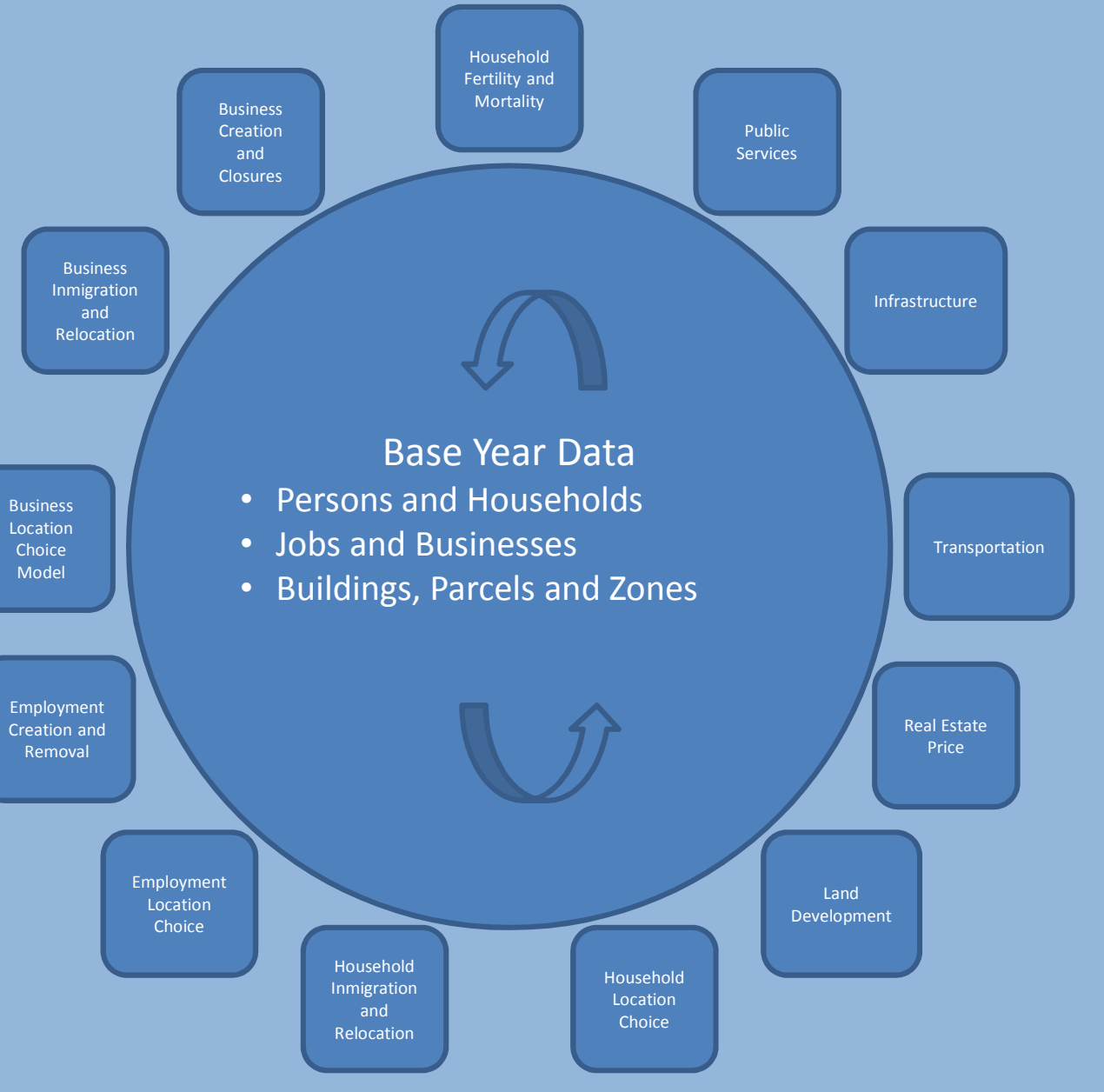
CONCEPT

Amongst all species, humans have a highly sophisticated, remarkably unique and individual set of rules which govern our behavioral patterns. The ultimate aggregate exhibition of these behavioral patterns is the phenomenon of urban dynamics and these interactions within the larger biosphere where the city resides. A countless number of activities comprise this complex urban system, implemented as individual choices taking place along different temporal scales as spatially distributed throughout the city. Some location choices involve decisions with short term implications, such as travelling to work, school or shopping, while other decisions have a more lasting and long term impact, such as accepting a job, moving into a new home, or deciding to start a business. These two fundamental components, activities related to temporary locations (transportation uses) and activities related to permanent locations (land uses) are the fundamental building blocks for modeling urban dynamics. By understanding these two basic types of activities which occur within the city, a template for simulating the urban dynamics of a city can be constructed. From this basic urban simulation system, models of infrastructures (transportation networks, electricity, water & sewer, solid waste, stormwater and parks and recreation) and institutions (public health, education and safety) can be implicitly incorporated.

An urban simulation system is comprised of several disaggregate data sets, integrated variables and sequential models, which are used to predict the choices of persons, households and businesses in order to spatially simulate the potential land development patterns forming throughout the region. Individual models forecast demographic and economic growth, predict the probability that a particular household, job or business will relocate, and if so, which available new location it will chose, project real estate prices and simulate potential new land developments. Primarily through the use of probabilities derived from surveys and data describing the physical geography of the city, an urban simulation system executes millions of predictions about the decisions of agents interacting with each other as well as within their spatial environment. The foundation of this system is these local interactions, which result in an emergent global structure representing some of the likely spatial land use and development scenarios which could occur in the future.

