difference in means between treatment and control groups. It might be asked whether it is acceptable simply to select a sample of roads very near improved roads, and not include any that are far away. The difficulty with that approach is that, prior to conducting the survey, we do not know the relationship of program impact to distance. Moreover, the "scope of inference" of such a sample would be much more restricted than for a treatment sample selected from the entire program area.

To have high precision for the double-difference estimate, it is desirable to introduce correlations between the "before" and "after" samples, and between the "treatment" and "control" samples. This is done by using a "panel-survey" approach for the "before" and "after" samples – it will be attempted to return to the same markets in later surveys. To increase the likelihood that a market will still exist in later surveys, sampling was restricted to localities that are of moderate size (year-2000 population of 1,000 or more, and the 20 largest localities in each district). Correlations will be introduced between the treatment and control samples by selecting control sample units using the technique of statistical matching. To achieve a high level of precision (and low bias), it is desirable to match *individual units* of the treatment sample with *individual units* selected from the control population. This approach produces what is called a "matched-pairs" sample. The precision of difference estimates based on "matched-pairs" samples is usually substantially higher than the precision of samples that are simply matched "distributionally" (i.e., for which overall characteristics (such as the mean, variance, or the "shape" of the distribution) match, but for which there are no matched pairs).

As was mentioned earlier, the goal in this project is to develop a model – a "mathematical model," or an "analytical model" – that describes program impact (price changes) as a function of distance from improved roads. In this study, "distance" refers to travel time as measured by a geographic information system (GIS) model, computing travel times directly through a complete GIS model of the Ghana trunk and feeder road system, considering variation in road surface quality and topography. The approach to the design of surveys for the purpose of constructing analytical models is quite different from the approach to designing surveys that are intended simply to describe a population of interest. The latter type of survey design is called a "descriptive survey" design, and the type of survey design used to collect data for development of an analytical model is called an "analytical survey" design.

In a descriptive survey, attention centers on estimation of overall population characteristics, such as means, proportions and totals. In the design of a descriptive survey, it is usually desired that each population unit be selected with nonzero probability, and that the selection probabilities be fairly uniform. Standard techniques involved in the design of descriptive surveys are stratification, multistage (or cluster) sampling, and the use of variable probabilities of selection of sample units (such as first-stage sample units). In an analytical survey, it is desired that the sample reflect high levels of spread, balance and orthogonality (low correlation) among variables that are important to selection of the program units (roads selected for improvement) or to program impact. The main technique used to achieve such sample characteristics is marginal stratification.

In a descriptive survey, it is important that all of the population of interest be subject to sampling, so that the scope of inference extends to the entire population. As mentioned, it is also desirable (in order to have high precision for estimation of population means and totals) that all units have comparable selection probabilities. In an analytical survey, it is much less important that population units have similar selection probabilities – or even that all units of the population be subject to sampling. It is far more important that the analytical model be well specified, and that the sample units collectively reflect the full range of variation and relationships of the dependent variables of interest (program impact measures) to independent (explanatory) variables. Because of this substantial difference, it is permissible in an analytical survey to exclude portions of the population that are very remote from the program, as long as the included units reflect a full range of variation among treatment and control units. Attention focuses on the scope of inference of the analytical model, not on the scope of inference of estimates of population means and totals.

2.2 Sample Frame, Sampling Units and Sample Sizes

Early in the project design, consideration was given to the choice of appropriate sampling units. In order to produce a high ratio of precision relative to survey cost, it is often desirable in descriptive surveys to conduct multistage sampling, in which a sample of first-stage units (primary sampling units, or PSUs) is selected, and a sample of second-stage units is selected from each first-stage unit. In some cases, sampling may be extended to more than two stages. While multistage sampling reduces survey costs, it also generally reduces the precision of sample estimates, because units within a given PSU may be more similar to each other than to units in the general population (as reflected in the "intra-unit correlation coefficient"). Another consideration in the choice of sampling units is the information that is available on various sample units prior to the survey, for use in constructing the sample design. Although multistage sampling is very useful in the design of descriptive surveys, it is generally avoided (or restricted to small first-stage units) in the design of analytical surveys, because its use substantially reduces the ability to control the distribution of low-level sample units, and to perform statistical matching.

Because of the desire to do matching of treatment and control sample units, a major consideration in the choice of sample units was the availability of data that could be used for matching. The data of interest is data that has some relationship to the selection of program roads and to program impact. Consideration was given to data that exist in government statistical reporting systems, in previous sample surveys, and in geographic information systems. Data sources included the national Census, the Ghana Living Standards Survey (GLSS), and geographic information system (GIS) data bases maintained by various organizations in Ghana (see Figure 1 (p.10) for a more complete listing). After review of these sources, it was determined that the best source of data for this evaluation project was an archive of Ghana GIS database provided by a number of Ghana government agencies, primarily because these data could be spatially geo-referenced, a key factor in allowing the computation of "accessibility" to road improvement "treatment" locations through the GIS calculation of travel-time indices.

The primary considerations in choosing the source of sample-design data were data coverage and level of aggregation: data were required for a substantial portion of the country, including MiDA program areas and other areas from which a control sample could be selected. In fact, the larger the spatial coverage of the data, the more robust would be the drawing of comparison observations to treatment observations, as this would allow for the inclusion of control observations farther from road improvements and thus theoretically subject to less influence from those improvements. Also, to facilitate control of the treatment sample (spread, balance and orthogonality) and to facilitate matching of control units to treatment units, the sample design data should be available at a relatively low level of aggregation. A wide range of demographic and physiographic GIS data was available for large portions of the country. The only problem was that the GIS data were not maintained by a single centralized organization, and it was necessary to coordinate with a number of organizations and individuals to access the data. A memorandum, "Preliminary Observations on Sample Design," dated 5 March 2009, was prepared to summarize the review of the data sources.

Based on the review of data availability, it was decided that the for our *localities* – our evaluation and sample observations – the most suitable data was a GIS database of the 20 largest localities in each Ghana District prepared from the Ghana 2000 Census by the Department of Feeder Roads. This dataset provides accurate geo-locations of Ghana localities as well as their population and names. While a number of other GIS datasets of Ghana localities were available (from the Ghana Survey Department, the Ghana Highway Authority (GHA), for example), none of these included accurate, verified locality populations. Although population data were available, linked to GIS data on Ghana 2000 Census Enumeration Areas, these data had the disadvantage of not providing accurate point geo-locations of locality centers, thus making it much more difficult to conduct surveys and to calculate travel-times access indices in the GIS. For these reasons, the GIS dataset of 20 largest localities in each District was selected as the key dataset providing observation/locality geo-locations, names and populations used for this sampling (and for the larger project evaluation).

Key to the sample design and the project evaluation is data on the locations of Ghana feeder roads, which are maintained and managed by the Ghana Department of Feeder Roads, as opposed to the national trunk road system maintained by the Ghana Highway Authority (GHA). Consequently, discussions were held between MiDA and the Department of Feeder Roads in May, 2009, to obtain GIS data on Ghana feeder road networks. The resulting data-sharing agreement provided highly accurate GIS feeder road data for all Ghana regions except the Western Region, and for this reason the Western Region was excluded from the evaluation and the sample design. This is not considered a serious drawback because the Western Region is located very far from MiDA program roads (which are spatially clustered in the eastern and northern areas of the country, while the Western Region is in the extreme southwest – see Figure 1) and because it represents a small portion of the country,. (Further comments on the exclusion of Western Region will be presented later.)

Whenever statistical matching is done, it is important to do it at as low a level of aggregation as possible, and to have a wide variety of match variables to consider. The available GIS data satisfied both of these conditions. Since it was desired to return to the same markets in later survey waves and markets are not stable for small localities and because it was important that places selected be large enough that most of the goods being priced in the survey would be present, sampling was restricted, as mentioned earlier, to localities having population 1,000 or more (according to the 2000 Census) and to the 20 largest localities in each district.

Following negotiations between NORC and MiDA, it was agreed that the sample size would be a total of 308 localities for the first (time 1) survey wave. This is the total for all localities, whether "treatment" or "control." From the point of view of statistical efficiency, it is desirable to balance the sample evenly between "treatment" units and "control" units (i.e., to select 154 treatment localities and 154 control localities). After consideration of design alternatives, it was decided that the treatment population would be all localities within 120 minutes estimated travel time of the nearest MiDA program road, and the control population would be all localities located more than 120 minutes estimated travel time from the nearest MiDA program road. (The estimated travel times were calculated using a GIS model of the Ghana road network (documented separately).) For both groups, as mentioned, consideration was restricted to localities having population 1,000 or more, and the largest 20 localities in each district.

Some additional comments will be made about the choice of 120 minutes as the cutoff point for defining the treatment and control populations. First, it should be recognized that these two populations do not have to be mutually exclusive, and they can even overlap. What is important is that the definition be such that the treatment population contain most of the localities for which travel times to points of interest might be significantly affected by the program intervention (road improvements), and that the control population contain a rather large number of localities that would be little affected by the program (a rather large number in order to facilitate statistical matching of control units to the units of the treatment sample). It is realized that in a national road network, all localities would be affected to some extent by road improvements anywhere in the country. The definition of treatment and control populations will hence not be "perfect." The objective is to divide the country into two parts – those significantly affected by the program, and those little-affected by the program.

The sample described in this report is in fact the second sample to be selected for the Ghana feeder roads program evaluation. For the first sample, the treatment population was defined to be all localities within MiDA program districts, and the control population was defined to be all localities located in districts adjacent to the MiDA program districts. When this initial sample was reviewed by MiDA in July, 2009, it was found to be unsuitable. First, it was discovered that GIS data shared with NORC by MiDA in late June, 2009, had included erroneous road segments that were not actually part of the MiDA "Tranche A" road improvement schedule. This was a significant problem, and the sample had to be redesigned and reselected because of it. Second, it was observed that the definition of the treatment and control populations used for the initial sample had three shortcomings.

1. Some of the adjacent districts were located quite close to MiDA program roads. This meant that the "control" population contained some localities that logically should be included in the "treatment" population.

2. The sizes of the treatment and control populations were not very large. It turned out that the treatment sample was a large proportion of the total treatment population – 174 out of 335. Because the sample design variables are not independent, this meant that there was not a lot of flexibility in controlling the spread and balance of the treatment sample (over the desired marginal stratifications of the design variables).

3. The control population was not very large (compared to the sample) – only 600 localities (from which 174 would be selected). This meant that there was not a very “rich” population from which to select controls.

Because of these shortcomings, it was decided, when constructing the second sample (described in this report) to use an alternative approach to defining the treatment and control populations. The treatment population would be defined to be all localities for which the estimated travel time to the nearest program road was less than or equal to 120 minutes, and the control population was defined to be all localities for which the estimated travel time to the nearest program road exceeded 120 minutes. (These estimated travel times were determined by a GIS model.) This definition resulted in a much “cleaner” distinction between the treatment and control populations, and also resulted in substantially larger population sizes – 675 treatment units (vs. 335 before) and 848 control units (vs. 600 before). These larger population sizes increased the flexibility to design the sample, by making it easier to approach the desired marginal stratifications and to match control units to selected treatment-sample units.

Note that it is not necessary that the treatment and control populations be mutually exclusive or exhaustive. It would have been permissible, for example, to define the treatment population as those localities having estimated travel times of 0-60 from the nearest program road, and the control population as those localities having estimated travel times 120-360 minutes. There are two problems with doing this. First, we are not really sure how far the significant area of effect of the program intervention is, and it is considered that 60 may be too restrictive. Second, the smaller the control population, the more difficult it becomes to find good matches of control units to treatment-sample units. The major case for restricting the control population is to reduce the geographic dispersion of the control sample units. Unfortunately, because of nature of the statistical matching procedure, the control sample is likely to be spread all over the control-population area, however it is defined, and so the only practical way to control the geographic spread of the control sample is to severely restrict the area of the control population. Since the number of potential control units is not very large to begin with (i.e., the entire locality population, less the treatment population: 1,523 – 675 = 848), this was not considered an acceptable thing to do. Hence, based on these considerations, it was decided to use 0-120 minutes to define the treatment population and 120+ minutes to define the control population. With this definition, the control population contains units that are unlikely to be much affected by the program intervention, and the size of the control population is large (so that matching is facilitated).

A final note is in order concerning the definition of the treatment and control populations. The objective of this project is primarily to construct a “double-difference” estimate of program impact. That estimator is a double difference in means. In addition, it is desired to construct an estimate of the relationship of program impact – price changes – to explanatory variables, such as travel time changes. The sample design constructed for this project enables both objectives to be accomplished. A double-difference estimate will be constructed by comparing treatment and control sample means before and after the program intervention, and an analytical model will be constructed (regression model or tables) to show the relationship of program impact to explanatory variables (such as changes in travel time).

It may seem strange at first glance that the sample contains many localities that are located far from program roads. There are two reasons for this. First, the control-population units are of necessity located far from program roads (to be little affected by the program intervention). Second, it is desired to have considerable spread in the travel times to the nearest program road, in order to facilitate estimation of the relationship of program impact (changes in prices) to changes in travel times. This condition is a necessary one, given the objectives of the project.

The extent to which an analytical survey design can be tailored to the design objectives depends on how large and varied the population is. If the treatment population and control populations are very large and varied, this is relatively easy to do. The smaller that these populations become, the more difficult it becomes to satisfy all of the design objectives. For the present survey, the total number of localities in all regions was 1,928. As mentioned, adequate GIS data could not be obtained for Western Region, which is located far from the MiDA program areas, and so this region was dropped from consideration. The number of localities remaining was 1,719. After restricting the population of interest to localities having population 1,000 or more, the population consisted of 1,523 localities: 675 in the treatment areas (travel time to nearest MiDA program roads less than or equal to 120 minutes) and 848 in the control areas (travel time to nearest MiDA program roads exceeding 120 minutes). This list is the *sample frame*, from which the treatment sample and matching control units are selected.

From this target population, it is desired to select a sample of 308 localities – 154 from the treatment population and 154 from the control population. (In order to allow for the possibility that some localities may not be acceptable, for example, the town was much smaller than indicated by the 2000 population data, or inaccessible because of rain, the sample size was increased to 174 treatment units and 174 control units, to allow for replacement of up to 20 sample units (of each type – because the sample is a "matched-pairs" sample, sample replacements must be made for matched pairs, not for individual sample units).) Since the size of the treatment sample is relatively large with respect to the size of the treatment population (174 vs. 675), the flexibility for using marginal stratification to achieve desired levels of spread and balance is somewhat limited. Since the size of the control sample is smaller with respect to the size of the control population (174 vs. 848), there is greater flexibility for matching the control sample to the treatment sample.

An additional comment is perhaps in order concerning the decision to restrict sampling to all of the country except Western Region. For a descriptive survey intended to produce estimates of means or totals for the entire country, restricting the target population in this manner would not be acceptable – the scope of inference of the sample would be narrowed. In this application, however, the objective is not to produce estimates of national means and totals, but to estimate the parameters of an analytical model of program impact, and of a double-difference estimate. For such an application, it is important that the sample be such that it includes sample units for which the distribution of the design variables is conducive to producing these estimates, and that the scope of the model be appropriate. The sample includes representation of localities that are both close to and far from program roads, and includes a matched sample of similar units. The quality and inferential scope of the model and double-difference estimate is not significantly improved by including control units that are very far away. While the quality of the matches might be slightly better if the control population covered the entire country, all that is really important is that the matching control units be similar with respect to variables that are important with respect to the model (i.e., with respect to program road selection and program impact), viz., with respect to the design (match) variables. (As discussed earlier, in defining the control population, an upper limit could have been imposed on the travel time to the nearest MiDA program road; this was not done, because it would have done relatively little to reduce the geographic dispersion of the sample units and would have substantially reduced the size of the control population (making matching more difficult).)

2.3 Sample Design Variables and GIS Data Preparation and Processing

As described above, GIS data were deemed to be integral to this sample design and to the larger evaluation of MiDA project feeder roads. The reasons for this included:

- Determination of treatment versus control localities using GIS-calculated travel-time accessibility indices was explicitly specified in the original RFP. These indices provide an excellent delineation of treatment from control for infrastructure improvement projects, but GIS spatial data on both improvement and locality locations is essential for their calculation.

- As was verified by the NORC team initial site visit in March, 2009, Ghana government agencies have generated superb, high-quality GIS datasets for the entire country. These include detailed physiographic GIS datasets with complete country-wide coverage. Use of these datasets allows the qualification of sample localities by extensive physiographic and spatial proximity variables that can be used to improve the robustness of the delineation of treatment from comparison communities by providing extensive descriptors to more accurately match treatment to comparison localities. The descriptive variables can be attached to each locality through spatial joining in a GIS, rather than through expensive and time-consuming survey data collection.

Consequently, an extensive GIS database was assembled by the NORC evaluation team for this purpose over a period of several months, through cooperation with MiDA, with a number of Ghana government agencies, and with assistance from local Ghanaian consultants hired for this data collection.

Next, these GIS datasets were evaluated and pre-processed to allow for the identification of a set of descriptive variables that could be calculated for all localities in the Ghana locality GIS database selected for sample observations (described above). The goal was to provide descriptive variables for all localities that described them in at least the following ways:

- provided the population of all localities in the relevant areas – a key variable for stratification and sample selection;

- provided a set of GIS-calculated variables describing each locality's access to major markets, to major infrastructure, and to MiDA treatment road locations. While there are multiple methods in the literature for calculating "access", for this study travel-times were calculated through the GIS road network, considering variation in road quality and topography (see Appendix A for a fuller description of these techniques).

- provided a set of physiographic descriptors of each locality, such as rainfall, temperature, and agricultural biophysical suitability. Since environmental and physiographic conditions are integral in agricultural production, these are very important variables to factor into the sample selection. GIS physiographic databases obtained from the Ghana

Soil Research Institute (SRI), the Ghana Meteorological Department, and CERSGIS were used to qualify all localities by this environmental variation.

As a result of this analysis and pre-processing, the set of variables listed in Table 2.1 were identified as usable for the sample delineation of treatment and control areas. The variable names are those used in the data base in which all of the data are stored.

2.3.1 GIS Data Pre-Processing. The GIS databases assembled from the various Ghana agencies were in different formats (including both raster and vector formats) and were furthermore in a variety of differing geographic and map projection systems. Extensive pre-processing time was spent unifying all layers into a common geographic projection system (Universal Transverse Mercator (UTM), suitable for highly accurate distance calculations at small scales), for the joining of spatial data between layers (for example, joining elevation values to locality positions) and for accurate calculation of surface distances.

