Appendix A

Purpose: This appendix provides an overview of the geographical demarcation techniques implemented in SCIP to assist in-field survey techniques in the selection and identification of target households. Through the use of satellite imagery, computer-assisted superset generation, support vector machines and automated image manipulation, we are able to generate accurate georeferenced maps that contain statistically-selected target houses.

Maps

The primary source of information that provides location and identification characteristics for this part of SCIP is a collection of satellite images, obtained from Digital Globe's 50 and 60-cm archive. These images provide an image resolution of 50 and 60 centimeters per pixel, respectively. At this resolution, man-made objects such as small dwellings, as well as larger types of vegetation, such as trees, become readily apparent and easily distinguishable. Figure A1 shows a sample of this map. The maps were acquired through one of Digital Globe's reseller partners, eMap International, which provided significant educational and institutional discount prices.

Figure A1 - Sample satellite image map

Each map was priced at 11.2 USD per square kilometer, with a minimum order of 10 km^2 of contiguous space. The necessary images were ordered by creating individual polygons that met Digital Globe's minimum area and feature size requirements (vertices greater than or equal to 2km, and fewer than 1000 vertices per polygon) based on georeferenced Enumeration Area (EA) shape files remitted to VIGH from Mozambique's National Institute for Statistics (INE.) eMaps returned satellite coverage of roughly 58 percent of the requested area, meaning that not all desired coverage of the individual enumeration areas was possible.

Identification

In order to provide interviewers on the ground a reasonable selection for their interview targets, the use of satellite maps performed two functions. The first is to provide the interviewing team a geographical reference map with familiar features and landmarks, based directly on the image data contained in the satellite maps. The secondary objective is to provide a subset of 15 target households in a statistically-random manner that allows the interviewers to conduct the survey with as little sample bias as possible. The second point represents the key technical challenge, and will now be covered in greater detail.

In order to effectively process a large number of possible households over a large sample area specified in the enumeration areas, there exist two methods for image analysis that allow for the production of maps containing the selected subsample of houses: 1.) Partially-computer-assisted, in which a user selects houses through a visual inspection process, and 2.) Fully-computer-assisted. Accuracy is the key goal in the selection of the full set of houses in each image; to reduce the possibility of bias in the subsample, two objectives must be met. The first is that all the dwellings that are in fact dwellings are properly identified such that as many of the actual dwellings in the map are present in the computer's list of identified dwellings. Secondly, the subset generated from the full set must be, at very least, quasi-random. In each analysis method, the randomized selection of the subset of dwellings for evaluation was done by inputting a list of identified households and generating a quasi-randomized permutation list through the use of a Pseudo-Random Number Generator (PRNG.) By using the input identifiers in a random, non-repeating order, the first 15 households in the list are selected as a quasi-random subset of the full list of houses.

Partially-assisted method

The partially-assisted identification involves the use of a computer program that was specifically written in MATLAB, an interpreted, functional programming language, for this purpose. The user must first generate a map in ArcGIS by importing the correct georeferenced satellite map into a corresponding regional map, with an enumeration area overlaid. The map is exported in the Tagged Information File Format (TIFF) and imported into the MATLAB program. A moving window is presented to the window in which he or she must click off user-identified houses. Upon completion of the identification in that window, a user input directs the computer to display the next portion, a 50% horizontal overlap, until the end of the horizontal width of the image is reached. A 50% shift on the vertical axis is then effected and the horizontal panning begins anew. Such display techniques reduce the chances of user error, in which a house may be missed on first inspection.

Each mouse click adds a normalized coordinate point to the program's "identified dwellings" buffer. This forms the location of known dwellings, which is the superset from which the selected subset is formed. Upon reaching the end of the image, which is logically the lower-

right-hand corner (LRHC) of the image, the computer performs the randomized selection as mentioned earlier. Those locations are then circumscribed by a 20-by-20-pixel crosshatch to mark the target dwelling. The resultant image is saved and then sent to post-processing.

Fully-assisted method

The fully-computer-assisted method does not require the user to select identified locations manually. Instead, the image is analyzed by an algorithm and candidate dwellings are identified and their locations stored in a buffer, as with the partially-assisted method. In particular, we have implemented the use of a Support Vector Machine (SVM) classifier. A support vector machine is a supervised machine learning algorithm that classifies inputs into binary categories (i.e., two categories). The SVM forms high or infinite-dimensional hyperplanes which classifies the input data. As inputs to the SVM are training data and a categorization matrix, which informs the classifier to which of the two categories each training datum corresponds.

This input is used to train the SVM classifier such that the actual testing data set inputs can then be analyzed and reported. Strictly speaking, an SVM classifier operates on scalar numbers that can be used to create support vectors that define the inputs. The support vectors define the hyperplanes. The operation of a support vector machine is dependent therefore on the quality of the training data, as well as the kernel used to create the support vectors and respective hyperplanes. The reader is encouraged to read Hsu et al[1] for a more through and mathematically-robust description of SVMs.

In our case, the training dataset is generated by manually identifying household locations on a target satellite map. In fact, this is done automatically during the partially-assisted method; selected dwelling locations are tagged, and 10-by-10-pixel regions are saved into training data images. These images are then analyzed to extract their color information. For each band in the RGB color space, we analyze the mean and median colors for the 100-pixel grid. This forms the basis for input data to the SVM. This gives us 6 feature sets on which the SVM can be trained. As a result, the SVM technique described here generally focuses on color information presented in the satellite map.