Central to the creation of an urban simulation system is the development of a number of data sets used in the model system which represent the existing condition of the city at the beginning point in time of a scenario run. Agent and Location data sets spatially describe the physical, demographic, and economic geography in terms of each person and household, job and business, as well as building, parcel and zone. The person, household, job and business tables are typically synthetically generated datasets which describe agent attributes including how they choose where to live, work and conduct business. The buildings table describes every physical structure, which is typically available from high resolution aerial photography, while newer approaches enable recording buildings in three dimensions and improved inference of use as derived from the signature of the building envelope. The parcels table describes the fundamental unit of property subdivision and ownership where each building is located and generally is found at the property recording agency survey department but is generally neither digitized, geographically projected nor topologically defined. The zones table is used as a means to describe spatial data at higher levels of aggregation when higher resolution data is not currently available or needed. Agent and Location data sets serve as the fundamental input for the models as each one executes its series of functions, generates output and then updates these tables based on the projection results.

Exogenous allocations and endogenous probabilities provide the number of new agents projected for immigration to the city as well as the probability a new agent will be introduced (births), removed (deaths) or decides to move from its current location. Allocation data sets include estimates for the total number of new households (by size) and businesses (by sector) projected in terms of a low, medium and high demographic and/or economic growth rate for each simulation year. Relocation data sets include the probability of a single household and business relocating within a given year, and are typically derived from a specialized survey. Statistical Model data sets specify how to spatially calculate a given variable used in a regression or location choice model, as well as the actual parameter used in that calculation.