In addition, the road network datasets were assembled, re-projected, and then merged into a single unified road network (including feeder and trunk roads), which was then extensively processed and cleaned to allow seamless calculation of GIS travel-time indices (see Appendix A for more details).

2.3.2 Extraction of Variable through Spatial Joining and Calculation of Distance and Travel Time Indices. Physiographic variables were extracted for all localities through spatial joining, whereby the value for the particular physiographic condition (for example agricultural suitability, as a function of soil and climate, or elevation) was calculated by computing an average of the nearby locations (pixels, or vector polygon, line or point objects) that were closest spatially to each locality position.

Distance variables were calculated using standard GIS Euclidean distance calculators. Travel-time indices were calculated using standardized GIS network least-cost path algorithms (such as the Djikstra algorithm), minimizing travel-time through the network as a function of road quality, road speed, number of lanes and topographic variation (see Appendix A for more details).

Table 2.1 List of Variables Derived from GIS Variables Used for Sample Delineation

Demographic Qualifiers:	***Description***	***Data Sources:***
Locality Name	Name of locality	Ghana 2000 Census
Population	Population of Locality from 2000 Ghana Census	Ghana 2000 Census
Male	2000 census male population	Ghana 2000 Census
Female	2000 census female population	Ghana 2000 Census
Population 1984	1984 census population	Ghana 2000 Census
Population 1070	1970 census population	Ghana 2000 Census
Total Households	Total number of households in locality	Ghana 2000 Census
Region Code	two-letter region code	CERSGIS
Region Name	Name of Ghana Region	CERSGIS
District Name	Ghana District name of locality	CERSGIS

Accessibility Variables: Travel-Time and Distance Variables

Accra Travel-Time	travel-time in minutes to Accra through Ghana road network	Ghana Highway Authority (GHA); Dept. of Feeder Roads
MiDA Travel-Time	travel-time in minutes to nearest point on nearest MiDA project road through road network	Ghana Highway Authority (GHA); Dept. of Feeder Roads
Feeder Road Distance	distance in meters to nearest Ghana feeder road (for this variable, distance more robust than travel-time)	Ghana Highway Authority (GHA); Dept. of Feeder Roads
Trunk Road Travel-Time	travel-time in minutes to the nearest Ghana trunk road	Ghana Highway Authority (GHA); Dept. of Feeder Roads
City 10 K Travel-Time	travel-time in minutes to the nearest Ghana town with population greater than 10,000	Ghana Highway Authority (GHA); Dept. of Feeder Roads

Physiographic Qualifiers:

Elevation	elevation in meters	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Slope	code for mean topographic slope of locality (codes: 1 = average slope of 0-2%; 2 = 2-4%; 3 = 4-8%; 4 = 8-16%; 5 = 16-30%; 6 = 30-45%; 7 = > 45%)	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Agricultural Suitability	agricultural suitability index, ranges from 1-100, with 100 having highest agricultural suitability	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Eco-Zone	econzone code (1 = Sudan Savanna Zone; 2 = Guinea Savanna Zone; 3 = Transition Zone; 4 = Deciduous Forest Zone; 5 = Moist Evergreen Zone; 6 = Coastal Savanna Zone; 7 = Wet Evergreen Zone)	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Agricultural Growing Period	Growing Period Pattern Zone Code - ranges from 1-3 with 3 having best growing period conditions, and likely best agricultural conditions (1 = one growing period per year 75% of years, 2 growing periods 25% of years; 2 = one growing period 55% of years, 2 periods 35% of years, 3 periods 10% of years; 3 = 2 growing periods 50% of years, one period 30% of years, 3 periods 20% of years)	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Thermal Zone	Thermal zone: number of days per year hotter than 35 degrees C, higher the code value the hotter (code: 1 = 0 to 5 days per year; 2 = 5 to 30; 3 = 30-90; 4 = 90-150; 5 > 150)	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Relative Humidity	Index of Annual Relative Mean Humidity: higher the index value, the more humid (can have non-linear impact on agricultural productivity, but in general the more humid the better for agriculture)	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Rainfall	Average Annual Rainfall in mm per year	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS
Temperature	average annual temperature in Degrees Celcius	Ghana Survey Department; Ghana Soil Research Institute (SRI); Ghana Meteorological Department; CERSGIS

Sample Design Process. As a first step in the sample design process, an analysis was done to determine the degree of association among the design variables. It was desired to combine or eliminate variables that were highly related. To this end, the variables were coded into quantiles, and the Cramer coefficient of association was calculated for each variable pair. Based on this analysis, it was decided to drop thermal zone and relative humidity from the list, since they were highly correlated with other variables. All of the remaining variables were not highly related, and retained for further analysis.

The next step was to recode the design variables into a small number of categories, to facilitate the process of marginal stratification to achieve a desirable level of spread and balance (orthogonality was not a concern, because, after dropping thermal zone and relative humidity, none of the remaining variables were highly correlated). The categories were defined using "natural" boundaries, as opposed to quantiles, defined in ways that were considered to promote estimation of the relationship of program impact to explanatory variables (design variables and survey variables). The data were re-coded as follows (with closed upper boundaries of internal categories). These re-codings are not used in the text of this report, but they are useful in interpreting the tables. The variables were re-coded as listed in Table 2.2. The short names for the variables used in later tables are in the middle column.

The preceding list identifies all variables used in construction of the sample design. There are two stages to the design construction: (1) marginal stratification to achieve desired levels of spread and balance (but not orthogonality, which has already been taken care of); and (2) statistical matching (of individual units) to make the distributions of the treatment and control samples as similar as possible with respect to the design variables. When the design variables are used for stratification, they may be referred to as "stratification variables"; when they are used for matching, they are usually referred to as "match variables."

Table 2.2 Coding of Variables Used in Matching Treatment and Control Localities

Variable	Short name	Coding
Treatment	TREAT	Not recoded (1 for treatment, zero otherwise)
Control	CONTRL	Not recoded (0 for treatment, 1 otherwise)
Population 2000	POP200	0:1,000; 1:1,000-5,000; 2:5,000-20,000; 3:20,000+
Travel time to Accra	TTACCRA2	0:0-30; 1:30-60; 3:60-120; 4:120+
Travel time MiDA feeder road	TTMIDA2	0:0-15; 1:15-30; 2:30-60; 3:60-120; 4:120+
Elevation	ELEVAT	0:0-500; 1:500-1,000; 2:1,000+
Agricultural suitability	SUITINDEX	0:1-3-; 1:30-60; 2:60+
Feeder distance	FEDDIST	0:0-500; 1:500-2,000; 2:2,000-5,000; 3:5,000+
Rainfall	RAINFALL	0:700-1,000; 1:1,100-1,300; 2:1,400-1,700
Temperature	TEMP	0:23-25; 1:25-27; 2:27+
Trunk road travel time	TRNKTME	0:0-15; 1:15-30; 2:30+
City 10K travel time	10KTME	0:0-15; 1:15-30; 2:30+
Eco Zone	ECOZONE	Not recoded (values 0-7)
Topographic slope	SLOPE	0:0-1; 1:2: 2:3; 3:4-6
Agricultural growing period	GROWPERIOD	Not recoded (values 0-3)
Region code	REGION	Not recoded (values 0-8

2.4 Sample Construction

Table 2.3 (at the end of the text in this section), entitled, "Population Frequencies," shows the number of population units in each stratum category, or "cell." It is clear from this table that the population is highly "skewed" or "unbalanced" with respect to many variables, such as population (pop2000) and travel time to program roads (ttmida2).

Table 2.4, entitled, "Desired Sample Frequencies before Matching," shows the sample sizes that are desired for each stratum cell, for the treatment-unit sample (i.e., before the sample is expanded to include the control-unit sample). The rationale for these allocations is as follows.

1. The total treatment sample size is 174 (154 "target" sample plus 20 replacement units).
2. An equal distribution of the sample was specified for all design variables except region (there was no reason to prefer any particular regional distribution – concern is for "explanatory" variables, and "region" is not one of them).

Once the desired stratum allocation had been specified, the selection probabilities were specified for each unit to make the expected sample frequencies (before matching) close to the desired allocation, subject to the requirement that all treatment units be selected with a minimum nonzero probability. Table 2.4, entitled, "Desired Sample Frequencies before Matching," shows the expected sample frequencies that resulted. Although there is reasonable spread of the sample units over the stratum cells for each of the design variables, the balance is not very close to the desired allocation. The reason why the balance is not very good is that the treatment sample represents a moderately large portion of the total treatment population, and the match variables are not independent, and so the extent to which the sample distribution can differ from that of the population is somewhat limited. For example, for some stratum cells the population size is less than the "desired" stratum allocation.

The fact that the balance is not perfectly even is not a serious concern. In order to facilitate estimation of the relationship of program impact to an explanatory variable, what is important is that the sample units be spread somewhat evenly over the range of variation – perfect balance is not necessary.

A probability sample of treatment units was selected, using the specified selection probabilities. That sample is shown in Table 2.5, "Actual Sample Frequencies before Matching." The actual sample frequencies are quite close to the expected sample frequencies (Table 2), and the comments made about Table 2.4 also apply to Table 2.5.

The next step in the sample design process is to select a matched control unit for each unit in the treatment sample. This is done by a matching algorithm that calculates a similarity measure ("distance" measure) for every treatment-sample-unit / control-population-unit pair, and selecting the pair that has the best measure (this is called "nearest-neighbor" matching). Once a unit in the control population has been paired to a treatment unit, both units are removed from further consideration (this type of matching is called "greedy" matching). The distance measure is a weighted linear combination of the difference between the (recoded) values of each design (match) variable for the treatment unit and the control unit under consideration. Table 2.6, entitled, "Matching Importance Weights," shows the weights. The variables are divided into three categories: demographic variables (population, region), physiographic variables (elevation, suitindex, rainfall, temp, ecozone, slope_code, and grow_perio) and travel-distance measures (travel time or distance) (ttaccra2, ttmida2, dist_feeder2, tt_trunkrd2, tt_city10k2), and each of these three categories is assigned the same total weight. Since the matching and control variables must necessarily differ by travel distance/time to program road (ttmida2) and region, those variables are not included in the matching process. Hence the number of each type of match variable is one demographic variable, seven physiographic variables and four travel-distance variables. The values of the weights for each of these three categories are 1.0 for the demographic variable, .14 for each of the seven physiographic variables, and .25 for each of the four travel-distance variables. (The weights within each category are not required to be equal, but there was no rationale for making them different.)

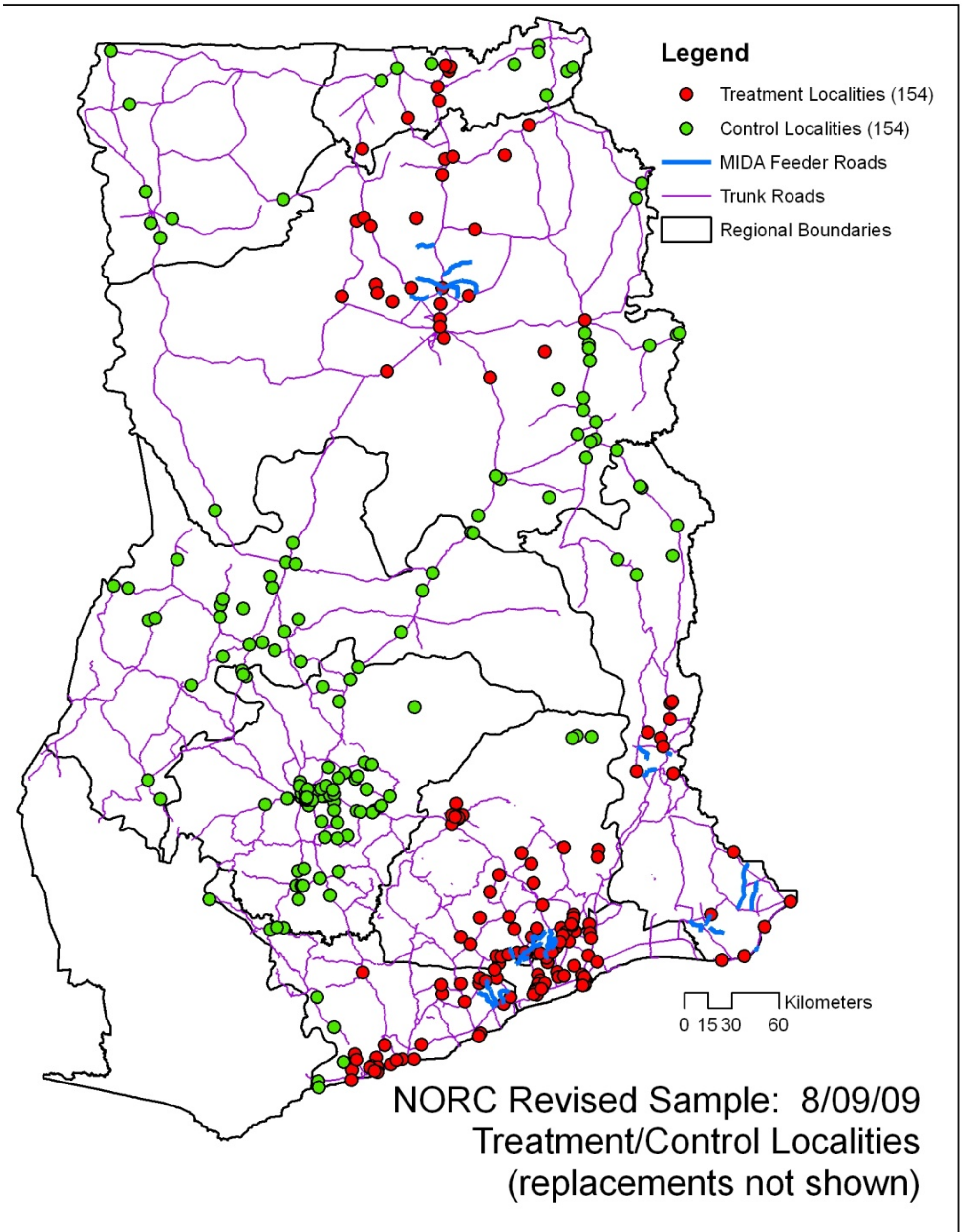
It is important to recognize that the matching process involves the matching of an *individual* control unit to each treatment-sample unit. This matching procedure produces what is called a "matched-pairs" sample. This type of sample leads to much more precise estimates of differences (such as the double-difference estimate of program impact) than matching procedures that simply match the distributions of the treatment and control units. (Note that this matching is *not* so-called propensity-score matching. When properly applied to match treatment and control distributions, propensity-score matching does result in matched distributions. When used to match individual units, however, propensity-score matching can produce absolutely terrible results, which may substantially decrease precision rather than increase it. Since we are matching individual units (to produce a matched-pairs sample), we do not employ propensity-score matching.

2.5 Description of Sample

Table 2.7, entitled, "Actual Sample Frequencies after Matching," shows the sample sizes in each design-stratum cell, after adding the matching control units to the treatment sample.

Figure 2.1, entitled, "Map of Sample Units for Ghana MiDA Feeder Roads Program Evaluation," presents a map showing the locations of the full sample (including both treatment-sample units and their individually matched controls). The map also shows the locations of the program feeder roads.

Figure 2.1: Map of Sample Units for Ghana MiDA Feeder Roads Program Evaluation



The sample localities occur at all distances from the program roads, since it was desired to have substantial variation in the travel time to the program roads (ttmida2) (in order to be able to develop an analytical model of program impact to travel time). What is rather striking about the map is the fact that the control units are spread all over. This is to be expected. The matching process involves matching on a variety of variables, but there is no requirement that the matched control units be located in specific or compact areas. Trying to impose such a requirement would severely hamper our ability to do matching, since the population of match units would be so restricted. In order to do effective matching (which is considered essential to achieve high precision for double-difference estimates), the control population must be as large and unrestricted as possible. It is hence inevitable that it would be "spread out." (The control population might have been restricted by some upper limit on the travel time to program roads, such as 360 or 420 minutes. Such a restriction would have had relatively little effect on the overall geographic distribution of the sample units, and would have the negative effect of making matching more difficult, so it was not imposed.)

In summary, the geographic distribution (spread) of the sample seems very reasonable. The geographic distribution of the sample is high for the control units, but this is unavoidable. Because of the sample design process, the sample has reasonable spread, balance and orthogonality for a large number of design variables. Also, the sample includes a control sample for which the units are individually matched to units in the treatment sample. The sample will be a very good one for use in estimating an analytical model showing the relationship of program impact (price changes) to the Ghana MiDA feeder-road improvements, and for estimating a double-difference estimate of program impact.

