# Supplement to Kianda Emerging: Heuristics for Synthetic Population Creation

By Tom Pike

## Foundational Data:

As discussed in the paper, Dr. Stefano Marras provided the data of the Map Kibera project. The data consists of detailed shapefiles of the Kianda neighborhood in Kibera slum, Nairobi Kenya. The data file contains 48 categories for each structure. Dr. Marras and his team surveyed every structure in the Kianda neighborhood and recorded their results at the substructure level, with each substructure record documenting physical characteristics, population characteristics and business characteristics. The physical characteristics of the house included information such as construction material (e.g. wood, sheet metal etc), as well as infrastructure information specifically toilet, water and electricity. For the population characteristics, the data file documented the number of people and family type by category. The population number consisted of total women (age 18 and over), total population and total children (17 and younger). The program calculated the males by subtracting the total people minus the women and children. The family type consisted of five categories, single, single parent, married, polygamous, and family elsewhere. The data also contained business categories to include rent information (mean, min and max), if a business was in the structure and what type of business (e.g. service, retail etc). A complete list of all the categories is available on the GitHub page. Figure 1 is an example of the data.

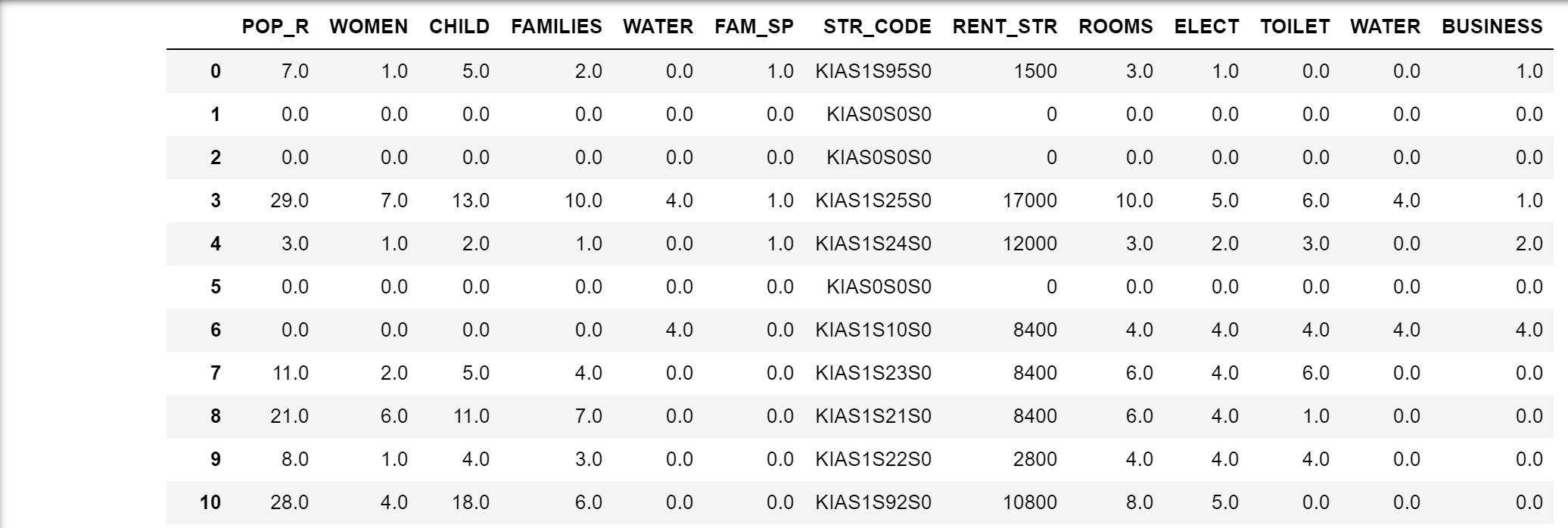


Figure 1: Example of data structure from Kianda data set.

The creation of the synthetic population took place in four phases. The first phase created families with rent and number of people by adult gender and children. The information used in this phase was based entirely on the dataset, while the data from phase two relied on data from the data file and aggregate statistics of Kianda. The second phase gave each family business ownership, infrastructure and tribe. Phases three and four relied on aggregate data from Nairobi slums to add the agent attributes. The third phase gave each member of the population an age range, and phase four gave each member of the population employment.

## Phase I: Family Characteristics and Rent

## Rent Calculation:

The information available for the Kianda neighborhood is highly detailed for each family but does not have specifics of the economic dynamics of the neighborhood. The rent details of the datafile then become a critical indicator for the income of each family and are used to calibrate pay and initialize wealth. Three data points from the data file, the rent max, rent mean and the rent min, were used to calculate the rent for each family. Due to the data provided and statistically small number of families in each structure (less than 20), the program uses the range rule to calculate the standard deviation of rent for the structure. The range rules is (rent max – the rent) /4. The program then created a normal distribution using the rent mean and standard deviation calculated using the range rule. In many cases, the rent is the same for all families so the standard deviation is 0 and all rents are the same.