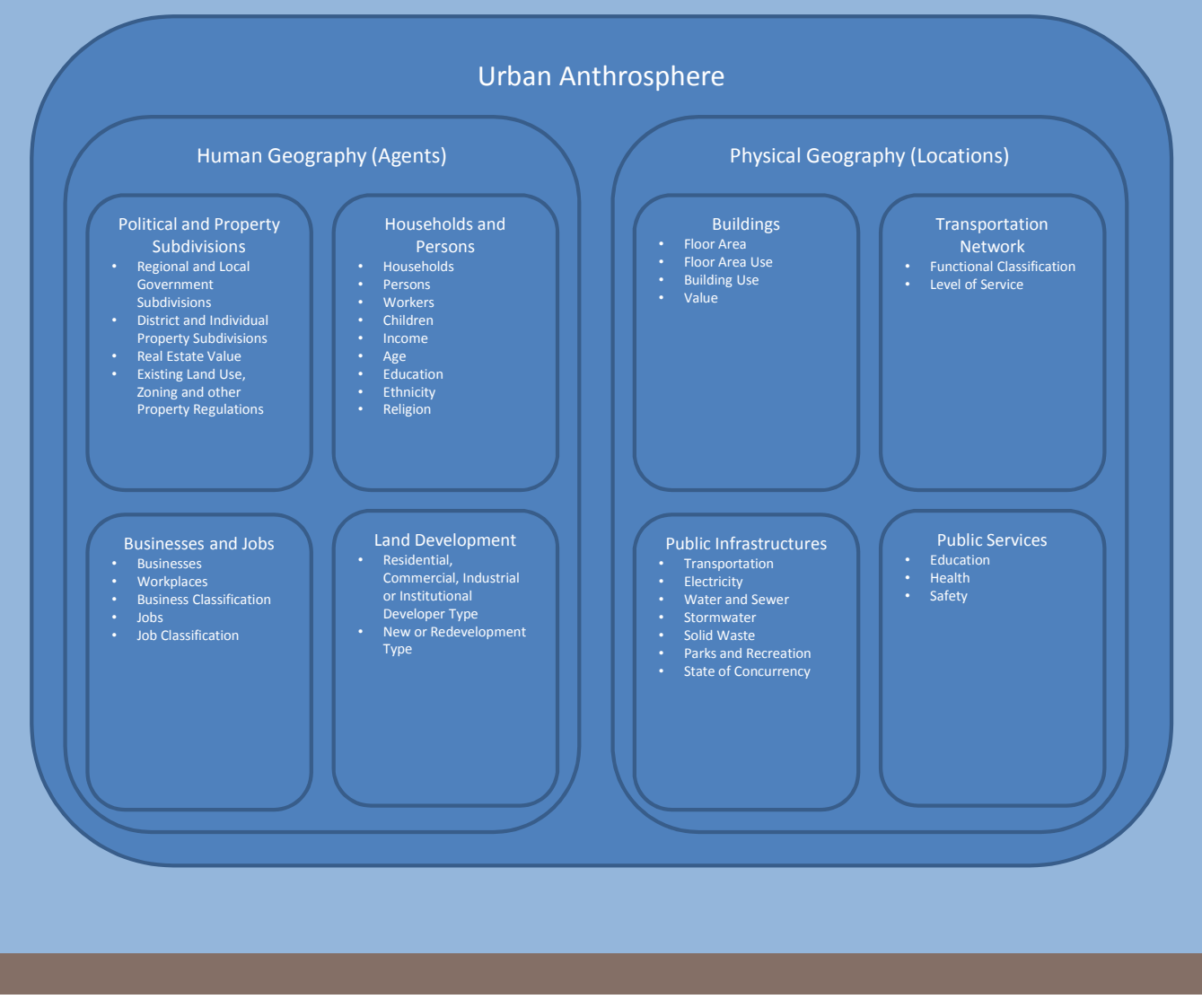
Base Year Data
- Persons and Households
- Jobs and Businesses
- Buildings, Parcels and Zones

Household Fertility and Mortality
Public Services
Infrastructure
Transportation
Real Estate Price
Land Development
Household Location Choice
Household Inmigration and Relocation
Employment Location Choice
Employment Creation and Removal
Business Location Choice Model
Business Inmigration and Relocation
Business Creation and Closures

An urban simulation begins its runtime environment by loading all base year data into a cache for use by each of the models. This model system for predicting land use and development patterns begins with household and business allocation models, continues with the household and business probability models, and is then followed by statistical models for determining real estate value, household and business location choice, and land development and infrastructure demand. As each model completes its initial run, the base year data sets within the cache are updated to reflect the results. This process continues through each annual cycle until the simulation has reached is scenario horizon date.