Table 2.3 Population Frequencies

Field Definitions:
IDNO: Record identification number
Name: Variable name & _Population
Last 10 columns: stratum code

IDNO	Name	0	1	2	3	4	5	6	7	8	9
1	Total_Population	1719	0	0	0	0	0	0	0	0	0
2	Treatment_Population	1044	675	0	0	0	0	0	0	0	0
3	Control_Population	871	848	0	0	0	0	0	0	0	0
4	POP2000_Population	196	1191	262	70	0	0	0	0	0	0
5	TTMIDA2_Population	151	122	209	284	953	0	0	0	0	0
6	TTACCRA2_Population	29	88	661	941	0	0	0	0	0	0
7	DIST_FEEDER2_Population	443	980	224	72	0	0	0	0	0	0
8	TT_TRUNKRD2_Population	1207	373	139	0	0	0	0	0	0	0
9	TT_CITY10K2_Population	653	500	566	0	0	0	0	0	0	0
10	ELEVATION_Population	647	856	216	0	0	0	0	0	0	0
11	SUITINDEX_Population	520	832	367	0	0	0	0	0	0	0
12	RAINFALL_Population	503	851	365	0	0	0	0	0	0	0
13	TEMP_Population	308	1176	235	0	0	0	0	0	0	0
14	ECOZONE_Population	76	37	430	189	817	7	163	0	0	0
15	SLOPE_CODE_Population	798	285	521	74	41	0	0	0	0	0
16	GROW_PERIO_Population	1	595	722	401	0	0	0	0	0	0
17	Region_Population	338	236	193	250	74	247	99	88	194	0

Table 2.4 Desired Sample Frequencies before Matching (Desired Sample Allocation)

Field Definitions:
IDNO: Record identification number
Name: Variable name & _Sample_Desired
Last 10 columns: stratum code

IDNO	Name	0	1	2	3	4	5	6	7	8	9
1	Total_Sample_Desired	174	0	0	0	0	0	0	0	0	0
2	Treatment_Sample_Desired	0	0	0	0	0	0	0	0	0	0
3	Control_Sample_Desired	0	0	0	0	0	0	0	0	0	0
4	POP2000_Sample_Desired	43	43	44	44	0	0	0	0	0	0
5	TTMIDA2_Sample_Desired	34	35	35	35	35	0	0	0	0	0
6	TTACCRA2_Sample_Desired	43	43	44	44	0	0	0	0	0	0
7	DIST_FEEDER2_Sample_Desired	43	43	44	44	0	0	0	0	0	0
8	TT_TRUNKRD2_Sample_Desired	58	58	58	0	0	0	0	0	0	0
9	TT_CITY10K2_Sample_Desired	58	58	58	0	0	0	0	0	0	0
10	ELEVATION_Sample_Desired	58	58	58	0	0	0	0	0	0	0
11	SUITINDEX_Sample_Desired	58	58	58	0	0	0	0	0	0	0
12	RAINFALL_Sample_Desired	58	58	58	0	0	0	0	0	0	0
13	TEMP_Sample_Desired	58	58	58	0	0	0	0	0	0	0
14	ECOZONE_Sample_Desired	24	25	25	25	25	25	25	0	0	0
15	SLOPE_CODE_Sample_Desired	43	43	44	44	0	0	0	0	0	0
16	GROW_PERIO_Sample_Desired	43	43	44	44	0	0	0	0	0	0
17	Region_Sample_Desired	0	0	0	0	0	0	0	0	0	0

Table 2.5 Actual Sample Frequencies before Matching

Field Definitions:
IDNO: Record identification number
Name: Variable name & _Sample_ActualBeforeMatching
Last 10 columns: stratum code

IDNO	Name	0	1	2	3	4	5	6	7	8	9
1	Total_Sample_ActualBeforeMatching	174	0	0	0	0	0	0	0	0	0
2	Treatment_Sample_ActualBeforeMatching	0	174	0	0	0	0	0	0	0	0
3	Control_Sample_ActualBeforeMatching	174	0	0	0	0	0	0	0	0	0
4	POP2000_Sample_ActualBeforeMatching	0	102	44	28	0	0	0	0	0	0
5	TTMIDA2_Sample_ActualBeforeMatching	41	41	37	55	0	0	0	0	0	0
6	TTACCRA2_Sample_ActualBeforeMatching	27	48	64	35	0	0	0	0	0	0
7	DIST_FEEDER2_Sample_ActualBeforeMatching	36	76	39	23	0	0	0	0	0	0
8	TT_TRUNKRD2_Sample_ActualBeforeMatching	146	18	10	0	0	0	0	0	0	0
9	TT_CITY10K2_Sample_ActualBeforeMatching	110	45	19	0	0	0	0	0	0	0
10	ELEVATION_Sample_ActualBeforeMatching	108	38	28	0	0	0	0	0	0	0
11	SUITINDEX_Sample_ActualBeforeMatching	83	65	26	0	0	0	0	0	0	0
12	RAINFALL_Sample_ActualBeforeMatching	85	62	27	0	0	0	0	0	0	0
13	TEMP_Sample_ActualBeforeMatching	21	123	30	0	0	0	0	0	0	0
14	ECOZONE_Sample_ActualBeforeMatching	23	4	31	2	69	0	45	0	0	0
15	SLOPE_CODE_Sample_ActualBeforeMatching	80	17	33	35	9	0	0	0	0	0
16	GROW_PERIO_Sample_ActualBeforeMatching	1	42	40	91	0	0	0	0	0	0
17	Region_Sample_ActualBeforeMatching	0	0	33	49	39	27	8	0	18	0

Table 2.6 Matching Importance Weights

Field Definitions:
IDNO: Record identification number
Name: Variable name & _MatchingImportanceWeights
Weight: Importance weight (coefficient for variable in distance measure)

IDNO	Name	Weight
1	Total_MatchingImportanceWeights	0
2	Treatment_MatchingImportanceWeights	0
3	Control_MatchingImportanceWeights	0
4	POP2000_MatchingImportanceWeights	1
5	TTMIDA2_MatchingImportanceWeights	0
6	TTACCRA2_MatchingImportanceWeights	0.25
7	DIST_FEEDER2_MatchingImportanceWeights	0.25
8	TT_TRUNKRD2_MatchingImportanceWeights	0.25
9	TT_CITY10K2_MatchingImportanceWeights	0.25
10	ELEVATION_MatchingImportanceWeights	0.142857
11	SUITINDEX_MatchingImportanceWeights	0.142857
12	RAINFALL_MatchingImportanceWeights	0.142857
13	TEMP_MatchingImportanceWeights	0.142857
14	ECOZONE_MatchingImportanceWeights	0.142857
15	SLOPE_CODE_MatchingImportanceWeights	0.142857
16	GROW_PERIO_MatchingImportanceWeights	0.142857
17	Region_MatchingImportanceWeights	0

Table 2.7 Actual Sample Frequencies after Matching

Field Definitions:
IDNO: Record identification number
Name: Variable name & _ActualAfterMatching
Last 10 columns: stratum code

IDNO	Name	0	1	2	3	4	5	6	7	8	9
1	Total_Sample_ActualAfterMatching	348	0	0	0	0	0	0	0	0	0
2	Treatment_Sample_ActualAfterMatching	174	174	0	0	0	0	0	0	0	0
3	Control_Sample_ActualAfterMatching	174	174	0	0	0	0	0	0	0	0
4	POP2000_Sample_ActualAfterMatching	0	204	89	55	0	0	0	0	0	0
5	TTMIDA2_Sample_ActualAfterMatching	41	41	37	55	174	0	0	0	0	0
6	TTACCRA2_Sample_ActualAfterMatching	27	48	121	152	0	0	0	0	0	0
7	DIST_FEEDER2_Sample_ActualAfterMatching	71	169	77	31	0	0	0	0	0	0
8	TT_TRUNKRD2_Sample_ActualAfterMatching	295	35	18	0	0	0	0	0	0	0
9	TT_CITY10K2_Sample_ActualAfterMatching	216	88	44	0	0	0	0	0	0	0
10	ELEVATION_Sample_ActualAfterMatching	139	159	50	0	0	0	0	0	0	0
11	SUITINDEX_Sample_ActualAfterMatching	126	175	47	0	0	0	0	0	0	0
12	RAINFALL_Sample_ActualAfterMatching	106	179	63	0	0	0	0	0	0	0
13	TEMP_Sample_ActualAfterMatching	41	266	41	0	0	0	0	0	0	0
14	ECOZONE_Sample_ActualAfterMatching	26	12	77	37	148	2	46	0	0	0
15	SLOPE_CODE_Sample_ActualAfterMatching	139	63	93	41	12	0	0	0	0	0
16	GROW_PERIO_Sample_ActualAfterMatching	1	106	142	99	0	0	0	0	0	0
17	Region_Sample_ActualAfterMatching	79	29	44	52	39	53	18	8	26	0

3. Conducting the Survey

As noted, this study focuses on how prices of goods sold at local markets (that are transported on improved roads) change over time. It also documents the changes in goods transport tariffs and passenger fares to market places served by the feeder roads.

Towards this end, field work to collect price information was conducted in 308 localities between August 12, 2009 and September 7, 2009. This section outlines the methodology followed, field period preparation, recruitment of field staff, training of field staff, field period review, issues encountered in the field, and recommendations for future phases.

3.1 Methodology

To help increase comparability and continuity, NORC priced a subset of items that were priced in the GLSS-5 and GLSS-4. As shown in Table 3.1, our survey priced 89 items divided between food and non-food items and transportation tariffs for passengers and goods each to three frequent destinations from each sample point. Our goods price survey instrument specifies the quantity of each item to be priced and, where applicable, the specific brand to be priced. For all items and observations, data on unit sizes (both standard and non-standard) were collected. Tariffs were gathered for passenger trips to the most frequent destinations from the sample point and for shipping a 100 kg bag of cement to three frequently goods-shipment destinations.

Table 3.1 Number of Goods and Destinations for which Prices Were Obtained

Type of Item Priced	No. Priced
Goods priced	
>>food items	70
>>non food items	19
Transportation tariffs	
>> passenger destinations	3
>>freight destinations	3

3.1.1 Price Index

General Approach. The goal was to obtain three independent price observations on each item, one from each of three retail outlets in each sample location, as was done in GLSS-4. Interviewers were instructed to go first to relatively formal retail establishments to obtain prices and then to progressively less formal outlets. They continued this process until they obtained three observations for each item. Interviewers recorded the name, location, and contact information of the outlet for each observation of each item as well as the price. We used the following scale to identify stores to take part in the survey:

1. Large retail outlets or supermarkets, followed by;
2. Kiosks of a smaller size, followed by;
3. Individual traders or stalls

In the follow-up survey rounds, interviewers will go to the same outlets to record the prices of the items. If the outlet no longer exists or no longer stocks an item, they will search for another outlet following the procedure outlined above and record the outlet's name along with the item's price. Every effort will be made to ensure that the same outlet is surveyed in each round.

In principle it is important that the surveys be done at the same time each year to avoid seasonal effects. However, due to the combination of the MiDA-established date for the Phase 3 final report and the delays in the Phase 1 field work, this will likely not be possible.

3.1.2 Transport Tariffs and Passenger Fares

General Approach. The first step in measuring tariffs and fares was to identify for each sample location the most frequent destinations for shipped goods and passenger travel. Our interviewers first located the transport hub for each sampled location. After locating the transport hub (possibly two hubs one servicing individuals and one dedicated to transporting goods), interviewers located knowledgeable informants to identify the three most common transport locations for individuals and goods. Knowledgeable informants in most cases were not ticket sellers but rather individuals who worked in close proximity to the transport or tariff hub (most likely workers in small kiosks near the hub) and had a working knowledge of where people and goods were going. After asking for a ranking of locations (1 being the most visited or shipped to place), interviewers identified the three most common responses (based on the ranking) which served as our data collection points. Three independent observations for each location were collected, when possible, with locating information collected as indicated in the discussion of market prices.

Additionally, for passenger travel the interviewers asked how long the trip normally takes and the type of vehicle used. This information on travel time is of interest in its own right as another indicator of road improvement outcomes but also will be used to calibrate the GIS models.

Passenger Fares. The price of a trip by bus or other modes were available locally. If tickets were only sold by the vehicle driver, the interviewers asked drivers about prices and attempted to confirm them with passengers, depending on whether trips were originating while the interviewers were in town. If tickets were sold at a transport hub kiosk, interviewers surveyed kiosk employees after identifying the three most common destinations.

Goods Shipments Tariffs. With respect to obtaining information on the cost of transporting goods from the village to the city, it was important as a first step to identify what good and what quantity should be priced. Interviewers asked about shipping 100 lbs of dry cement, a commonly shipped good with a standard weight. Similar to the steps outlined above, interviewers located knowledgeable informants to rank the most common destinations for transported goods from the sampled location. Three independent observations (where possible) were collected for each identified final destination.

3.2 Preparation

3.2.1 Initial visit. On March 4, 2009 NORC's Ghana Market Survey Field Manager (NORC FM) visited localities in Ghana in order to assess conditions on the ground, sub contractor capabilities, and cultural nuances that may impact data collection. As part of this trip, the NORC FM visited Kasoa, Bawjiase, Swedru, and Nsawam to observe local market conditions and transit hubs.

In Bawjiase, the NORC FM met with local MiDA staff to discuss the Ministry of Food and Agriculture's bi-weekly market data collection efforts in the area. During this meeting the NORC FM reviewed the data collection methodology and survey instruments being used for insights on how to design the current data collection effort.

These field visits proved invaluable in determining the scope of our current study. The NORC FM was able to define the pros and cons of sampling markets vs. localities as well as the differences between conducting data collection on market and non-market days. In addition the NORC FM observed the nuances between large urban markets and smaller local markets which allowed for a careful multifaceted data collection design accounting for the different local realities. The limitations and future recommendations based on the initial field visit will be described in later sections of this report.

3.2.2 Questionnaire Design. During the initial visit the NORC FM identified a subset of items on the GLSS surveys to identify and price in the market. This initial pricing and observation allowed for a detailed understanding of the impediments interviewers may encounter during data collection. After observing local conditions the NORC FM met with his counterparts on the local subcontractor team (Pentax Management and Consulting) to carry out an item by item review of the GLSS survey. Through this review NORC and Pentax were able to refine the GLSS survey to meet the needs of the current study. Standard weights and product types were identified for the majority of products, non important items were deleted in order to reduce the time of the survey, and possible fielding issues were discussed with resolutions identified.

3.2.3 Additional Meetings. During the initial meetings conducted in March of 2009, NORC met with the head of the Institute of Statistical, Social and Economic Research (ISSR) to learn about overall MiDA impact evaluation and identify synergies and overlaps, if any. This meeting also included a discussion of the GLSS5+ data and this data set and full documentation was obtained for it. The NORC FM was able to design survey instruments that are in congruence with the overall evaluation design after these meetings.

3.2.4 Recruitment of Field Staff. The key to the successful execution of a survey is in the quality, commitment and training of the field staff—field interviewers and supervisors. Field interviewers must be drilled to deliver the questions in exactly the way that they were designed and must fully understand the meaning and context of the questions. The uniformity of survey application is best ensured by keeping the field team as small as possible consistent with the time available for the study.

Pentax, with the oversight of NORC, was responsible for advertising, interviewing, and hiring of all interviewing and data entry team members. To the maximum extent possible, Pentax drew on its roster of field interviewers and supervisors with whom it has previously worked in order to ensure the highest level of field staff quality.

We recruited 25 supervisors and interviewers to oversee and conduct the survey tasks within a relatively short four-week data collection period. Although 25 field staff were recruited, only 24 individuals successfully completed training (described more fully below). All field interviewer and supervisor candidates were interviewed by Pentax's Data Collection Field Manager to establish their experience, interpersonal skills, understanding of the basic concepts used in socio-economic surveys, ability to record accurately information on the questionnaires, capability to identify the appropriate people for the interviews, professionalism and neutrality, and capacity to understand the necessity of avoiding directing the respondents replies. Interview staff was also recruited based on knowledge of the local languages used during the survey.

For the Supervisor candidates, their leadership qualities and objectivity were also evaluated. Supervisors were responsible for ensuring that respondents were correctly identified, making certain interviewers comply with all consent and confidentiality requirements as approved by the IRB, and verifying the completeness and internal consistency of the questionnaires before they were returned from the field to the central office for data entry.

Based on the criteria below, the best candidates were invited to participate in the training:

(a) Field Supervisors

1. Approximately three (3) years experience in collecting data on market surveys and/or agriculture-related evaluations, demonstrating excellent interviewer skills;
2. Demonstrated experience in team management;
3. Excellent organization and data management skills;
4. Previous experience in conducting or helping implement training regimes;
5. Ability to review and correct field data;
6. Establish excellent liaison with the Data Collection Team Leader;
7. First University degree in the Social Sciences or Statistics, preferred;
8. Excellent interpersonal skills and demonstrated ability to effectively interact with various stakeholders including government officials, project implementers, project beneficiaries, among others;
9. Ability to speak and understand two or more of the major languages spoken in the three MiDA Intervention Zones; and
10. Proficiency in written and oral English.

(b) Field Interviewers

1. Experience in collecting data on market surveys and or/agriculture-related evaluations, demonstrating excellent interviewer skills;

2. Ability to review and correct field data;

3. Establish excellent liaison with the Data Collection Team Leader;

4. First University degree in the Social Sciences or Statistics, preferred;

5. Excellent interpersonal skills and demonstrated ability to effectively interact with various stakeholders including government officials, project implementers, project beneficiaries, among others;

6. Ability to speak and understand two or more of the major languages spoken in the three MiDA Intervention Zones; and

7. Proficiency in written and oral English preferred.

3.3 Training

Training was conducted from August 3 – August 7, 2009 at Ange Hill Hotel in Accra. The NORC FM and Pentax FM led the training sessions which were attended by 25 interviewers. Of these interviewers, 24 completed the training satisfactorily with 1 interviewer not invited to participate in the full data collection. MiDA staff attended each day of training and supported the NORC/Pentax staff when needed.

In addition to classroom exercises, the interviewers conducted a short pre-test of the survey protocols in Winneba and Swedru. The pre-test was invaluable in cementing field procedures and identifying problem areas.

Additional ad-hoc training was conducted on August 10 and August 11, 2009. Extra training was carried out to take advantage of the two day delay in beginning field work.

During the opening session of training, the MiDA chief statistician gave a detailed over view of MiDA, the current study, as well as the overall evaluation framework. NORC appreciated this overview and felt it helped to set a positive tone for the remainder of the training.

The training focused on the following themes:

- Survey design and methodology
- Sampling and enumeration
- Role playing and interview techniques
- Survey logistics
- The Price Module
- The Tariff Module
- The Transport Module

Additional training was carried out with the supervisors covering the following items in further detail:

- Team management
- Logistics
- Field editing

- Data management processes
- The Tariff Module
- The Transport Module

Moreover, Ms. Abigail Abandoh-Sam (Director of M&E, MiDA) attended several sessions and spoke at length to the interviewers and supervisors on the importance of the current survey. Reiterating the importance of high data quality reinforced for interviewers the significance of the current study. The training program is outlined below.