As multiple dwelling types exist with different color variations, the classifier must be run with different iterations of training data to identify the different houses. Note that the “not house” datatype is the same across all classification categories. Thus, one pass of the SVM focuses on thatched-roof huts, while a second pass may focus on sheet metal-roofed houses, and so on. Each pass of the SVM aggregates data into the identified houses buffer. As a result, the aggregated maps can have an identification accuracy of close to 98%. However, this is dependent on the

[1] Hsu, Chih-Wei, Chang, C., Lin, C., "A Practical Guide to Support Vector Classification", National Taiwan University, Taipei, Taiwan, 2003, Revised 2010

uniformity of the satellite image's lighting and exposure conditions. As some areas are large, this is not always the case, and dwellings may be missed. For this reason, the bulk of the maps are generated with the partially-assisted method. The results from either method looks similar from a visual perspective.

Post-processing

Once the images have selected subsamples, they are tiled into ninths (3x3 image matix), and legends and titles are superimposed. By using a uniform naming strategy, the labeling process reads in a file containing the names of the enumeration areas, preceded by a 10-digit code containing information about the district code and enumeration area number. This allows the labelizer to automatically label the input images without requiring manual intervention. In addition to the name of the EA, a compass heading and a tile map is presented, allowing the interviewer to identify clearly to which part of the main image the tiled sub-image corresponds. An example is seen in Figure A2. The resultant maps are compressed into a spanned archive and uploaded to a file repository where it can be downloaded and subsequently printed for use in the field.

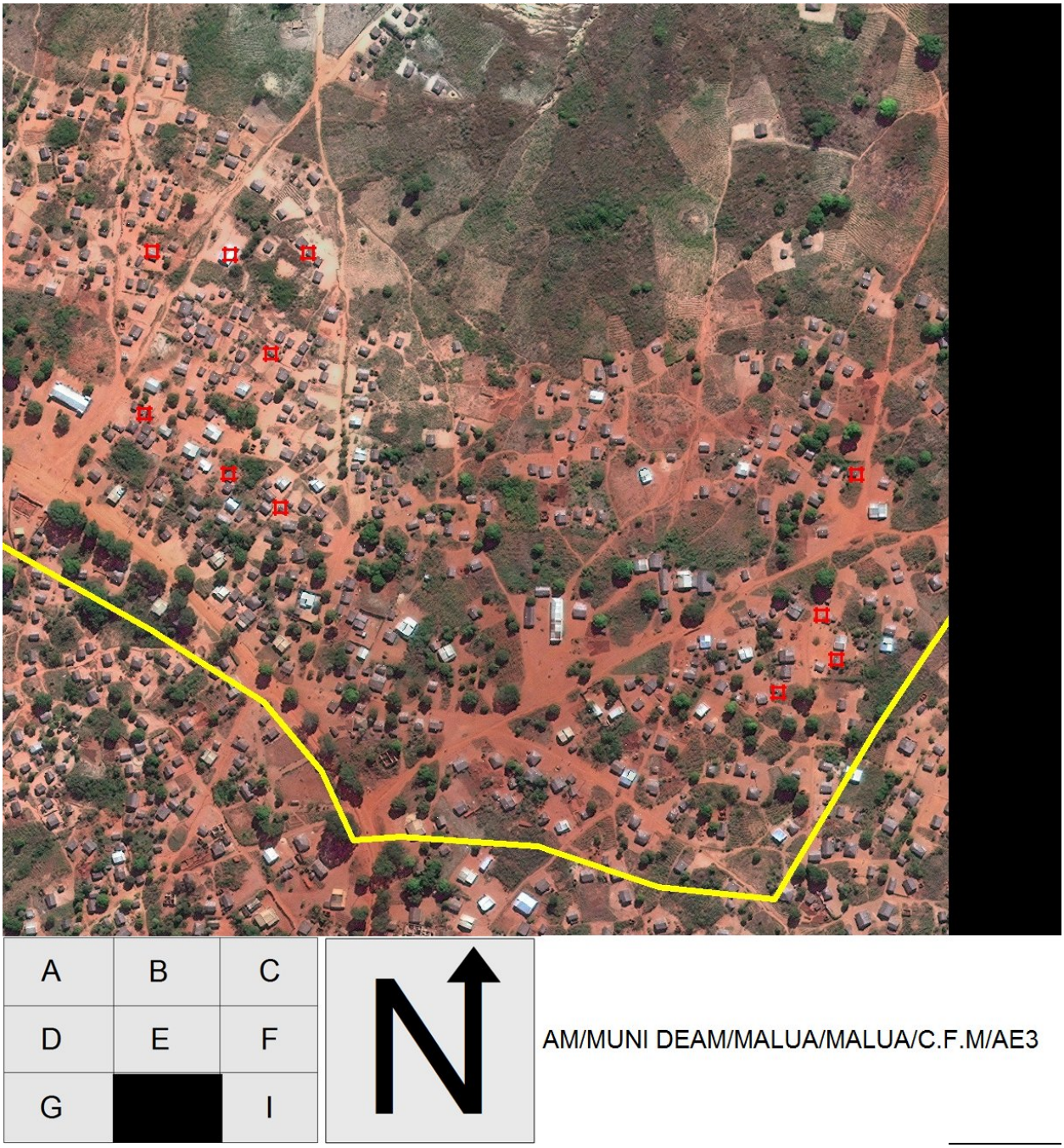

Figure A2 - Selected target houses in a sub-image