The model system begins with the Household Transition Model, which estimates allocations for how many new households will move to the urban area during the initial year. The household transition model compares the actual number of households residing in the city with the household control totals table, subtracts the difference, and queues this number of new households for subsequent input to the household location choice model. At the beginning of the simulation, the urban simulation exists as thousands of uninhabited residential locations waiting to be occupied by tens of thousands of households; therefore, all of the synthetically simulated households from the households table are queued for input to the household location choice model during the initial simulation year. Beginning the following year and with all subsequent years, the Household Transition Model also randomly selects households for removal, when relocating households have chosen to move outside of the Metropolitan Area.

Following the household transition model, the Business Transition Model estimates how many new businesses will move to the city during the initial simulation year. The business transition model compares the actual number of businesses with the business control totals table, subtracts the difference, and queues this number of new businesses for subsequent input to the business location choice model. As with the household transition model, at the beginning of the simulation, the city exists as thousands of unoccupied business locations waiting to be occupied by tens of thousands of potential businesses; therefore, all synthetically simulated businesses are initially queued for input to the business location choice model. As local data becomes more readily available and reliable, synthetically generated business data is less likely to be needed, and the focus will turn to describing workers occupying each workplace.



Urban Anthrosphere

Human Geography (Agents)

Political and Property Subdivisions
- Regional and Local Government Subdivisions
- District and Individual Property Subdivisions
- Real Estate Value
- Existing Land Use, Zoning and other Property Regulations

Households and Persons
- Households
- Persons
- Workers
- Children
- Income
- Age
- Education
- Ethnicity
- Religion

Businesses and Jobs
- Businesses
- Workplaces
- Business Classification
- Jobs
- Job Classification

Land Development
- Residential, Commercial, Industrial or Institutional Developer Type
- New or Redevelopment Type

Physical Geography (Locations)

Buildings
- Floor Area
- Floor Area Use
- Building Use
- Value

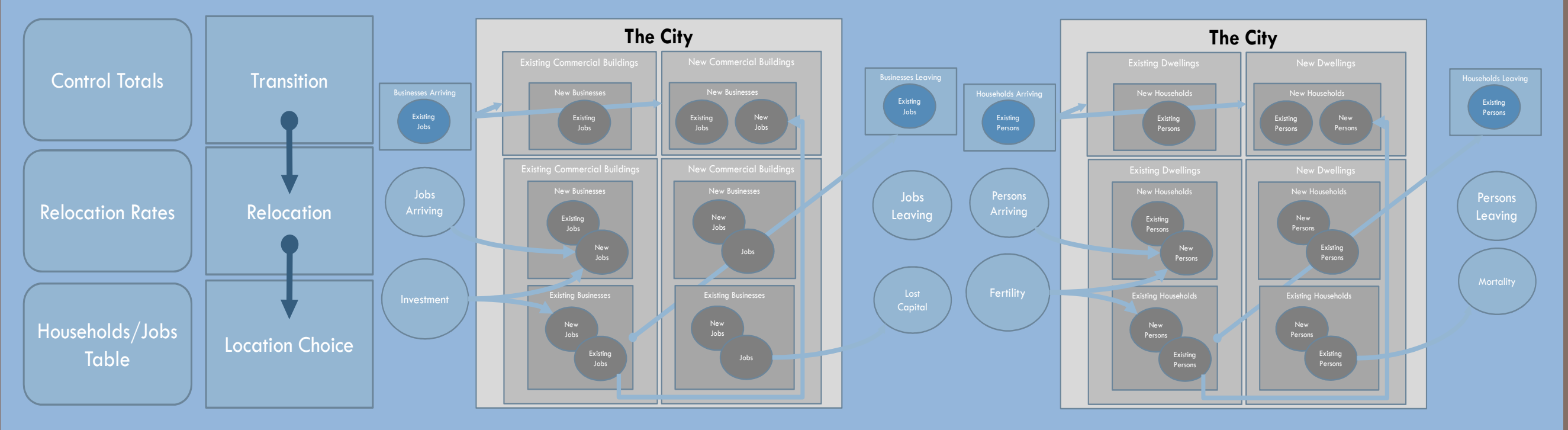
Transportation Network
- Functional Classification
- Level of Service