TRAINING ON "IMPACT EVALUATION OF FEEDER ROADS ON MARKET PRICES"

Date: Monday, August 3rd – Friday, August 7th, 2009 at ANGE HILL HOTEL

DATE / DAY	9:00am – 10:30am	10:30am – 11:00 am	11:30am – 1:00 pm	1:00 pm – 2:00 pm	2:00 pm – 4:00 pm
DAY 1 **Monday, 3rd August,2009**	***Session 1*** • Registration • Introduction to the Project • Overview of the Price Module **MM, AAS & KK**	**SNACK BREAK**	***Session 2*** • Overview of Tariff Module • Into to Interview Techniques **KK & GC**	**LUNCH**	***Session 3*** • Field Controls • Field Management • Training the Interviewers **KK & GC**
DAY 2 **Tuesday, 4th August,2009**	***Session 4*** • Introduction to the current project • Overview of the Price Module **KK& GC**		***Session 5*** • Overview of the Tariff Module • Introduction to Interview Techniques **KK & GC**		***Session 6*** • Field Control **KK & GC**
DAY 3 **Wednesday, 5th August,2009**	***Session 7*** • In-depth Coverage of the Price Module • In-depth Coverage of Tariff Module **KK& GC**		• Interview role playing and Techniques **KK & GC**		• Field Test Preparation • Group Discussion **KK & GC**
DAY 4 **Thursday, 6th August,2009**	**Pilot – Test on the Field**		**Pilot – Test on the Field**		**Pilot – Test on the Field**
DAY 5 **Friday, 7th August,2009**	**De-Briefing & Assignment** **KK & GC**		**De-Briefing & Assignment** **KK & GC**		**De-Briefing & Assignment** **KK & GC**

* MM: Dr. Mohammed Muslim (COO, Pentax Management Consultancy Services Limited)
* AAS: Abigail Abandoh-Sam – (Director of M&E, MiDA) *KK – Kareem Kysia (Survey Expert, NORC) * GC – Godfrey Crentsil (Data Collection Team Leader, Pentax

3.4 Field Period

Field work was carried out from August 12, 2009 through September 7, 2009 in nine Ghanaian regions: Upper East, Upper West, Northern, Brong-Ahafo, Ashanti, Volta, Central, Eastern, and Greater Accra. Our sample was distributed across regions with an equal amount of treatment and control groups composing a total of 308 localities.

3.4.1 Field Teams. There were 4 field teams composed of 16 interviewers, 4 associate field supervisors, and 4 field supervisors for a total of 24 field staff. The following table outlines the staffing for the study per team:

Team 1 – Brong-Ahafo & Ashanti Region	
Name	**Position**
Justice Kojo Antwi	Supervisor
Justice Maddy	Associate Supervisor
David Kojo Rockson	Interviewer
Rejoice Bawa	Interviewer
Maame Esi Aikins	Interviewer
Rita Adjei	Interviewer
Team 2 – Northern, Upper East, & Upper West Region	
Name	**Position**
John Kwosi Emmanuel	Supervisor
Bukari Drahamani	Associate Supervisor
Abubakari Mugisu	Interviewer
Shaibu Issifu	Interviewer
Johnson A. Etipana	Interviewer
Daniel Salifu	Interviewer
Team 3 – Volta & Eastern Region	
Name	**Position**
Jennifer Acquah-Mensah	Supervisor
Maxwell Akannah	Associate Supervisor
Eunice Torto	Interviewer
Clement Osei Manu	Interviewer
Christopher Odoi	Interviewer
Pertual Joan Som	Interviewer
Team 4 – Greater Accra & Central Region	
Name	**Position**
Akua Sekyibea Akuffo	Supervisor
Naku Adama	Associate Supervisor
Rebecca Ama Larbi	Interviewer
Jones Eric Amo	Interviewer
Andrew Botchway	Interviewer
Anita Asante	Interviewer

The field teams remained stable over the data collection period with no interviewers or supervisors needing to be replaced. Each field team was comprised of at least two members who spoke the necessary languages to conduct the field work in a satisfactory manner.

A total of 308 localities were surveyed with the following breakdown per field team:

Team	Number of Localities
1	95
2	72
3	71
4	70

3.4.2 Logistics. Field members were given a housing, food, and travel stipend to cover costs during the data collection period. In situations where the travel stipend was insufficient additional monies were distributed. (Recommendations for improved travel arrangements will be discussed in following sections.) In addition, each field member was issued with an ID card identifying them as MiDA certified field interviewers as well as introductory letters for district officials.

Field teams were also provided with detailed regional maps with each locality identified with a unique ID which tied into their field data control sheets. Field teams were given an opportunity to comment on their assignments in preparation for the data collection period and develop a logistical plan for carrying out the data collection.

Field supervisors consulted their maps and their data control sheets before handing out assignments for each day of data collection. These assignments were made the evening before each day of data collection. Each interviewer and supervisor used local transportation to reach particular localities. Most teams divided into sub-teams of three persons each (a supervisor or associate supervisor accompanying each team) in order to cover the maximum amount of localities per day without impacting data quality.

At the end of each day of data collection the teams came together in a central location (usually the hotels they were staying at) to discuss lessons learned from the day and to develop a plan for the next day of data collection.

Supervisors collected completed questionnaires and receipted them in their data control sheets. Each evening the supervisors were contacted by central office staff who enumerated the completed interviews and localities on central office data control forms for reporting to the NORC FM and MiDA.

Each survey module was identified with a unique locality ID to ensure that each survey could be connected to the appropriate location. Completed and edited questionnaires were placed within a manila envelope corresponding to each locality visited. Each supervisor ensured the safe oversight of envelopes which were delivered to Pentax for data entry at the end of the field period.

3.4.3 Data Collection. A detailed description of our methodological approach can be found in earlier sections of this report. However, it is important to point out a few items that arose during the actual fielding of the surveys.

The methodology used during data collection differed in interesting ways dependent on the size of the locality being surveyed. In small localities, when first entering a village, the interviewers often requested a meeting or introduction with village leaders. They would then introduce themselves and the purpose of the study. This proved important in gaining the cooperation of retailers who wanted assurances that the proper village leaders had been consulted. Moreover, interviewers in small localities recruited a knowledgeable informant to help locate retail locations within the village. This informant would ensure that small retailers, often working out of their own homes, were surveyed if needed. These introductions and informants were of less value in large and less personalized market locations.

In addition, in smaller localities it was difficult to find many of the produce items listed on the questionnaire. This finding was expected since smaller localities often relied on home grown produce and a barter system. Interviewers were reminded that they only had to fill out survey items based on what was actually available in particular markets.

Larger localities also had distinct characteristics. Oftentimes the survey would take considerably longer in large localities with many stalls. Time was spent ensuring that the methodology was followed and the largest retailers were identified first.

3.4.4 Oversight. During the first week of data collection, the NORC FM, Pentax FM, and MiDA supervisor visited each field team. These visits proved invaluable in identifying field issues early and instituting fixes to ensure the highest quality data collection. Following the first week, the Pentax FM was in continuous contact with the field teams as well as the NORC FM. All identified issues were handled appropriately to ensure high quality and a successful conclusion to the data collection period.

These meetings were of particular importance for Team 2 which visited the most distant and difficult to reach localities in the Upper East and Upper West regions. Team 2 met with the director of the MiDA Office in the Northern Agricultural Zone in order to receive local information on the area and assistance in defining a detailed work plan.

Of the 308 sampled localities only one locality was removed from the sample because we were unable to locate it. This locality, Choo #0155, was not located and was removed along with its matching pair, Sabiye #0159. These localities were replaced with Suame #0812 and Ogbodzo #1264. All other localities were located and surveyed.

3.5 Challenges & Adjustments

Data collection was carried out with highest possible quality and was completed largely successfully. However, as with any large scale research project, there were challenges identified in the data collection. Below we discuss those identified during the data collection period and then those that became evident when the data set was examined carefully. Where appropriate, we note recommendations to solve these challenges and ensure successful data collection in future phases.

3.5.1 Issues Identified in the Field Work Phase

3.5.1.1 Transportation. Although NORC made a best faith effort to secure 4-wheel drive vehicles for use during the field period, in the end field staff had to use local transportation to reach localities. Relying on local transportation had the largest impact on Teams 2 and 3.

The NORC/Pentax team provided Team 2 with transportation to the two farthest localities in the Upper West region during the field review using 4-wheel drive vehicles. This helped in the successful enumeration of those localities. Unfortunately, the original travel stipend for Team 2 was insufficient due to the fact that the localities in the Upper East, Upper West, and Northern region were widely spread out.

Team 3 encountered difficulties in reaching two localities during the field period. Both these localities were underserved by local transportation and impacted by the rainy season.

For Phase 2 & 3 NORC and Pentax will develop a detailed transportation plan well in advance of the field work that states clearly how the interview teams will be transported to all localities. The Plan should be provided sufficiently early to MiDA and NORC for review and approval. In identified situations vehicles will be provided as needed. In addition, field staff will be provided with sufficient travel funds to carry out the data collection without delay.

3.5.1.2 ID Cards, Brochures. Although field staff were provided with ID cards identifying them as MiDA enumerators, these ID cards were not of a high enough quality to ensure a professional appearance and mitigate respondent refusal.

For Phase 2 & 3, NORC and Pentax will design high quality ID badges with identifying photographs of each field staff.

In addition, MiDA brochures identifying the current work being conducted in Ghana proved to be invaluable in gaining respondent cooperation. Unfortunately, MiDA was unable to provide the NORC/Pentax team with a large volume of brochures.

For Phase 2 & 3, NORC and Pentax will work closely with MiDA to ensure a sufficient quantity of brochures.

3.5.1.3 Locating Sheets. Locating respondents in subsequent phases may turn out to be one of the major challenges we face. A subset of market stalls lacked detailed and identifiable locating information. In addition, after data collection began, we discovered that many respondents became suspicious when interviewers asked for identifying information. Interviewers were instructed to stop requesting locating information if the respondent became distrustful.

The extant of this problem is not yet known, however our methodological approach of finding particular stalls by decreasing size should mitigate the impact of poor locating information. In addition, the issue is most pronounced among small traders without a set stall. The analytic power of the current evaluation should not be impacted if these small traders are replaced with traders of an equal size in Phase 2.

NORC plans to spend an extended amount of time during the Phase 2 supplementary training on describing how to relocate particular traders and how to overcome locating issues in the field.

3.5.1.4 Data editing. During the first week of data collection the NORC FM and MiDA supervisor observed two data quality issues. The first was the use of transcriptions among a small number of interviewers. Interviewers were writing down survey responses in notebooks and then transcribing those results onto the survey several hours later.

The NORC FM stressed the problematic nature of this practice and ensured that it would not be repeated. During supplementary training NORC plans to stress the problems with transcribing answers which should mitigate the possibility of this problem recurring.

A second issue dealt with the late data editing and review by supervisors. In a subset of cases, supervisors were waiting until they had left a particular village before conducting data editing tasks. Late data editing more or less defeats the purpose of data editing in the first place. If supervisors have already left a particular locality then they cannot return to correct possible data entry errors.

After identifying this issue early in the field period, the NORC FM spoke with each supervisor and stressed the problematic nature of this practice. During supplementary training sessions, NORC will stress the importance of data editing and how to conduct it properly.

3.5.1.5 Standardized Weights. The description of sizes or weights of food items like corn dough, cassava dough, and groundnut paste, is ambiguous and therefore does not provide fair basis for comparison.

We believed this issue to be quite limited owing to three factors. First, few products are sold in quantities as unspecific as "bunches." Our initial count was that less than 10 percent of all items fall into this class. Our view was that after data entry we would have a stronger understanding of exactly how many items are measured using standards more prone to measurement error. Measures such as baskets or olonka were believed to be quite standard within a locality and will remain so over time.

Second, observation indicated that there is strong uniformity within a market for a bunch at a point in time—competition generally is sufficient to produce this outcome. Third, in the double difference calculations, a bunch price in a locality is contrasted with the price of the same bunch in the same locality, not another locality.

The issue arises *IF* the quantity (size) of the bunch changes over time. One could imagine that if prices rose, vendors could lower the size of the bunch to keep the bunch affordable to more customers. In this case, the bunch price would overstate the price reduction due to the feeder road improvement.

Before examining the data from Round 1 (described below), we planned to weigh bunches in phases 2 and 3. It was not clear if we could use this information in conjunction with information from round 1. Weighing in phases 2 and 3 would, however, at a minimum have informed us of the price changes per kilo between those two observations and by implication over time more generally.

3.5.2 Issues Identified in the Data Set. Examination of the data files containing the price information revealed issues that were not previously evident.

3.5.2.1 Heterogeneity in Items Priced. Two issues of this type were encountered. First, for fresh food items it was found that many different measures were priced, e.g., an olonka, standard bowl, and margarine tin were all used to measure equivalent items.. Because each measurement instrument has a different weight the prices are not comparable. Second, for other food items and for non food items the protocol had called for interviewers to obtain prices for a certain item, e.g., a package of cigarettes. When the descriptions of what was priced were studied it was found that interviewers priced different versions of the good, e.g., 1 cigarette, 3 cigarettes, as well as packages. Because of economies in per unit packaging costs and in the shopkeeper's time associated with the sale, smaller quantities command higher prices. Therefore, these are not the same good.

We were able to address the first issue, different quantities of food stuffs, by sending out a team after the initial data review to weigh the amount of each food item in each of the containers where it had been priced. The product weights, net of the container's weights, were then applied to each observation and a consistent price per kilo was obtained for each observation.

The second issue, different versions of the same good priced, is a greater problem. Prices for goods that did not meet the basic description of the good had to be treated as missing values, e.g., 1 cigarette could not be multiplied by 20 to get the price of a packet. Staff has gone through the descriptions of each item listed by the interviewers and separated them into separate lists of acceptable and non acceptable versions of the items. These lists will be used during training to ensure that interviewers collect the highest quality data in subsequent rounds.

The issue of different quantities of a good is particularly severe for four oils: coconut oil, groundnut oil, palm kernel oil, and palm oil. Contrary to our understanding when the survey was being designed, each of these is sold in a variety of containers. The survey forms where price information is recorded lists, for example: beer bottle, coke bottle, Frytol bottle, schnapps bottle, liter, and gallon. A fundamental problem is that some of these bottles come in several sizes. A "Frytol bottle," for example, comes in three sizes. The survey forms do not record the size bottle for each price. As a result the four oils were dropped from the list of items being priced. The only compensating factor is that the oils were available only in a very small number of localities. In total six items had to be dropped.[1]

The team sought to avoid cases where there would be no price observation for a good in a locality because the good priced was invalid. In such cases we would not have the ability to compute the change-in-price metric essential for the evaluation. We have imputed one price observation for each of these cases.

The imputation procedure involves the estimation of a separate regression model for each good for which imputed values are needed. The observations used in the estimation are all valid observations for the good. The independent variables are factors expected to influence the price of the good in different locations. These include the size of the locality, the distance to the closest town with a population of over 10,000, and the accessibility of the locality measured in different ways. Estimates for the same type of model are presented in the next section and the rationale for including each of these variables discussed there.

The estimated models are presented in Annex C. In reviewing them it is important to bear in mind that the objective in this case is to maximize the explanatory power of the models and therefore the quality of the price predictions made using them. They are not being employed to test hypotheses about what determines the prices observed at different locations. Models with that objective are presented are presented in the next section.

[1] The other two are local honey and Rice (imported) Texas long grain medium size bag

Altogether values were imputed for 1,628 price observations. The entries in Table 3.2 summarize the distribution of items with imputed values. The table shows, for example, those prices of 35.8 percent of items (29 out of 82) were imputed for 1-10 localities; 15 percent of items had prices imputed in over 50 localities. The imputed values have been incorporated into the baseline prices data set that is described below

Table 3.2 Percent Distribution of Imputed Prices by Number of Localities

No. of Localities	Percent Distribution of Items with Imputed Prices
0	25.9
1-10	35.8
11-50	23.5
More than 50	14.8

3.5.2.2 Missing Observations. Interviewers were unable to obtain three observations for a good or transportation tariffs in many cases. This is understandable because many of the localities in the survey are small villages, i.e., places with as few as 1,000 inhabitants. Data on this point are presented below.

3.6 Summary of Data Set Quality

Table 3.3 presents an overview of the final data set in terms of the issues noted above and its overall quality. For each good or tariff priced the table provides two types of information. The first set of data columns shows the distribution of observations among three possible categories. We sought 3 observations in each of 308 localities or 924 observations. The table shows the percentage distribution among three categories:

- valid observations,
- invalid--meaning the interviewer priced the wrong version of the item or a version that could not be properly converted to a standard price, and
- missing--meaning that the good was not sold in the locality or not sold at three locations.

The second set of data columns organizes the information by locality showing the number (out of 308) that had only missing values, i.e., the good was not sold in a locality, there were only invalid observations, and the number of localities that have at least one valid price observation after some values were imputed. The number of "with invalid only" localities corresponds to the number of localities for which the price of the item was imputed. The figures in the first and third columns sum to 308, i.e., the number of sample localities.