The code trace for this process is (1) population folder, (2) dbfread module, (3) calc\_rent function.

*What is the implication for the six children in structure "KIAS2S72S8” with a rent of 1/6 800 for decision and food?- orphanage?*

## Children Assignment:

The challenge with the children is to disperse a reasonable number of children across the families within the structure. To do this a Poisson distribution of 0.95 lamdba was used to create a vector of children per household. Randomly adding or subtracting from values in the vector then corrected the distribution to make it equal to the actual number of children in structure. The assignment process of children to families then assigned based on the number of parents in the structure. In some cases, the vector had too much dispersal, meaning the number of children for a family was too low. To account for this situation values in the array were combined until all family types with children, had children. In addition, there were some cases were the children were orphans or at least were living with single adults not considered their parents and were under the age of 18. So, a structure had all single adults and an array of children. In this case the children were each assigned to their own family.

The code trace for this process is (1) population folder, (2) dbfread module, (3) calc\_children and row\_to dict functions.

## Adult Assignment:

For the adult vector, the program divided men and women among the family types. To create the adult vectors, the program simply iterated through family type assigning a male or female until there were no adults left to assign. The function only allowed one male and female for each married couple and one male and two females for each polygamous relationship. Depending on the breakdown of males and females a polygamous couple could have more females added later. The program did not allow for one female and multiple males as this is not an observed practice.

The code trace for this process is (1) population folder, (2) dbfread module, (3) divvy\_men and divvy\_women and row\_to dict functions.

## Family Creation:

To create families the program merged the child and rent vectors with the adult vector. With these three vectors, families were created based on the family type and rent vector for each structure. For example, a single adult male family would consist of a family number, one male adult and a rent scalar. Or, a married couple would consist of a family number, one adult male, one adult female, a child scalar and a rent scalar. This process was repeated for each house. The result was a family creation for a substructure which had the same features of family type and number of people as documented in the dataset.

The code trace for this process is (1) population folder (2) dbfread module (3) row\_to\_dict function

## Phase II: Tribe, Infrastructure, Business and Ownership

## Infrastructure, Business and Ownership Assignment:

The second phase of the synthetic population creation consisted of assigning family owners, businesses, infrastructure and tribes. The datafile defines ownership as someone who owns the structure in which they live and businesses are commercial enterprises within the structure, which may or may not be owned by the residents. The program uses ownership and businesses as a heuristic for wealth in the structure. As there was no indication of which family was the owner, the program assigned ownership randomly to families within the house. For infrastructure there may be one room in a structure with electricity or water. To account for ownership if there was infrastructure the program assigned ownership of the infrastructure as those who were also the owners of the building. If the amount of infrastructure exceeded the number of owners, the remaining infrastructure was assigned randomly.

The code trace for this process is (1) population folder, (2) build module, (3) add\_busi and add\_infra functions.

## Tribe Assignment:

The specifics of the tribal data are not disseminated due to the potential use for targeted ethnic violence. However, the aggregate statistics of tribes in Kianda are available (Marras, 2009). The economic dynamic of Kenya, however long term dependencies of ethnic stratification would make a random assignment of tribes inaccurate. It is well understood the Kikuyu tribes are the primary owner is Nairobi and the Kibera slum (Dafe, 2009; de Smedt, 2009; Dercon & Gutiérrez-Romero, 2012; Marras, 2008). Therefore, tribal assignment is purposely biased based on this information. To create the tribal assignment, the program creates an array of all known tribes in Kianda with the percentages match to the number of people in the population. The program then assigns owners first to the Kikuyu tribe. As the ownership exceeds the number of Kikuyu, the program assigns tribe randomly to the other owners. The program then randomly assigns tribe affiliation to rest of the population as no there were another other clear distinctions discussed in the literature. Due to the lack of information on the integration of tribes the program employed two additional heuristics for tribal assignment. First, every family was of the same tribe. Although this is not strictly true in Kenya, intertribal marriages are a minority (citation). In addition, there was no distinction made between structures, so any tribe can live with any other tribe. Although this is also not strictly true, as the model does not account for dynamics such as two families in the same structure sharing the challenges of food consumption or caring for children to young for school. The impact of the overmixing of tribes is negligible. In addition, future iterations of the model plan to replicate the riots which occurred in 2009 and an essential piece of this dynamic is the tribes segregating along racial lines. A mixed population therefore will not negatively impact the dynamics of the model when recreating patterns of life, while allowing for further validation when the model is extended to include riots.