Public Infrastructures
- Transportation
- Electricity
- Water and Sewer
- Stormwater
- Solid Waste
- Parks and Recreation
- State of Concurrency

Public Services
- Education
- Health
- Safety

Once the business transition model has completed its run, the Household Relocation Model predicts the probability that a household will move from its current location or remain in place during that particular simulation year. The household relocation model determines the relocation probability for each household and then randomly selects a number to determine if each individual household has been chosen to move from its current location within that simulation year. As households are identified for relocation, they are combined with the new households moving to the region as allocated from the household transition model for input into the household location choice model. After the household relocation model, the Business Relocation Model is initiated in order to predict the probability that a job will move from its current location during that particular simulation year. The business relocation model determines which businesses will be scheduled for relocation based on the probabilities in the relocation rates table.

DATA



The Real Estate Price Model uses a linear regression model to predict real estate value as dependent to the spatial variables found in the real estate price model specifications table. These six variables have been specified to describe building, household, and employment attributes by zone for each building type and then used to calculate the average_value_per_unit attribute in the buildings table. Model coefficients found in the real estate price model coefficients table are generally estimated from tax assessor or real estate professional's data. The Household Location Choice Model predicts in which particular residential structure a new household (from the Household Transition Model) or an existing household (from the Household Relocation Model) will be located. The household location choice model uses the six spatial variables found in the household location choice model specifications table to calculate the probability of a household selecting a particular location from a set of 30 to 50 available dwelling units. Once a household chooses a location, the buildings table is updated to reflect occupancy of that particular residential unit. Model coefficients parameterizing building area, population density, job density, number of households, work travel time and household income are found in the household location choice model coefficients table.

The Business Location Choice Model predicts in which particular institutional, industrial or commercial structure a new job (from the employment transition model) or an existing job (from the employment relocation model) will be located. The employment location choice model uses the three spatial variables found in the employment location choice model specifications table to calculate the probability of a job selecting a particular location from a set of 30 to 50 available business locations. Once a job chooses a location, the buildings table is updated to reflect occupancy of that particular workplace. Model coefficients parameterizing population density, number of jobs, and work travel time are found in the employment location choice model coefficients table.

The Infrastructure demand model projects infrastructure consumption or demand per parcel from a linear regression model which is typically estimated from historic monthly consumption data. Aggregated demographic and economic attributes can be used to parameterize to projected demand, as dependent to variables describing the number of households and their demographic composition as well as the total number of businesses and jobs and their economic composition. Regression coefficients are applied to aggregated results from the Household and Employment Location Choice models to project total demand.

Close-to-Reality Synthetic Population

Synthetically generated population data is generally an important first step in running microsimulations or agent based models used to predict urban dynamics and/or transportation activities. Microsimulation models often attempt to reproduce the behavior of individual persons, households or firms over the course of several years in order to quantitatively and qualitatively visualize potential scenarios which could occur as well as their associated costs and benefits. In order to reduce potential prediction error, using population data that most closely reflects the existing population inhabiting the geographic area of study is desirable. Generating this synthetic population has typically been achieved by either repeatedly drawing samples from sample data or using the iterative proportional fitting method (IPF), a common method employed by transportation models. By simulating the existing population data, a realistic framework for comparing the implementation of different policy cadres (business as usual, weak sustainability, strong sustainability) under different growth scenarios (low, medium or high economic or demographic growth rates) can be projected.