Table 3.3 Summary of the Survey's Coverage by Item

	Observations (percent distribution[a])			Number of localities		
Fresh food items	Valid	Invalid	Missing	with missing only	with invalid only[b]	At least one valid obs after imputation
Guinea corn/sorghum	18.8	0.5	80.6	215	2	93
Maize	34.0	3.8	62.2	129	14	179
Millet	22.9	0.8	76.3	195	4	113
Rice (Local)	43.3	0.6	56.1	113	3	195
Gari	70.6	3.5	26.0	18	13	290
Kolanuts	8.3	6.8	84.8	231	33	77
Ginger	43.6	10.6	45.8	87	53	221
Flour (wheat)	13.4	0.4	86.1	234	3	74
Maize ground/corn dough	35.0	3.4	61.7	142	14	166
Cocoyam	26.3	0.6	73.1	179	4	129
Beef	14.4	15.8	69.8	170	73	138
Goat meat	15.3	2.3	82.5	219	15	89
Mutton	13.0	2.4	84.6	232	15	76
Chicken	10.8	2.4	86.8	231	16	77
Red fish (fresh)	6.5	6.4	87.1	237	34	71
Fish (smoked tuna)	23.2	0.0	76.8	193	0	115
Fish (fried)	14.5	0.6	84.8	217	3	91
Cassava	44.6	0.0	55.4	116	0	192
Plantain	47.8	0.0	52.2	112	0	196
Puna yam	46.6	0.0	53.4	96	0	212
Groundnuts (raw)	55.5	1.3	43.2	53	7	255
Groundnuts (paste)	21.9	34.2	43.9	76	143	232
Chicken eggs	55.7	1.3	43.0	82	5	226
Red beans	40.0	1.3	58.7	110	5	198
Cassava – dough	35.3	0.6	64.1	152	3	156
Coconut	11.0	0.0	89.0	247	0	61
Banana	19.3	0.3	80.4	188	2	120
Oranges/tangerines	17.3	1.3	81.4	204	10	104
Pineapple	12.7	0.0	87.3	242	0	66
Mango	7.9	0.0	92.1	271	0	37
Avocado pear	17.7	0.5	81.7	220	3	88
Cocoyam leaves	37.0	0.0	63.0	159	0	149
Garden eggs	52.3	10.0	37.8	81	40	227
Okra	52.4	9.1	38.5	56	44	252
Carrots	17.9	0.0	82.1	219	0	89
Pepper	68.4	2.3	29.3	44	8	264
Onions	59.4	9.6	31.0	47	46	261
Fresh tomatoes	56.6	11.9	31.5	49	53	259
Palm nuts	37.7	0.6	61.7	141	4	167

Packaged food items	Observations			Number of localities		
	Valid	Invalid	Missing	with missing only	with invalid only	At least one valid obs after imputation
Sugar (granulated)	51.8	23.1	25.1	35	97	273
Ice cream (vanilla fan ice - small, sachet, large)	10.9	1.5	87.6	234	7	74
Chocolate (Medium, small-sized chocolate)	18.6	0.9	80.5	213	7	95
Salt (small sachet, sachet)	49.8	23.1	27.2	38	83	270
Biscuits (Digestive - medium)	21.3	2.7	76.0	194	11	114
Fish (Titus canned sardines)	73.9	0.0	26.1	43	0	265
Corned beef (Exeter - large)	12.9	11.7	75.4	201	57	107
Milk powder (small sachet, sachet)	43.6	25.1	31.3	53	97	255
Baby milk (lactogen - medium)	33.1	0.6	66.2	156	4	152
Margarine (Medium)	39.7	8.1	52.2	113	32	195
Tinned milk (unsweetened)	46.8	0.0	53.2	138	0	170
Tinned milk (evaporated)	82.5	0.3	17.2	18	1	290
Tomato puree (salsa canned)	61.0	11.7	27.3	54	43	254
Bread (big, medium, sugar loaf)	44.2	0.6	55.2	104	3	204
Coffee (Nescafe)	23.9	23.9	52.2	110	91	198
Chocolate drinks (Milo-medium)	19.6	25.2	55.2	130	106	178
Lipton tea (25 bags)	74.9	4.7	20.5	15	19	293
Soft drinks (coke or fanta)	76.7	0.0	23.3	23	0	285
Malt drinks (Malta Guinness)	75.1	0.0	24.9	26	0	282
Mineral Water (Voltic bottled large)	61.1	0.5	38.3	77	4	231
Akpeteshie (Beer bottle)	46.6	0.4	52.9	86	3	222
Palm wine/Raffia palm wine (Beer bottle)	5.2	0.0	94.8	275	0	33
Pito/Brukutu (beer bottle)	5.2	0.0	94.8	274	0	34
Beer (Star lager)	50.4	0.0	49.6	74	0	234

	Observations			Number of localities		
Non-food items	Valid	Invalid	Missing	with missing only	with invalid only	At least one valid obs after imputation
Cigarette (1pack)	38.7	6.3	55.0	93	27	215
Charcoal (small bunch)	32.4	5.0	62.7	131	26	177
Key soap (by the bar)	73.5	0.8	25.8	25	4	283
Lux (1 bar)	46.2	21.8	32.0	46	90	262
Dettol (medium)	49.0	1.8	49.1	99	8	209
Insecticides (1 packet)	75.2	4.3	20.5	11	17	297
Matches (1 small box)	65.4	18.0	16.7	10	61	298
Toilet papers (1 roll)	73.9	0.3	25.8	43	1	265
Candles (1 stick)	76.1	1.3	22.6	23	4	285
Pain killers (1 sachet)	38.9	5.3	55.8	62	35	246
Anti malaria medicines (malafin/4 piece sachet)	40.0	0.8	59.2	73	4	235
Condoms (champion 3 pack)	39.6	0.0	60.4	82	0	226
Petrol (1 liter)	20.0	0.0	80.0	185	0	123
Diesel (1 liter)	17.4	0.0	82.6	198	0	110
Exercise books (small exercise book)	58.0	0.1	41.9	50	1	258
Mesh/wigs	19.6	0.1	80.3	188	1	120
Toothpaste (pepsodent large)	69.0	3.2	27.7	37	12	271
Razor blades (5 blade sachet)	72.6	0.0	27.4	38	0	270
Sure deodorant	22.5	0.0	77.5	192	0	116
Transportation tariffs						
Passenger						
Destination 1	68.6	--	31.4	19	--	289
Destination 2	64.8	--	35.2	36	--	272
Destination 3	58.4	--	41.6	62	--	246
Freight					--	
Destination 1	65.5	--	34.5	27	--	281
Destination 2	61.8	--	38.2	44	--	264
Destination 3	55.7	--	44.3	68	--	240

a. Entries in each row add to 100.
b. Corresponds to the number of localities for which the price of the item was imputed.

The final panel of the table 3.3 presents the data for transportation tariffs. There are no invalid cases for transportation tariffs. There are, however, a small number of localities with only missing price observations, i.e., there is no regularly scheduled transport service for these places. As expected, the number of localities without service increases for the second and third destinations, indicating that for some localities there is service only to one or two destinations at which the passenger must transfer to proceed to other destinations.

The results for goods priced are summarized in Table 3.4 which presents the distribution of results (e.g., valid, missing) among the sample localities. The two panels show the changes in distributions attributable to imputing some prices. In Panel A, for example, 28 percent of items had less than 20 percent valid price observations (rather than missing or invalid), 20.7 percent of items had 21-40 percent of its observations valid, and so on up to only 1.2 percent of items had over 80 percent of their observations as valid. Just over 40 percent of all price observations are valid.

The entries in Panel B show that invalid responses of the type described earlier were not common: over 90 percent of items had 20 percent or less of their observations invalid. Overall, 4.5 percent of observations are invalid. With the lessons learned in this round about the need to carefully supervise the specification of each item priced by the interviewers this rate should fall in subsequent rounds. Note that the invalid responses remaining after imputations (Panel A) arise when there was more than one invalid observation for an item in a locality.

The share of price observations missing in the data is large—it constitutes 57 percent of all price observations.

Table 3.4 Percentage Distribution of Price Observations by Result Type for Goods[a]

(Columns sum to 100 percent)

A. After Imputations

Percent of Observations	Price Observation Result Type		
	Valid %	**Invalid %**	**Missing %**
<=20 percent	28.0	100.0	2.4
21-40	20.7	0	26.8
41-60	28.0	0	25.6
61-80	22.0	0	19.5
>80	1.2	0	25.6
Mean percent	41.0	2.4	56.5

B. Before Imputations

Percent of Observations	Price Observation Result Type		
	Valid %	**Invalid %**	**Missing %**
<=20 percent	30.9	91.4	2.5
21-40	24.7	8.6	25.9
41-60	25.9	0	25.9
61-80	17.3	0	19.8
>80	1.2	0	25.9
Mean percent	38.6	4.5	56.9

a. Excludes transportation tariffs.

3.7 Double Difference Analysis

As noted in Section 2, the evaluation will compute the double difference in the price of a good between a treatment and control locality and between two point in time.

$$DDi = (Ptj,z2,i - Ptj,z1,i) – (Pcj,z2,I - Pcj,z1,i)$$

where

DD =	double difference
P=	price observation; average based on up to three observations
i =	good i
tj=	treatment locality j
tc=	matched control locality j
z2=	time 2, i.e., a price observation at time 2.

In the double difference analysis for groups of goods or transportation services, it is possible to include every observation on a good or service for which DD can be computed. The analysis will be for the four groups of goods and transportation services shown in Table 3.3 and for all goods and services combined.

As a concrete example, we can consider packaged food items. Twenty-one such items are included in the list of goods priced. If on average DD can be computed for half of the included localities, then the sample size is 1,617 (21 x 77). The sample sizes will be sufficiently large to indentify significant differences at reasonable confidence levels.

4. Findings

This section reviews the baseline price data gathered from two perspectives. The first examines the variation in prices and tariffs observed in localities and between the treatment and control localities. The second analyzes variation in prices and tariffs associated with localities' size and decree of accessibility to roads and other markets. The objective here is to understand if prices at baseline vary in ways that one would anticipate. When data from the second and third data collection rounds are available, double difference price computations of the type outlined earlier will be undertaken.

4.1 Variations in Observed Prices and Tariffs

Full information on the price information obtained is presented in Table 4.1, which shows for all locations combined and separately for treatment and control localities the mean price in Ghanaian cedis, the coefficient of variation (COV),[2] and the percentage distribution of *normalized* prices. Normalized prices were computed as the mean observed price of a good in a locality divided by the mean price of the good observed in all localities.[3] The table consists of prices in four groups: fresh food items, packaged food items, non-food items and transportation tariffs.

The data in the table for transportation tariffs are included for completeness. Because the destinations priced differ from locality to locality, the prices are not comparable across locations. (The double difference measure of price change remains valid, however.)

There are a number of clear differences in the extent of variation among localities, and these variations are consistent for treatment and control locations. Among fresh food items, for example, local rice (item A.4) has a low COV while kolanuts (A.6) and ginger (A.7) have comparatively higher price variance. Among packaged foods, corned beef (B.7) and tinned milk by Ideal (B.11) have low variance while milk powder (B.8) is much higher. Many of the lowest variances are observed for non food items; among these are those for Key soap (C.3), Lux soap (C.4), Dettol (C.5), petrol and diesel (C.13 and 14), and Pepsodent toothpaste (C.17). It may be that distributors have pricing agreements with retailers for such items.

[2] The coefficient of variation (COV) is the standard deviation of the variable divided by its mean. The computation yields a measure of variance independent of the absolute variance.
[3] Tests for statistically significant differences between the price values for treatment and control areas have not been calculated because the distributions underlying the price observations do not conform to the "normal distribution." The treatment sample was designed to provide high precision for the double-difference estimates. For this sample the selection probabilities are highly variables, although they are all non zero. Additionally, the control group was selected by matching localities from the control population to the treatment sample units. This was done locality-by-locality so that a "matched pairs" sample was obtained. As a consequence one does not know the selection probabilities of the control sample at all. In short, while the sample is very good for the ultimate goal of estimating project impact and for estimating analytic models, it is not suited for estimating overall characteristics of the baseline population using just the sample data.

Table 4.1 Mean and Distribution of Observed Prices by Item and Treatment versus Control Localities

Fresh Foods	Mean[a]	COV[b]	Distribution of Normalized Values[c]				N
			0-.7 (%)	.71-1.0 (%)	1.1-1.3 (%)	>1.3 (%)	
A.1 Guinea corn/sorghum (per kg)							
All locations	0.83	59.1	39.8	26.9	18.3	15.1	93
Treatments	0.84	47.3	29.8	27.7	23.4	19.1	47
Controls	0.80	70.5	50.0	26.1	13.0	10.9	46
A.2 Maize (per kg)							
All locations	0.55	32.2	11.7	35.8	35.2	17.3	179
Treatments	0.59	30.9	8.3	28.1	37.5	26.0	96
Controls	0.50	32.0	15.7	44.6	32.5	7.2	83
A. 3 Millet (per kg)							
All locations	0.91	30.9	8.0	54.0	28.3	9.7	113
Treatments	0.92	23.2	3.2	54.0	34.9	7.9	63
Controls	0.89	39.1	14.0	54.0	20.0	12.0	50
A. 4 Rice (local – per kg)							
All locations	1.20	22.4	9.1	39.6	42.6	8.6	197
Treatments	1.20	26.3	15.4	32.1	42.3	10.3	78
Controls	1.19	19.4	5.0	44.5	42.9	7.6	119
A. 5 Gari (per kg)							
All locations	0.88	41.1	12.0	53.1	27.1	7.9	292
Treatments	0.94	43.3	4.1	54.1	31.1	10.8	148
Controls	0.82	36.5	20.1	52.1	22.9	4.9	144
A. 6 Kolanuts (per kg)							
All locations	6.30	62.4	35.1	23.4	7.8	33.8	77
Treatments	7.30	61.9	32.1	17.9	3.6	46.4	28
Controls	5.70	60.4	36.7	26.5	10.2	26.5	49
A. 7 Ginger (per piece)							
All locations	0.20	66.0	45.7	23.5	0.5	30.3	221
Treatments	0.18	67.6	51.6	27.4	.8	20.2	124
Controls	0.23	61.2	38.1	11.7	0	43.3	97
A. 8 Flour – wheat (per kg)							
All locations	1.96	32.9	16.2	37.8	17.6	28.4	74
Treatments	2. 08	26.5	7.7	35.9	23.1	33.3	39
Controls	1.82	39.4	25.7	40.0	11.4	22.9	35
A. 9 Maize ground/corn dough (per kg)							
All locations	0.81	73.8	50.0	12.0	9.6	28.3	166
Treatments	0.86	77.6	48.0	9.2	9.2	33.7	98
Controls	0.74	65.2	52.9	16.2	10.3	20.6	68
A. 10 Cocoyam (per kg)							
All locations	0.76	42.2	31.8	5.4	48.8	14.0	129
Treatments	0.85	44.0	23.4	3.1	51.6	21.9	64
Controls	0.67	34.1	40.0	7.7	46.2	6.2	65
A. 11 Beef (per kg)							
All locations	5.32	22.1	9.4	25.9	63.3	1.4	139
Treatments	5.01	25.6	10.8	44.6	43.2	1.4	74
Controls	5.66	16.5	7.7	4.6	86.2	1.5	65
A. 12 Goat meat (per kg)							
All locations	5.72	24.3	10.1	34.8	49.4	5.6	89
Treatments	5.53	19.4	6.5	47.8	43.5	2.2	46
Controls	5.89	28.0	14.0	20.9	55.8	9.3	43
A. 13 Mutton (per kg)							
All locations	5.70	20.9	9.2	28.9	60.5	1.3	76
Treatments	5.44	20.4	9.1	43.2	47.7	0	44

Fresh Foods	Mean[a]	COV[b]	Distribution of Normalized Values[c]				N
			0-.7 (%)	.71-1.0 (%)	1.1-1.3 (%)	>1.3 (%)	
Controls	6.05	20.4	9.4	9.4	78.1	3.1	32
A. 14 Chicken-fresh/high quality (per kg)							
All locations	4.36	37.9	15.6	48.1	19.5	16.9	77
Treatments	3.88	34.8	16.7	62.5	14.6	6.3	48
Controls	5.14	35.5	13.8	24.1	27.6	34.5	29
A. 15 Red fish – fresh (per kg)							
All locations	4.10	22.2	6.8	38.4	50.7	4.1	73
Treatments	4.40	23.9	8.3	19.4	63.9	8.3	36
Controls	3.82	16.8	5.4	56.8	37.8	0	37
A. 16 Fish - smoked tuna (per piece)							
All locations	2.00	55.3	33.6	23.3	19.0	24.1	116
Treatments	2.48	40.2	8.8	26.5	27.9	36.8	68
Controls	1.32	66.7	68.8	18.8	6.3	6.3	48
A. 17 Fish - fried (piece)							
All locations	0.71	57.6	29.1	32.6	7.0	31.4	86
Treatments	0.74	58.9	26.0	36.0	6.0	32.0	50
Controls	0.67	55.7	33.3	27.8	8.3	30.6	36
A. 18 Cassava (small bunch)							
All locations	0.19	100.9	45.8	25.0	10.4	18.8	192
Treatments	0.23	78.9	23.8	32.7	15.8	27.7	101
Controls	0.15	133.1	70.3	16.5	4.4	8.8	91
A. 19 Plantain (5 fingers)							
All locations	0.86	76.2	32.1	39.8	9.7	18.4	196
Treatments	0.83	54.6	21.0	46.0	14.0	19.0	100
Controls	0.88	92.1	43.8	33.3	5.2	17.7	96
A. 20 Puna Yam (1 tuba)							
All locations	0.98	48.4	23.6	28.8	22.2	25.5	212
Treatments	1.16	42.8	14.3	21.4	24.5	39.8	98
Controls	0.84	48.7	31.6	35.1	20.2	13.2	114
A. 21 Groundnuts – raw (per kg)							
All locations	1.58	34.9	20.0	20.8	43.9	15.3	255
Treatments	1.72	32.8	16.7	10.3	52.4	20.6	126
Controls	1.44	35.1	23.3	31.0	35.7	10.1	129
A. 22 Groundnuts – paste (per kg)							
All locations	1.23	32.5	11.6	48.7	28.9	10.8	232
Treatments	1.13	37.0	19.4	64.5	6.5	9.7	124
Controls	1.36	25.4	2.8	30.6	54.6	12.0	108
A. 23 Chicken eggs (per crate)							
All locations	5.55	13.7	1.3	49.1	45.6	4.0	226
Treatments	5.70	14.5	2.4	35.7	57.1	4.8	126
Controls	5.39	11.9	0	66.0	31.0	3.0	100
A. 24 Red beans (per kg)							
All locations	1.20	30.1	18.2	27.2	41.9	12.1	198
Treatments	1.32	28.1	14.7	11.6	53.7	20.0	95
Controls	1.09	29.4	21.4	42.7	31.1	4.9	103
A. 25 Cassava – dough (per kg)							
All locations	0.71	68.7	30.8	36.5	17.3	15.4	156
Treatments	0.70	52.1	25.5	37.8	22.4	14.3	98
Controls	0.73	89.1	39.7	34.5	8.6	17.2	58