The code trace for this process is (1) population folder, (2) build module, (3) add\_tribes functions.

## Phase III: Age and Child Gender

Phase III represents the complete transition from the detailed data file to assigning population attributes based on aggregate statistics.

As the model represents daily life over a week period the aging of agents is not a consideration. The primary need for age was to determine what role the agent serves within the family. Are they able to work? Do they need to be nurtured? Can they go to school? Based on this need agents were grouped into asymmetric categories as follows: 0-4, 5-9, 10-14, 15-18, 18-25, 25-50 and 51+. With these broad categories, the program created a reference dictionary which used the population of the previously created population and assigned ages based on a 2012 study of informal settlements in Nairobi conducted by the African Population and Health Research Center. With this information, the program then assigned ages heuristically based on the family situation. For example, a large family with six children would have adults in the 25-50 range and have children ranging in all four categories. Single people were more likely to be either younger 18-25 or older 51+. Children on their own were more likely to be 14-18 and the program did not allow children 0-4 to be on their own. In this phase the program also assigned gender to the children. Gender was assigned randomly based on the age-related statistics. The program iterated through all families assigning age and matching with known population statistics of Nairobi slums.

The code trace for this process is (1) population folder, (2) agentize module (3) make\_agents, make\_males, make\_females, make\_children.

## Phase IV: Jobs, Physiology and Maps

In phase IV each agent is given a job, sleep, caloric and hydratiom requirements and straight-line distance to different types of shops within the Kianda neighborhood.

## Job Assignment:

The program assigns employment by leveraging the attributes of the agents and matching them to documented statistical patterns of employment within the population. Employment is grouped into six categories: regular, casual, fixed, student, child labor, and business owners. The rate of employment in each area is based on a survey of Nairobi informal settlements conducted by Suyami Gulyani and his team and published in 2006. From this information and the population size, the program creates employment targets for each age group and gender. The program then assigns employment heuristically to each family. For example, the program assigns married agent males with high rent and good infrastructure regular employment, where wives are typically unemployed. Unemployment is also higher for younger individuals and the program prioritizes fixed incomes for the 51+ age group. All children form 5 -18 are employed as students and some 18-25 year olds are also employed as students.

The code trace for this process is (1) population folder (2) jobs module

## Physiology:

Physiological needs are assigned based on age and gender for daily caloric requirements, sleep requirements and hydration. For each category of age and gender the program creates a normal distribution with targeted requirements as the mean and an appropriate standard deviation for the category. The food requirements were based on the total recommended food requirements of the U.S. Department of Agriculture using the basal metabolic rate (calories used during rest) (Citation). The water requirements were based recommendations through studies form the National Academy of Sciences (citation), while the sleep requirements are based on sleep recommendations from the National Sleep Foundation (citation). This approach allows for the creation of a heterogenous population of physiological sleep, food and water needs.

The code trace for this process (1) population folder (2) phenotypes module

## Maps:

The final portion of agent creation is to give each agent a map to all the food shops and casual labor recruitment sites in the neighborhood. For simplicity these shops are broken into five categories and assigned heuristically based on type and location. The shops (designated as a commercial business in the datafile) are assigned one of four categories, 24\_cooked (street food vendor), 24\_uncooked (grocery store), cooked (street vendor less then 24 hours) and 24 uncooked (grocery less than 24 hours). In addition, several cooked locations are identified as an informal gathering for casual or unemployed individuals. These locations are assigned heuristically based on their location in Kianda. 24 hours businesses are closer to the northern boundary of Kianda to gather more traffic from Nairobi city. This is also the location of the casual labor collection. The non-24-hour stores are located in the interior of Kianda and would serve the local population instead of trying to gain more street traffic from outside the neighborhood.

Due to the density of Kianda and lack of clear roads throughout the neighborhood, calculating specific routes in impractical. Therefore, for each agent the straight-line distance from the centroid of their house to the centroid of the shop structure is calculated. The list is then sorted nearest to furthest and stored as an agent attribute. On drawback is shops or locations outside Kianda cannot be take into effect, making the neighborhood in this simulation isolated form the outside world.

The cord trace for this process is (1) population folder (2) shops module, movement module and familyagents module.

The synthetic population for the Kianda Emerging model consisted of four phases. From the data file a detailed population was created at the family level per structure. The program then heuristically mapped aggregate statistics of Nairobi slum populations, to include employment, wealth, age, physiological features and shop distances, to the population. The main module which initiates the synthetic population creation is the Main\_synthetic module in the population folder. This module will then output pickle files of the population as well as a plain text data file.