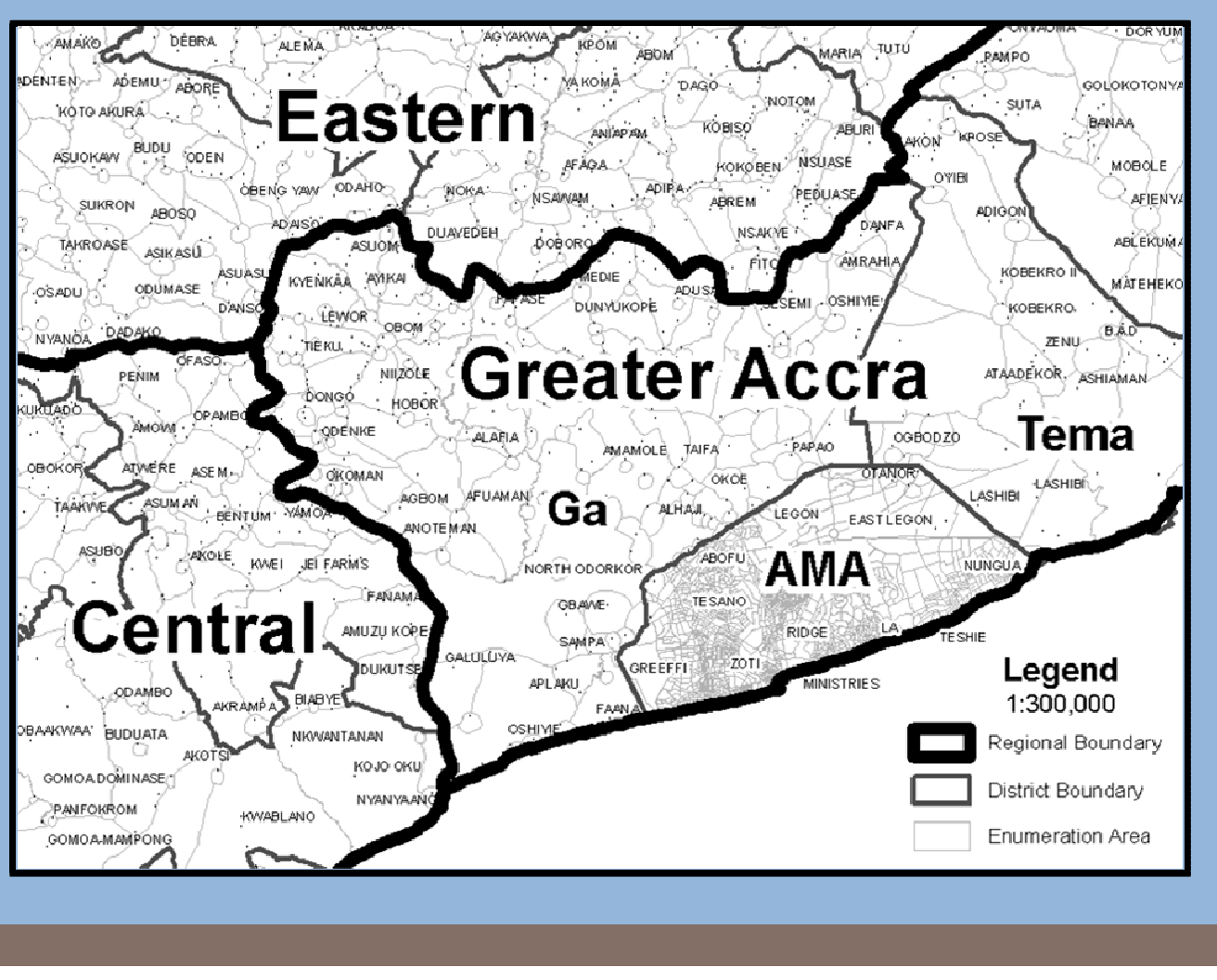
One of the advantages of synthetic data is its cost effectiveness when compared to comprehensive and detailed population data, which is in effect nearly impossible to obtain for every living person inhabiting a significantly sized urban geographical area. Additionally, generating synthetic data serves to meet the need for observing statistical disclosure limitations. Generating synthetic data not only presents the researcher with the base year data needed to simulate different potential urban simulation scenarios it also presents the public statistician with the means for releasing these base year datasets for practical application, while protecting rights to privacy as well as maintaining the likelihood of receiving authentic data from individual survey observations.