Fresh Foods	Mean[a]	COV[b]	Distribution of Normalized Values[c]				N
			0-.7 (%)	.71-1.0 (%)	1.1-1.3 (%)	>1.3 (%)	
A. 26 Coconut (medium size- per piece)							
All locations	0.40	48.1	24.6	42.6	23,0	9.8	61
Treatments	0.70	43.2	17.9	35.7	32.1	14.3	28
Controls	0.37	51.6	30.3	48.5	15.2	6.2	33
A. 27 Banana (1standard bunch)							
All locations	0.96	63.8	33.3	35.8	10.8	20.0	120
Treatments	0.93	56.3	29.4	38.2	16.2	16.2	68
Controls	1.00	71.6	38.5	32.7	3.8	25.0	52
A. 28 Oranges/tangerines (per piece)							
All locations	0.14	97.6	35.6	35.6	7.7	21.2	104
Treatments	0.14	102.1	30.2	39.7	11.1	19.0	63
Controls	0.13	91.4	43.9	29.3	2.4	24.4	41
A. 29 Pineapple (per piece)							
All locations	0.96	44.7	25.0	15.0	40.0	20.0	60
Treatments	0.93	49.8	27.5	15.0	40.0	17.5	40
Controls	1.02	34.7	20.0	15.0	40.0	25.0	20
A. 30 Mango (per piece)							
All locations	0.65	70.8	35.1	10.8	8.1	45.9	37
Treatments	0.87	24.5	0	20.0	15.0	65.0	20
Controls	0.39	138.0	76.5	0	0	23.5	17
A. 31 Avocado pear (per piece)							
All locations	0.62	68.4	45.5	35.2	6.8	12.5	88
Treatments	0.58	42.3	24.0	48.0	10.0	18.0	50
Controls	0.68	85.9	73.7	18.4	2.6	5.3	38
A. 32 Cocoyam leaves (standard bunch)							
All locations	0.59	49.0	12.8	56.4	8.7	22.1	149
Treatments	0.70	41.1	1.2	50.6	14.8	33.3	81
Controls	0.47	51.5	26.5	63.2	1.5	8.8	68
A. 33 Garden eggs (per piece)							
All locations	0.07	69.5	30.4	39.2	9.3	21.1	227
Treatments	0.06	65.9	29.4	39.5	8.4	22.7	119
Controls	0.07	73.0	31.5	38.9	10.2	19.4	108
A. 34 Okra (per piece)							
All locations	0.03	64.1	26.5	39.7	15.1	18.7	219
Treatments	0.03	38.6	12.7	48.3	21.2	17.8	118
Controls	0.04	84.5	42.6	29.7	7.9	19.8	101
A. 35 Carrots (per piece)							
All locations	0.24	52.8	21.6	34.1	15.9	28.4	88
Treatments	0.27	46.2	10.0	34.0	20.0	36.0	50
Controls	0.19	59.6	36.8	34.2	10.5	18.4	38
A. 36 Pepper (fresh per kg)							
All locations	2.00	82.7	62.5	5.3	5.7	26.5	264
Treatments	1.89	82.9	68.1	4.3	5.1	22.5	138
Controls	2.12	82.2	56.3	6.3	6.3	31.0	126
A. 37 Large onions (per bulb)							
All locations	0.26	56.2	28.0	34.9	11.5	25.7	261
Treatments	0.26	55.6	24.8	37.6	9.8	27.8	133
Controls	0.25	57.0	31.3	32.0	13.3	23.4	128
A. 38 Fresh tomatoes (per piece)							

Fresh Foods	Mean[a]	COV[b]	Distribution of Normalized Values[c]				N
			0-.7 (%)	.71-1.0 (%)	1.1-1.3 (%)	>1.3 (%)	
All locations	0.15	65.7	40.2	18.5	5.4	35.9	259
Treatments	0.15	50.1	35.1	23.1	8.2	33.6	134
Controls	0.15	79.8	45.6	13.6	2.4	38.4	125
A. 39 Palm nuts (per kg)							
All locations	0.44	53.2	21.0	51.5	9.0	18.6	167
Treatments	0.42	49.0	19.8	56.0	9.9	14.3	91
Controls	0.47	56.7	22.4	46.1	7.9	23.7	76
Packaged Foods							
B. 1 Sugar (granulated per kg)							
All locations	1.36	33.7	9.3	45.7	40.4	4.6	280
Treatments	1.44	36.9	3.4	47.0	41.6	8.1	149
Controls	1.27	26.6	16.0	44.3	38.9	0.8	131
B. 2 Ice cream (Vanilla Fan Ice- Large)							
All locations	0.31	9.5	0	76.0	24.0	0	75
Treatments	0.31	9.2	0	79.7	20.3	0	64
Controls	0.32	11.1	0	54.5	45.5	0	11
B. 3 Chocolate (medium)							
All locations	0.92	31.3	4.2	63.2	16.8	15.8	95
Treatments	0.84	28.7	4.8	77.8	9.5	7.9	63
Controls	1.09	28.2	3.1	34.4	31.3	31.3	32
B. 4 Salt (1 sachet)							
All locations	0.23	18.6	5.2	41.1	50.0	3.7	270
Treatments	0.23	20.0	7.0	28.0	60.8	4.2	143
Controls	0.23	17.0	3.1	55.9	37.8	3.1	127
B. 5 Biscuits (Digestive-Medium)							
All locations	2.52	35.9	6.1	73.6	10.1	10.1	148
Treatments	2.75	31.5	3.3	69.2	14.3	13.2	91
Controls	2.08	39.6	10.5	80.7	3.5	5.3	57
B. 6 Fish (Titus canned sardines)							
All locations	1.09	15.1	0	51.7	44.9	3.4	265
Treatments	1.15	12.5	0	33.6	63.5	2.9	137
Controls	1.04	16.5	0	71.1	25.0	3.9	128
B. 7 Corned beef (Exeter – large)							
All locations	4.60	6.1	0	45.8	54.2	0	107
Treatments	4.56	6.2	0	53.3	46.7	0	60
Controls	4.63	6.2	0	36.2	63.8	0	47
B. 8 Milk (powder)							
All locations	0.17	52.3	42.3	0	35.2	22.4	196
Treatments	0.16	41.2	48.6	0	32.4	18.9	111
Controls	0.19	59.2	34.1	0	38.8	27.1	85
B. 9 Baby milk (Lactogen – medium)							
All locations	7.04	10.7	2.0	51.3	`46.7	0	152
Treatments	7.17	6.8	0	46.3	53.7	0	82
Controls	6.88	13.9	4.3	57.1	38.6	0	70
B. 10 Margarine (Blue band – medium)							
All locations	1.48	19.4	5.1	39.0	50.8	5.1	195
Treatments	1.50	19.4	3.7	39.8	50.0	6.5	108
Controls	1.46	19.4	6.9	37.9	51.7	3.4	87
B. 11 Tinned milk – Ideal (unsweetened)							

Fresh Foods	Mean[a]	COV[b]	Distribution of Normalized Values[c]				N
			0-.7 (%)	.71-1.0 (%)	1.1-1.3 (%)	>1.3 (%)	
All locations	0.88	11.1	1.2	45.3	52.9	0.6	170
Treatments	0.86	9.9	0	57.7	42.3	0	78
Controls	0.88	12.0	2.2	34.8	62.0	1.1	92
B. 12 Tinned milk – Ideal (evaporated)							
All locations	0.92	7.1	0	48.6	51.4	0	290
Treatments	0.92	7.2	0	52.8	47.2	0	144
Controls	0.93	6.9	0	44.5	55.5	0	146
B. 13 Tomato puree (salsa canned – medium size)							
All locations	1.12	28.6	1.6	70.5	13.0	15.0	254
Treatments	1.02	21.3	1.6	81.6	12.0	4.8	125
Controls	1.21	30.6	1.6	59.7	14.0	24.8	129
B. 14 Bread (1 loaf)							
All locations	1.13	29.4	7.4	60.8	15.2	16.7	204
Treatments	1.19	32.0	5.9	52.0	22.5	19.6	102
Controls	1.07	24.7	8.8	69.6	7.8	13.7	102
B. 15 Coffee (Nescafe – medium tin)							
All locations	2.50	13.4	4.5	23.2	72.2	0	198
Treatments	2.50	12.7	4.4	25.7	69.9	0	113
Controls	2.48	14.3	4.7	20.0	75.3	0	85
B. 16 Chocolate drinks (Milo – medium)							
All locations	3.40	7.7	1.7	48.9	49.4	0	178
Treatments	3.93	7.1	2.2	60.2	37.6	0	93
Controls	4.07	7.9	1.2	36.5	62.4	0	85
B. 17 Lipton tea (25 bags)							
All locations	1.42	23.0	1.0	65.2	24.2	9.6	293
Treatments	1.42	21.1	.7	65.5	25.7	8.1	148
Controls	1.40	24.9	1.4	64.8	22.8	11.0	145
B. 18 Soft drinks (Coke or Fanta)							
All locations	0.58	13.0	0	50.2	47.7	2.1	285
Treatments	0.56	9.8	0	57.6	41.7	.7	144
Controls	0.59	15.0	0	42.6	53.9	3.5	141
B. 19 Malt drinks (Malta Guinness)							
All locations	0.99	7.6	0	98.7	0	1.3	79
Treatments	0.97	5.9	0	100.0	0	0	51
Controls	1.01	8.6	0	96.4	0	3.6	28
B. 20 Mineral water (Voltic bottled – large)							
All locations	0.95	20.3	0	32.2	66.1	1.7	180
Treatments	0.93	9.5	0	27.8	72.2	0	97
Controls	0.96	27.4	0	37.3	59.0	3.6	83
B. 21 Akpeteshie (Beer bottle)							
All locations	1.95	16.9	4.1	30.6	61.7	3.6	222
Treatments	1.96	18.5	3.5	38.1	51.3	7.1	113
Controls	1.93	15.1	4.6	22.9	72.5	0	109
B. 22 Palm wine/Raaffia palm wine (Beer bottle)							
All locations	0.91	63.9	36.4	12.1	27.3	24.2	33
Treatments	1.08	53.9	21.7	8.7	34.8	34.8	23
Controls	0.49	57.3	70.0	20.0	10.0	0	10
B. 23 Pito/Brukutu (Beer							

Fresh Foods	Mean[a]	COV[b]	Distribution of Normalized Values[c]				N
			0-.7 (%)	.71-1.0 (%)	1.1-1.3 (%)	>1.3 (%)	
bottle)							
All locations	0.87	59.9	41.2	5.9	20.6	32.4	34
Treatments	1.17	30.7	14.3	0	28.6	57.1	14
Controls	0.66	78.9	60.0	10.0	15.0	15.0	20
B. 24 Beer (Star Lager)							
All locations	1.47	5.9	0.4	34.6	64.5	0.4	234
Treatments	1.48	6.0	0	31.1	68.0	.8	122
Controls	1.46	5.9	0.9	38.4	60.7	0	112
Non Food Items							
C.1 Cigarette (1 pack)							
All locations	1.50	43.0	33.5	20.5	18.1	27.9	215
Treatments	1.37	42.5	39.5	24.6	15.8	20.2	114
Controls	1.63	41.6	26.7	15.8	20.8	36.6	101
C. 2 Charcoal (small bunch)							
All locations	0.18	43.9	6.2	65.5	2.3	26.0	177
Treatments	0.17	45.6	5.6	61.8	2.2	30.3	89
Controls	0.14	40.7	6.8	69.3	2.3	21.6	88
C. 3 Key soap (by the bar)							
All locations	2.24	7.1	0	61.1	38.2	0.7	283
Treatments	2.25	7.2	0	60.8	38.5	.7	143
Controls	2.22	7.0	0	55.8	34.4	0.7	140
C. 4 Lux (1 bar)							
All locations	0.43	12.9	1.9	59.5	37.0	1.5	262
Treatments	0.42	12.9	3.0	53.8	42.4	.8	132
Controls	0.43	12.8	0.8	65.4	31.5	2.3	130
C. 5Dettol (medium)							
All locations	2.29	14.2	7.2	30.6	60.3	1.9	209
Treatments	2.25	15.3	10.3	30.8	58.1	.9	117
Controls	2.33	12.7	3.3	30.4	63.0	3.3	92
C. 6 Insecticides – 1 packet (coil)							
All locations	0.93	24.8	5.1	44.1	45.8	5.1	297
Treatments	0.93	26.9	4.0	47.0	44.3	4.7	149
Controls	0.91	22.5	6.1	41.2	47.3	5.4	148
C. 7 Matches (1 small box)							
All locations	0.08	118.9	76.8	9.1	5.4	8.7	298
Treatments	0.06	43.0	70.7	16.7	6.0	6.7	150
Controls	0.09	135.5	83.1	1.4	4.7	10.8	148
C. 8 Toilet papers (1 roll)							
All locations	0.47	20.5	1.9	49.1	43.0	6.0	265
Treatments	0.43	21.5	3.7	72.6	20.0	3.7	135
Controls	0.51	16.3	0	24.6	66.9	8.5	130
C. 9 Candles (1 stick)							
All locations	0.20	13.2	2.1	68.8	27.7	1.4	285
Treatments	0.21	12.8	1.4	60.4	37.5	.7	144
Controls	0.20	13.5	2.8	77.3	17.7	2.1	141
C. 10 Pain killers (1 sachet/10 capsules)							
All locations	0.12	24.5	0.8	67.5	24.0	7.7	246
Treatments	0.13	20.7	0	71.1	24.8	4.1	121
Controls	0.12	26.9	1.6	64.0	23.3	11.2	125
C. 11 Anti malaria medicines (4 piece sachet)							
All locations	0.51	44.1	12.3	68.8	11.9	7.2	235
Treatments	0.48	21.1	8.8	78.9	8.8	3.5	114

Fresh Foods	Mean[a]	COV[b]	Distribution of Normalized Values[c]				N
			0-.7 (%)	.71-1.0 (%)	1.1-1.3 (%)	>1.3 (%)	
Controls	0.53	55.3	15.7	58.7	14.9	10.7	121
C. 12 Condoms (3 pack)							
All locations	0.22	49.4	27.6	46.7	5.0	20.6	199
Treatments	0.23	52.9	21.1	49.5	5.3	24.2	95
Controls	0.21	44.9	33.7	44.2	4.8	17.3	104
C. 13 Petrol (1 liter)							
All locations	1.26	12.9	0	49.6	48.0	2.4	123
Treatments	1.22	12.0	0	63.4	35.2	1.4	71
Controls	1.32	12.6	0	30.8	65.4	3.8	52
C. 14 Diesel (1 liter)							
All locations	1.23	10.3	0	53.6	46.4	0	110
Treatments	1.20	8.9	0	63.6	36.4	0	66
Controls	1.28	11.0	0	38.6	61.4	0	44
C. 15 Exercise books (small)							
All locations	0.30	22.7	8.5	37.2	43.4	10.9	258
Treatments	0.28	28.0	13.4	48.0	26.0	12.6	127
Controls	0.30	16.4	3.8	26.7	60.3	9.2	131
C. 16 Mesh/wigs (Nina weaves)							
All locations	2.91	25.6	2.5	63.3	21.7	12.5	120
Treatments	2.75	19.6	1.3	73.3	20.0	5.3	75
Controls	3.17	30.0	4.4	46.7	24.4	24.4	45
C. 17 Toothpaste (Pepsodent large)							
All locations	1.44	9.3	1.5	43.2	55.0	0.4	271
Treatments	1.46	9.3	1.5	39.7	58.1	.7	136
Controls	1.43	9.0	1.5	46.7	51.9	0	135
C. 18 Razor blades (5 blade sachet)							
All locations	0.43	83.5	65.2	3.3	8.9	22.6	270
Treatments	0.38	89.7	75.4	4.2	2.8	17.6	142
Controls	0.49	76.3	53.9	2.3	15.6	28.1	128
C. 19 Sure deodorant							
All locations	2.36	36.8	27.6	17.2	45.7	9.5	116
Treatments	2.58	33.6	14.3	20.0	55.7	10.0	70
Controls	2.03	37.8	47.8	13.0	30.4	8.7	46
Transportation tariffs							
Passengers							
Destination 1	1.23	146.9	53.1	19.3	9.0	18.6	290
Destination 2	1.31	146.0	55.1	12.9	15.1	16.9	272
Destination 3	1.77	164.7	63.8	9.8	7.7	18.7	246
Freight							
Destination 1	0.87	139.2	50.2	10.9	25.3	13.7	285
Destination 2	1.00	169.4	50.2	9.7	22.5	17.6	267
Destination 3	1.21	175.0	53.9	22.2	8.6	15.2	243

a. In Ghanaian cedi
b. Definition: (standard deviation/mean)*100
c. Normalized prices were computed as the mean observed price of a good in a locality divided by the mean price of the good observed in all localities.

The COVs for transportation are an order of magnitude larger than for the goods. This is expected because the destinations are specified for each locality.

We can explore the variation in transportation tariffs by examining how tariffs vary with the time traveled. Recall that in each locality interviewers identified the three most common destinations from that locality for passengers and for freight and then obtained three price quotes for each destination. The tabulation in Table 4.2 shows regularity for passenger trips: prices rise steadily with the time required to reach the destination. The pattern is not nearly as strong for freight haulage but the most distant destination regularly costs more than the closest.

Table 4.2 Transportation Tariffs by Time Traveled to the Destination
(Average cost per trip in cedis)

Time traveled (minutes)	Passenger tariffs by destination			Freight tariffs by destination		
	1	2	3	1	2	3
<10	.54	.79	.96	.56	.69	.80
11-30	.91	1.05	1.36	.88	1.22	1.36
31-60	.94	1.26	1.05	.73	.88	.85
61-120	1.70	1.69	2.30	.91	.84	1.13
120+	3.44	2.53	4.54	1.54	1.24	2.03

4.2 Price Variation with Local Conditions

To what degree can price variations be associated with simple measures such as the size of the locality, measured by population or the distance from the next large community? Based on the price patterns observed in Table 4.1, the response might be "not much," since the price variance for a number of items was really quite low.

Nevertheless, it is reasonable to explore the kind of explanations suggested in the previous paragraph. The table below list several hypotheses that we can test with the data at our disposal.

Table 4.3 Hypotheses on Reasons for Price Variability and Corresponding Variables

Hypothesis	Variable specification	Variable name
Larger localities will have lower prices because of greater competition among sellers. This may be offset, however, by higher wages and shop rents in bigger towns.	Population of locality in 2000 in thousands	POP2000
Localities located close to larger towns will have lower prices because other options for making purchases are reasonably convenient	Time in minutes to the closest town of over 10,000 population	TT_CITY10K2
More accessible localities have lower prices	Travel time in minutes to Accra divided by 100	TTACCRA2
	Travel time in minutes to nearest trunk road	TRNKTME2
	Distance to nearest feeder road in kilometers	DIST_FEEDER2
Localities receiving greater amounts of rain are likely to have higher prices because roads are impassable or difficult to traverse some of the time	Average annual rainfall in meters per year	RAINFALL
Prices are lower on market day because of greater competition	Variable = 1, if prices were recorded on a market day	MKTDAY

We estimated standard regression models for three composite prices to test these hypotheses.[4] The three composites are: fresh food items, packaged food items, and non-food items. A composite price was computed for a locality as the sum of the normalized prices of the included items divided by the number of items included in the composite. The same set of items had to be included for each locality for consistency. Hence, the set of items included in each composite depended on the number of localities where we had valid observations for all items in the composite bundle of items. Missing observations are scattered widely among the 308 sample localities so that in the end both the number of items in a composite price and the number of localities for which it could be computed were small, as shown in the last panel of Table 4.4.

The results are displayed in Table 4.4. In terms of identifying conditions associated systematic price variation, the model for fresh food items is the worst performer. This makes sense because a larger share of fresh food items is available locally than for package food items and non food items. The only significant variable is locality population. The positive sign of the coefficient indicates in terms of the hypotheses stated earlier that prices are higher in larger places owing to higher rents and wages that offset the effects of more sellers increasing competition. The quantitative effects of an increase in locality population has a small effect on prices: based on the estimated price elasticity with respect to population, a 100 percent increase in average population (from about 10,000 to 20,000) results in just a 2 percent increase in composite price of fresh foods.

[4] In addition to the linear models shown in table, we also estimated models in which the independent variables were specified in quadratic form. These did not yield different results. The variable TRNKTME is not included in the model because it was found to be highly correlated with other accessibility measures.

In the models for both packaged foods and non food items the size of the settlement is not significant. We interpret this as meaning that the effects of the two factors just noted are offsetting each other. The fact that it was market day in some localities when the price data was collected did not have a significant effect on prices after controlling for the other factors in the models. This is certainly a different result than expected, particularly for fresh food items. Additionally, the amount of rainfall where a locality is situated does not influence pricing. The effect of rainfall depends critically on road conditions and if roads are in good condition, the effect could be negligible. Road conditions are likely being accounted for to a major extent in the travel time variables.

Table 4.4 Results of Estimated Regression Models for Composite Normalized Prices

	Composite Groups		
	Fresh food items	**Packaged food items**	**Non food items**
Independent variables			
Constant	0.96	0.87	1.06
POP2000	.002*	.001	.001
TTACCRA2	-.024	-.029	-.101
TT_CITY10K2	.003	.014**	.034*
DIST_FEEDER2	.014	.009**	.001
RAINFALL	.002	.001	.001
MKTDAY	.170	-.016	.032
Significance indicators			
R square	.343	.254	.178
F-statistic	2.26	3.30	2.50
Significance level	0.069	0.007	.030
Memorandum items			
N	33	65	76
Number of items included	17	16	14
Mean of dependent variable	1.06	.88	.99

*Statistically significant at the 0.10 level
**Statistically significant at the 0.05 level or higher

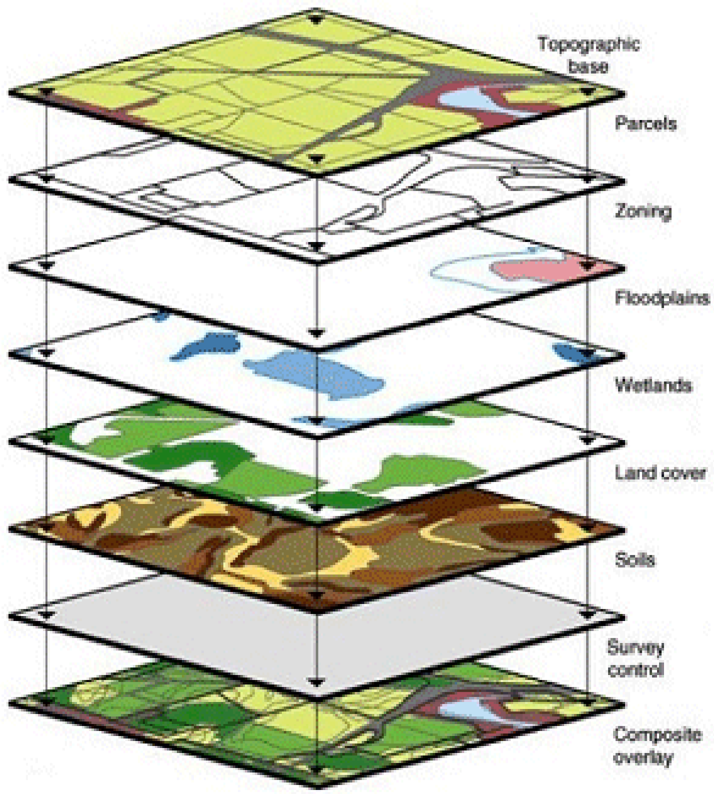
In the model for packaged food items, two variables measuring distance from opportunities for more venders to offer products are highly significant: travel time to the nearest city of at least 10,000 population (TT_CITY10K2) and the kilometers the closest feeder road (DIST_FEEDER2). Both indicate that the greater the locality's isolation, the higher the prices for packaged goods. Again, however, the quantitative effects are small based on elasticity estimates: doubling the travel time to Accra raises packaged food prices by 8 percent and doubling the distance to the closest feeder road from 1.9 to 3.8 kilometers pushes these prices up by 2 percent.

Finally, for the non food items model, the only significant variable in explaining price variations is travel time from the locality to Accra (TTACCRA2). The quantitative effect here while still small is larger than for packaged foods: doubling the travel time to Accra increases these prices by 9.5 percent.

Overall, the models substantially confirm the kind of systematic price variation one would anticipate. This in turn lends confidence to the quality of the data collected.

Annex A:
GIS Systems and GIS Accessibility

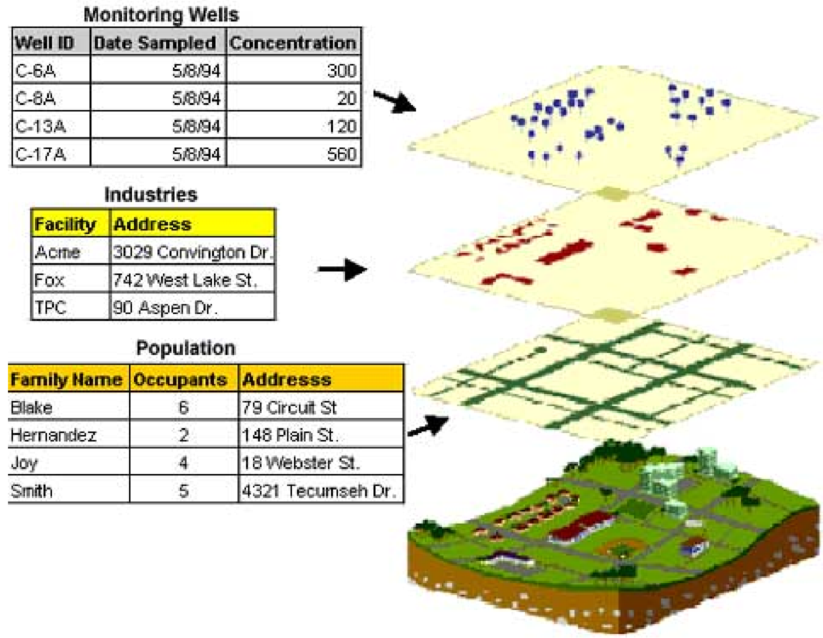
Geographic Information Systems (GIS)



A Geographic Information System (GIS) is a computer geo-database system for capturing, storing, checking, integrating, manipulating, analyzing and displaying data related to positions on the Earth's surface. In essence, a GIS is simply a standard database (running on Oracle or SQL Server platforms, for example) that also has the functionality of incorporating digital geo-spatial data (for query, display and data storage purposes). These geo-spatial data variables and inputs might be represented as several different "layers", where each layer holds data about a particular kind of feature. Each feature is linked to a position on the graphical image on a map and a record in an attribute table. By layering information such as road networks, village or community locations, and population, spatial relationships among the objects being mapped can be emphasized. A GIS differs from other information systems because it combines common database operations such as query and statistical analysis with the benefits of visual and geographic analysis offered by maps.

GIS can relate otherwise disparate data on the basis of common geography, revealing hidden patterns, relationships, and trends that are not readily apparent in spreadsheets or statistical packages, often creating new information from existing data resources.

Although quite complex and robust GIS data formats exist and are increasingly used in an ever-growing range of industries and sectors, in a simplified representation GIS represents real-world objects in four major categories: points, arcs (straight or curved line segments), polygons, or in a grid/pixel system. The point, line and polygon data model is referred to use the umbrella term "vector", while the grid/pixel model (such as with a digital air photo or satellite image that is "geo-rectified" to a cartographic projection system, such as latitude/longitude) is referred to as the "raster" data model. In each case, the GIS computer database keeps track of the geo-location of each object, and recognizes each point, line, polygon or raster pixel as a distinct object, having its own properties. Thus, for example, the computer keeps track of the geo-location, length and starting and ending points of each individual curvilinear line segment in an integrated road network. In addition, tabular data can be attached to each GIS object, and is incorporated into the central GIS database in "attribute tables". Thus, for example, if cities or towns are represented by spatial point objects, then any of a host of socioeconomic variables describing those cities or towns can be included in the database and integrated (such as city/town population, number of households, date of incorporation, etc.). In the case of a curvilinear line segment in a GIS road network, variables describing the pavement type, approximate road quality, legal speed limit, or date of last maintenance can be included in the GIS database.



Transportation Modeling and GIS

GIS has been used for decades for transportation modeling, and there is a rich precedent and literature for GIS methodologies for transportation modeling, including the calculation of travel-times through a road network and accessibility indices (de Dios Ortuzar and Willumsen 2002; Miller and Wu 2000; Liu et al. 1994) GIS has been used, for example, by city transportation planning agencies in the US for several decades, and currently almost every major city and town in the US and in most developed countries performs transportation modeling, maintenance and management using GIS systems. Most frequently, transportation and traffic flows are modeling in GIS through a "vector" road network database. Data on traffic flows and volumes is inputted into the attribute tables for the road network, and then traffic movements and flows, incorporating travel-cost and congestion models (see below) are used to "paint a picture" of road usage and traffic flows as an initial evaluation step. This initial analysis seeks to monitor immediate direct effects related to changes in road usage and traffic flows in the improvement area, and the goal is to see which groups of beneficiaries are using the road initially and over time for which kinds of purposes in terms of trips and repeat trips. For example, which households are using the improved road sections for trips from home to work, or to shopping?

What are the home to work traffic flow patterns and travel times, the existing usage and directional flows? The analysis is often supported by integrating digital vector road networks, community and industry locations, physiographic data and road usage “zones” in a spatial GIS database.

Travel-Time, Travel-Cost and Accessibility Indices

There is extensive spatial economic theory (agglomeration theory) that describes the fact that spatial access to markets, controlled by transportation costs, is crucial in economic development. In this sense, evaluation of impact of the road upgrades is unlike other impact evaluations in that the level of treatment is not a discrete binary function (road or no road) but a continuous one, particularly where the treatment is not the construction of a new road where one did not previously exist, but the upgrading of an existing road. In such cases, the degree of treatment varies in two ways:

- ***Degree of access to the road.*** Because roads have a fixed spatial placement, access to a road of a particular household or community is a function of location. The level of access for a household located adjacent to the road is better than that of a household located some distance from the road.
- ***Quality of access to the road.*** The second dimension of access is quality; is the means of access one that allows for easy and efficient travel and transport of good to the road? Higher quality access to the road is expected to yield a greater impact.

GIS can be used to assess accessibility as a function of road quality, the time of road building, geography, topography, and other factors that aid or hinder access ("quality of access"), including political or administrative policies or traffic congestion (Miller and Wu 2000; Rosero-Bixby 2004) Using GIS to give continuous accessibility values to observational units, regressions between continuous indices and selected impact variables could be run, to illuminate trends and patterns, establish correlations, and bolster and support conclusions. Furthermore, multivariate regression models could be constructed with the inclusion of controls that might influence or mitigate true accessibility, such as tax policies or after-effects of natural disasters.

Spatial economic theory as far back as the famous Von Thunen (von Thünen and Hall 1826) land rent model has been based on the assumption that spatial access to markets, controlled by transportation costs, is crucial for economic development. In principle, improved access to consumer markets (including inter-industry buyers and suppliers) will increase the demand for a firm's products, thereby providing the incentive to increase scale and invest in cost-reducing technologies. Marshall (Marshall 1890) showed that the geographical concentration of economic activities can result in a "snowball" effect, where new entrants tend to agglomerate to benefit from higher diversity and specialization in production processes. Workers and firms would benefit from gaining access to an agglomeration as they could expect higher wages and to have access to a larger set of employers. Furthermore, access to markets or economic city/town agglomerations can determine if a household is able to afford the cost of shipping products for sale, earning potentially higher wages in agglomeration centers, or gaining access to information spillovers or technology advances, further reducing costs.

There is a rich body of literature on the benefits to firms from gaining improved access or proximity to other firms in the same industry (Henderson 1974, 1988; Carlino 1979). Theoretical and empirical work on urban economics and economic geography (Henderson, Shalizi, and Venables 2001)suggests that the net benefits of industry concentration and location in dense urban areas are disproportionately accrued by technology intensive and innovative sectors. This is because the benefits of knowledge sharing (ideas) and access to producer services (e.g., venture capital) are considerably higher in these sectors than in low-end manufacturing that employs standardized production processes. As a result, these innovative sectors can afford the high wages and rents in dense urban locations and industry clusters. Paradoxically however, we find a considerable range of standardized industrial activity in most developing country countries. One explanation for this is the lack of inter-regional transport infrastructure linking small centers to large urban areas, thereby reducing the opportunities for efficient location decisions and de-concentration of large urban areas. In a recent empirical study, (Henderson and Box 2000)documents the linkages between improvements in inter-regional infrastructure and growth of smaller agglomerations outside of larger city centers.

In general, "access" to markets is determined by the household's or village's true cost of traveling to or accessing market centers. This could include the cost of transporting goods for sale, transporting (back to the village) key inputs for production or consumption, or the cost of transporting people for migratory or more permanent employment. Thus, effective access to urban markets also depends on the willingness and ability to afford transport costs, and these in turn are directly a function of road quality as well as actual measured road distance, topography, climate, rivers or any other potentially inhibiting (and thus more costly) exogenous geo-physical barriers.

The classic gravity model which is commonly used in the analysis of trade between regions and countries states that the interaction between two places is proportional to the size of the two places as measured by population, employment or some other index of social or economic activity, and inversely proportional to some measure of separation such as distance. Following (Hansen 1959)

$$I_i = \sum_j \frac{S_j}{d_{ij}^b}$$

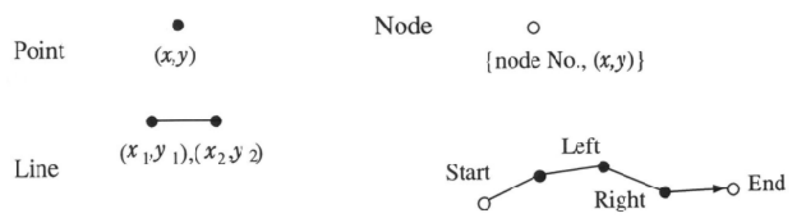
where I is the "classical" accessibility indicator estimated for location i (for example, a village), S is a size indicator at a market destination *j* (for example, population, purchasing power or employment), and *d* is a measure of distance (or more generally, *friction*) between origin i and destination *j*, while *b* describes how increasing distance reduces the expected level of interaction. Empirical research suggests that simple inverse distance weighting describes a more rapid decline of interaction with increasing distance than is often observed in the real world (Weibull 1976), and thus a negative exponential function is often used.

There are several options for developing accessibility indicators depending on the choice of distance variables used in the computation. These include: (a) indicators based on "straight-line" or Euclidean distance; (b) indicators incorporating topography; (c) indicators incorporating the availability of transport networks; (d) indicators incorporating the quality of transport networks; and (e) movement across a "cost surface".

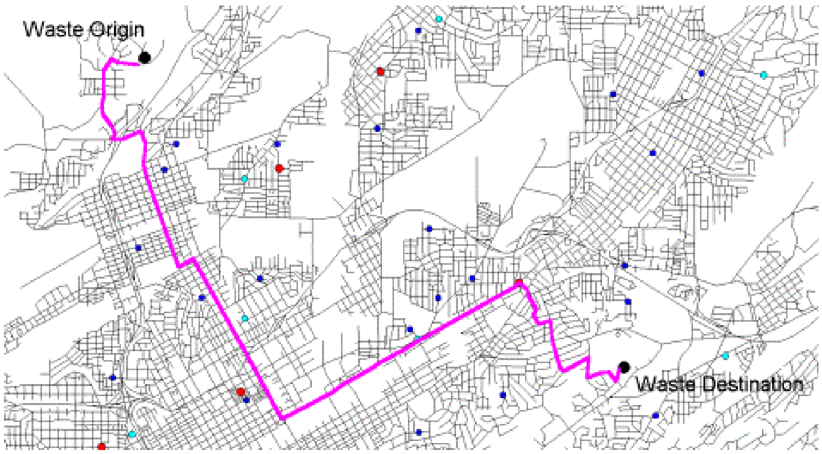
A better alternative is to use actual measured distance along road networks as the basis of the inverse weighting parameter and to incorporate information on the quality of different transportation links. Feasible travel speed and thus travel times will vary depending on each type of network link. A place located near a national highway will be more accessible than one on a rural, secondary road. The choice of the friction parameter of the access measure will therefore strongly influence the shape of the catchment area for a given point, i.e., the area that can be reached within a given travel time. This, in turn, determines the size of potential market demand as measured by the population within the catchment area.

In studies related to agglomeration economies and economic geography (H. Hanson 2005)the distance measure of choice is usually the straight-line (Euclidean, or “as-the-crow-flies”) distance, which has the advantage of computational simplicity. However, this assumption of uniform (isotropic) plane is clearly unrealistic, particularly in countries where topography and sparse transport networks of uneven quality greatly affect the effort required to move between different parts of the country. Such an access index takes no account of the fact that hills and mountains greatly reduce travel times and greatly increase travel costs. Nor does it take into account the fact that people and goods move along road networks – not across a uniform plane. Topographic data (such as from contour lines digitized from paper maps, or from spot samples taken on the ground by surveyors, or from airborne or satellite instruments) can be converted using GIS algorithms to a continuous elevation *surface*. In that case, distance *across topography* can be calculated: the GIS calculates Euclidean distance, but then further calculates the actual distance on the ground considering topographic variation. This is partially illustrated by this graphic at right, where distance is measured both across the two-dimensional x,y surface, but also across the topographic z surface, calculating actual distance traveled (in meters, kilometers, etc.).

A far better alternative, however, is to use actual measured distance along existing road networks, considering the fact that goods and people move predominantly along infrastructure networks. This can be accomplished by obtaining an accurate digital GIS road network. Such a road network has all roads digitized into GIS digital “vector” objects. That is, rather than simply a graphical image of the roads, the road network is actually made up of many individual line segments, connected to each other at the end points, which are called “nodes”. Each individual road segment in the larger network is “seen” by the GIS as an individual digital object. The GIS can calculate the exact length, direction and curvature of each line segment (just as it can for polygonal objects). These graphics illustrates the underlying road network structure, which is technically referred to as “vector line topological structure”:



The GIS keeps track of the exact geo-location of each node connecting linear road segments, as well as the exact curve of each line segment. Thus the GIS can calculate precisely the exact distance along each segment in any desired unit (such as meters, kilometers, etc.). Thus, using advanced algorithms, the GIS can calculate travel distance *through* the road network from any node to any other node. Other algorithms (such as the Djikstra algorithm (Dijkstra 1959; Liu et al. 1994) or variants) will pick a “shortest path” through the network to get from node X to node Y, minimizing travel distance, as in this graphic below. Here the GIS has simply found the shortest path through the road network assuming that all road network segments are equal in terms of road quality or road speed. However, data on road quality or road speed of each individual road segment is often available, and can be entered into the GIS database and attached to each road segment (in fact any amount of information on road segments – or any other object in the GIS – can be entered into the database, such as data on road segment names, date of paving, cost per segment, number bridges per segment, etc – all of this information is kept track of in the GIS database). If data on road quality is available, then approximate road speeds can be estimated. Typically, road maps categorize roads into categories as in this map from Portugal (below).



For example, if a road is categorized as "one-lane paved", then an approximate road speed of 45 miles per hour could be assigned to all road segments with that categorization. Once categories of roads are assigned approximate road speeds, then travel times through the road network *considering road speed/road quality* can be calculated. This is a simple calculation: road length divided by road speed. For example, if the road segment is 50 kilometers long, and the road speed of that segment is 25 kilometers per hour, then the travel time (or travel "cost" if the definition of cost here is time) would be 50/25 or 2 hours of travel time. Often, this results in a different "least cost" or "least time" (if the "cost" is in terms of speed) pathway than the minimum distance pathway along all road networks. For example, it may be quicker in terms of time/cost to drive onto a highway and then exit to get to a destination than to travel along intermediate roads even though they provide a more direct link. Thus, the pathway of minimum distance may not always be the same as the pathway of minimum time or cost. This graphic shows the *fastest* route through a network from one destination to another, rather than the minimum distance route.

Topographic information could be combined with road network speed information, so that the road network segments are weighted by elevation or slope. For example, one might burn less gas or put less stress on a truck (lower "cost") to drive around a mountain than across it, even though the minimum distance pathway is across the mountain. In this case, the path of "accessibility" would likely be around the mountain.

Physiographic Data To Weight Travel Cost Estimations

While measuring distance along road networks incorporating data on varying road quality or varying road speed is a far superior method than measuring access "as the crow flies" (Euclidian distance) or even along road networks without considering road quality, the accuracy of the computed access indices can be further enhanced by incorporating weights that reflect further variable that impede travel, adding travel cost and time. For example, topography (as well as slope angle) is an extremely important variable that could dramatically alter travel times and costs, but might not be considered at all if only road network distance and road quality were considered. While digital data on a road network might indicate that a particular stretch of road was paved at high quality, with an official speed limit of 80 kilometers per hour, nonetheless in reality that stretch might involve movement up and down steep hills, in effect slowing travel time and increasing travel cost beyond what is measured simply by the road network data. Furthermore, a flat stretch of road in a low-lying area that rarely encounters debilitating weather such as snowstorms might overall be much easier (and cheaper) to travel than a similar flat stretch of identical road quality located at high elevations. On the latter, travel may frequently be inhibited by severe snow or ice, thus dramatically increasing travel costs.

Other important physiographic factors can affect actual travel costs and times, including land cover, climate, rainfall amounts, and the presence of lakes, rivers, streams and glaciers, which may periodically overflow, or swell during certain times of the year. Furthermore, a road network map may not indicate that certain areas are restricted because they are protected – either for conservation or military purposes, for example – and thus travel through them is impractical. In that case, the road network will need to be digitally altered to reflect the actual travel routes.

By the same token, certain physiographic factors can provide exogenous drivers of village economic productivity, such as inherently fertile soils that would result in higher agricultural productivity, or favorable rainfall patterns or climate, etc. Villages located in areas with good access to clean water, or with less intimidating (and costly) topography for villagers to drive and navigate, might have an inherent (exogenous) advantage over other villages with very similar socioeconomic measures. In their absence, for example, an economic increase in one village over another might be falsely attributed to superior road access, rather than to superior soil fertility, which may be the true driver. Or the reverse could occur, blunting the effective measurement of true, positive road benefits. Ignoring such physiographic conditions for villages could also ignore another key element of "accessibility": the fact that market access may be more valuable for some communities than for others. For example, a community with inherently poor soil fertility may benefit more from access to a fertilizer market than a community with inherently rich soils, but having the same level of access as measured by road distance, quality and even topography.

Once these data are assembled in the GIS, along with geo-locations of impacted communities (such as villages), then the GIS can quickly "map" to each community variables describing the respective physiographic conditions for each. Also, these data inputs can be used to weight the road network segments, as well as the areas of land leading to the nearest road network (in the case of villages that have no road network connection, if these exist).

References

Carlino, G. A. 1979. “INCREASING RETURNS TO SCALE IN METROPOLITAN MANUFACTURING*.” *Journal of Regional Science* 19:363-372.

Dijkstra, E. W. 1959. “A note on two problems in connexion with graphs.” *Numerische mathematik* 1:269-271.

de Dios Ortuzar, J., and L. G. Willumsen. 2002. *Modelling transport*. Wiley.

H. Hanson, G. 2005. “Market potential, increasing returns and geographic concentration.” *Journal of International Economics* 67:1-24.

Hansen, W. G. 1959. “How accessibility shapes land use.” *Journal of the American Planning Association* 25:73-76.

Henderson, J. V. 1974. “The sizes and types of cities.” *The American Economic Review* 640-656.

Henderson, J. V. 1988. *Urban development: theory, fact, and illusion*. Oxford University Press, USA.

Henderson, J. V., and B. Box. 2000. “The effects of urban concentration on economic growth.” *NBER working paper*.

Henderson, J. V., Z. Shalizi, and A. J. Venables. 2001. “Geography and development.” *Journal of Economic Geography* 1:81.

Liu, B. et al. 1994. “Finding the shortest route using cases, knowledge, and Djikstra'salgorithm.” *IEEE Expert* 9:7-11.

Marshall, A. 1890. *Principles of economics*. Macmillan and Co.

Miller, H. J., and Y. H. Wu. 2000. “GIS software for measuring space-time accessibility in transportation planning and analysis.” *GeoInformatica* 4:141-159.

Rosero-Bixby, L. 2004. “Spatial access to health care in Costa Rica and its equity: a GIS-based study.” *Social Science & Medicine* 58:1271-1284.

von Thünen, J. H., and P. Hall. 1826. *Von Thünen's isolated state*. Pergamon.

Weibull, J. W. 1976. “An axiomatic approach to the measurement of accessibility.” *Regional Science and Urban Economics* 6:357-379.

Annex B:
Survey Instruments

Millennium Development Authority

REPUBLIC OF GHANA

GHANA MARKET SURVEY
PRICE MODULE

PHASE 1: 2009

REGION: □□□ **MARKET NUMBER**: □□□□

GHANA MARKET SURVEY
PRICE MODULE

PHASE 1: 2009

REGION.. ☐☐ DISTRICT ....…. ☐☐☐☐

NAME OF LOCALITY:____________________________________

MARKET NUMBER: ☐☐☐☐

DATE: ☐☐ ☐☐ ☐☐☐☐
DAY MONTH YEAR

INTERVIEWER:_________________________________ CODE: ☐☐☐

SUPERVISOR:______________________________________ CODE: ☐☐☐

IS IT MARKET DAY?: YES NO

EDIT CHECK: COMPLETE ☐

INCOMPLETE ☐

REMARKS: ____

____

____

CODE	ITEM	1ST OBSERVATION	2ND OBSERVATION	3RD OBSERVATION

Market ID:_______________

I. FOOD PRICES

		SIZE	PRICE	SIZE	PRICE	SIZE	PRICE
001	Guinea corn/sorghum						
002	Maize						
003	Millet						
004	Rice (Local)						
005	Gari						
006	Kolanuts (Small Basket)						
007	Ginger (Small Basket)						
008	Flour (wheat)						
009	Maize ground/corn dough						
010	Cocoyam (small bunch/10 pieces)						
011	Beef						
012	Goat meat						
013	Mutton						
014	Chicken – fresh/high quality						
015	Red Fish (fresh)						
016	Fish (smoked tuna)						
017	Fish (fried)						
018	Cassava (small bunch/10 pieces)						
019	Plantain (high quality/5 fingers/non-ripe)						
020	Puna Yam (1 tuba)						

Market ID: _______________

I. FOOD PRICES (Continued)

CODE	ITEM	1ST OBSERVATION		2ND OBSERVATION		3RD OBSERVATION	
		SIZE	PRICE	SIZE	PRICE	SIZE	PRICE
021	Groundnuts (raw)						
022	Groundnuts (Paste)						
023	Chicken eggs (1 dozen)						
024	Coconut oil						
025	Groundnut oil						
026	Palm kernel oil						
027	Palm oil						
028	Red Beans						
029	Cassava - dough						
031	Coconut (high quality – medium size)						
032	Banana (1 standard bunch)						
033	Oranges/tangerines (Small Basket)						
034	Pineapple (Small Basket)						
035	Mango (Small Basket)						
036	Avocado pear (Small Basket)						
037	Cocoyam leaves (kontomire/standard bunch)						
038	Garden eggs (Small Basket)						
039	Okro (Small Basket)						
040	Carrots						
041	Pepper (fresh)						
042	Large Onions (Small Basket)						

043	Fresh Tomatoes (Small Basket)						
044	Palm nuts						

Market ID:_________________

I. FOOD PRICES (Continued)

CODE	ITEM	1ST OBSERVATION		2ND OBSERVATION		3RD OBSERVATION	
		SIZE	PRICE	SIZE	PRICE	SIZE	PRICE
045	Sugar (cube, granulated)						
046	Local Honey (Large Bottle)						
047	Ice cream (Vanilla Fan Ice - Large)						
048	Chocolate (Golden Tree– medium)						
049	Salt (1 Sachet)						
050	Biscuits (Digestive – Medium)						
051	Fish (Titus canned sardines) Normal Standard Size						
052	Corned beef (Exeter – Large)						
053	Milk (powder)						
054	Baby milk (Lactogen –Medium)						
055	Margarine (Blue Band – Medium)						
056	Tinned milk - Ideal (unsweetened)						
057	Tinned milk – Ideal (evaporated)						
058	Tomato puree (Salsa canned – medium size)						
059	Bread – sugar bread (1 loaf)						
060	Rice (Imported) Texas Long Grain Medium Size Bag						
061	Coffee (Nescafe – medium tin)						
062	Chocolate drinks (Milo – medium)						
063	Lipton Tea (25 bags)						

064	Soft drinks (Coke or Fanta)						
065	Malt drinks (Malta Guinness)						
066	Mineral water (Voltic bottled - large						

Market ID:______________

I. FOOD PRICES (Concl'd)

CODE	ITEM	1ST OBSERVATION		2ND OBSERVATION		3RD OBSERVATION	
		SIZE	PRICE	SIZE	PRICE	SIZE	PRICE
067	Akpeteshie (Beer bottle)						
068	Palm wine/Raffia palm wine (Beer bottle)						
069	Pito/Brukutu (Beer bottle)						
070	Beer (Star Lager)						

II. NON - FOOD PRICES

CODE	ITEM	DESCRIPTION	1ST PRICE	2ND PRICE	3RD PRICE
071	Cigarette (1 pack)				
072	Charcoal (small bunch)				
073	Key Soap (by the bar)				
074	Lux (1 bar)				
075	Dettol (medium)				
076	Insecticides – 1 PACKET (coil)				
077	Matches (1 small box)				
078	Toilet papers (1 roll)				
079	Candles (1 stick)				
080	Pain killers (1 sachet/10 capsules)				
081	Anti malaria medicines (Malafin/4 piece sachet)				
082	Condoms (champion 3 pack)				
083	Petrol (1 liter)				
084	Diesel (1 liter)				

085	Exercise books (small exercise book)				
086	Mesh/wigs (Nina weaves)				
087	Toothpaste (Pepsodent large)				
088	Razor blades (5 blade sachet)				
089	Sure deodorant				

Millennium Development Authority

REPUBLIC OF GHANA

GHANA MARKET SURVEY

TARIFF & TRANSPORT MODULE

PHASE 1: 2009

REGION: | | | MARKET. NUMBER: | | | | |

GHANA MARKET SURVEY
TARIFF & TRANSPORT MODULE

PHASE 1: 2009

REGION.. [ ][ ]

DISTRICT ....…….. [ ][ ][ ][ ]

NAME OF LOCALITY:____________________________________

MARKET NUMBER: [ ][ ][ ][ ]

DATE: [ ][ ] [ ][ ] [ ][ ][ ][ ]
DAY MONTH YEAR

WHAT IS TRANSPORTED

PEOPLE ONLY….1
GOODS ONLY….2
PEOPLE & GOODS…3

[ ]

INTERVIEWER:_________________________________ CODE: [ ][ ][ ]

SUPERVISOR:______________________________________ CODE: [ ][ ][ ]

REMARKS: ____

____

____

Transport Costs

Market ID:___________________

1ST OBSERVATION	
Station Name	

Destination 1			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Destination 2			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Destination 3			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Was information on goods also collected at this location?

☐ Yes

☐ NO

Transport Costs

Market ID:___________________

2ND OBSERVATION	
Station Name	

Destination 1			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Destination 2			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Destination 3			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Was information on goods also collected at this location?

☐ Yes

☐ NO

Transport Costs

Market ID:____________________

3RD OBSERVATION	
Station Name	

Destination 1			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Destination 2			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Destination 3			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Passenger Cost

Was information on goods also collected at this location?

☐ Yes

☐ NO

Tariff Costs

Market ID:___________________

1ST OBSERVATION	
Station Name	
Transporter Name	
Locating Information	
Type of goods transported	

Destination 1			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Destination 2			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Destination 3			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Was passenger transport information also collected at this location?

☐ Yes

☐ NO

Tariff Costs

Market ID:__________________

2ND OBSERVATION	
Location Name	
Respondent Name	
Locating Information	
Type of goods transported	

Destination 1			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Destination 2			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Destination 3			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Was passenger transport information also collected at this location?

- ☐ Yes
- ☐ NO

Tariff Costs

Market ID:__________________

3RD OBSERVATION	
Location Name	
Respondent Name	
Locating Information	
Type of goods transported	

Destination 1			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Destination 2			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Destination 3			
Locating Information			
Type of Vehicle			
Starting Point	Ending Point	Travel Time	Tariff

Was passenger transport information also collected at this location?

☐ Yes

☐ NO

Millennium Development Authority

REPUBLIC OF GHANA

GHANA MARKET SURVEY

LOCATING SHEET

2009

REGION: MARKET NUMBER:

Market ID:___________________________

Owners Name	
Store Name	
Stall Number	
Locating Information	

Items Priced

ITEM CODE	**Please Select the Observation Number for Each Item Recorded (select one only)**		
	1ST OBSERVATION	**2ND OBSERVATION**	**3RD OBSERVATION**

Please select the size of the store: **Large** ☐

Medium ☐

Small ☐

Annex C:
Selected Results of Regression Models for Imputing Values for Invalid Observations[5]

[5] Variable definitions follow the tables.

Fresh food items

Variable	Gari	Maize	Beef	Onions	Pepper
Constant	0.970	0.710	4.242	.250	3.410
POP2000	1.086E-6	1.373E-6	4.542E-6	6.613E-7	6.196
TTACCRA2	.000	.001	.001	.000	.004
TT_CITY10K2	.001	.000	-.008	.000	-.005
DIST-FEEDER2	-6.918E-6	8.114E-6	.000	6.394E-6	.000
RAINFALL	8.153E-7	-6.859E-5	.001	-2.827E-6	-.002
MRKTDAY_P1	-.125	-.041	-.327	.032	1.323
R2	.054	.034	.111	.034	.261
F-statistic	2.590	.920	1.231	1.223	14.661
Significance	.019	.482	.303	.296	.000

Packaged food items

Variable	Sugar	Salt	Milk powder	Coffee	Tomato puree
Constant	2.070	.238	7.019	2.643	.860
POP2000	8.018E-6	-1.928E-7	3.409E-6	2.170E-6	3.872E-6
TTACCRA2	.000	-5.864E-5	-.001	-6.774E-5	.000
TT_CITY10K2	.000	.000	.004	.004	-.002
DIST-FEEDER2	4.734E-5	2.723E-6	3.312E-5	-2.603E-5	5.690E-6
RAINFALL	-.001	1.913E-6	4.838E-6	.000	.000
MRKTDAY_P1	.035	.010	-.011	-.136	-.027
R2	.120	.074	.048	.095	.136
F-statistic	3.831	2.380	1.173	1.748	5.346
Significance	.001	.031	.324	.118	.000

Non-food items

Variable	Cigarette	Lux	Matches	Pain killers	Toothpaste
Constant	.975	.446	-.021	.134	1.503
POP2000	3.500E-6	-9.190E-7	6.591E-7	-6.265E-8	-1.034E-6
TTACCRA2	6.887E-5	.000	1.047E-6	2.116E-5	-5.594E-5
TT_CITY10K2	-.001	.000	.000	-7.714E-5	-1.269E-5
DIST-FEEDER2	-1.018E-5	-1.430E-6	-1.858E-6	-1.208E-6	-2.988E-6
RAINFALL	.000	-1.837E-5	8.202E-5	-1.523E-5	-1.888E-5
MRKTDAY_P1	.241	-.032	-.002	.002	-.033
R2	.046	.196	.053	.040	.035
F-statistic	1.451	6.720	2.147	1.428	1.546
Significance	.198	.000	.049	.205	.164

Variable Definitions

Variable name	Variable specification
POP2000	Population of locality in 2000
TT_CITY10K2	Time in minutes to the closest town of over 10,000 population
TTACCRA2	Travel time in minutes to Accra divided
DIST_FEEDER2	Distance to nearest feeder road in meters
RAINFALL	Average annual rainfall in millimeters per year
MKTDAY	Variable = 1, if prices were recorded on a market day