In order to generate a synthetic population from a sample such as the GLSS5, several conditions need to be met. First the actual size of regions and strata must be reflected in the survey weights. Secondly, marginal distributions and interaction between variables should be reflected correctly, while heterogeneities between subgroups, especially regional aspects, should be allowed. Finally, pure replication of units from the underlying sample should be avoided, as this generally leads to extremely small variability of units within smaller subgroups. Following these conditions, the synthetic data should include univariate distributions overall and in subpopulations as well as multivariate relations among the variables. In order to meet these conditions, multinomial logistic regressions can be used to predict possible outcomes of a dependent variable from probabilities derived from a given set of independent variables. Also used in synthetically generating the household structure, categorical and continuous variables for Great Accra is the conditional probability distribution which is the probability distribution of variable Y when variable X is known to be a particular value.

The *Ghana Living Standards Survey-Round Five* (GLSS 5), like earlier ones, focuses on the household as a key socio-economic unit and provides valuable insights into living conditions in Ghana. The fifth round of the GLSS was conducted by the Ghana Statistical Service (GSS) from 4th September 2005 to 3rd September 2006. A nationally representative sample of 8,687 households in 580 enumeration areas, containing 37,128 households members were covered in GLSS5. Detailed information was collected on demographic characteristics of respondents and all aspects of living conditions including health, education, housing, household income, consumption and expenditure, credit, assets and savings, prices and employment. For the purposes of this work, sections on Demography, Education and Employment were used.

The synthetic population generation cannot be applied to the GLSS5 directly if the data includes missing attributes from observations. While in the univariate case the observations with missing information could simply be deleted, this can result in a severe loss of information in the multivariate case. Multivariate observations usually form the rows of a data matrix, and deleting an entire row implies that cells carrying available information are lost for the analysis. Instead of deleting observations with missing values, it is better to fill in the missing cells with appropriate values, which is possible with multivariate data sets.

Many different methods for imputation have been developed over the last few decades. While univariate methods replace the missing values by the coordinate-wise mean or median, the more advisable multivariate methods are based on similarities among the objects and/or variables. A typical distance based method is k-nearest neighbor (KNN) imputation, where the information of the nearest k>=1 complete observations is used to estimate the missing values using the Aitchison distance for measuring compositional datasets. While kNN is numerically stable it has some limitations. First the optimal number of k nearest neighbors needs to be determined, by randomly setting observed cells to missing, estimating these values and measuring the error. Secondly, kNN imputation does not fully account for the multivariate relations between the compositional parts, which are only considered indirectly when searching for the k-nearest neighbors. A next step in this process will be to apply a model-based imputation procedure which relies on a more realistic estimation of the multivariate data structure.



MODEL

Application of SimPopulation to the GLSS5

The household structure is simulated separately for each combination of stratum k and household size l. First, the number of households is estimated using the Horvitz-Thompson estimator, indexing the set of households in stratum k of the survey data with household size l, in accordance to the corresponding household weights. To prevent unrealistic structures in the population households, basic information from the survey households is resampled. Using the R package simPopulation, we start our analysis using the function simStructure().

```
gamaP <- simStructure(gamal, hid = "hhid", w = "weight", strata = "cluster", additional = c("age","sex"))
```

Additional categorical variables are simulated using the simCategorical() function which estimates the conditional distribution with multinomial logistic regression models for each stratum using survey indices to fit responses and predictors while incorporating survey weights. In order to reduce computation time, data is aggregated or the number of categories are reduced into categorical groups. The argument basic specifies existing generated variables found in the household structure, while the argument additional specifies the variables to be simulated in this step.

```
basic <- c("ageCat","sex","hsize")

gamaP_Cat <- simCategorical(gamal, gamaP, w = "weight", strata = "cluster", basic = basic, additional = c
("nation", "ethnic", "religion", "highest_degree", "occupation"))
```

Next the function simContinuous() is used to simulate the variable annual income with the basic argument modified to include additional predictor variables. This approach is able to handle semi-continuous variables, i.e. variables that contain a large amount of zeros, which is true with regard to the variable annual_income in the GLSS5. Following the approach used for simulating categorical variables, the continuous variable is discretized by breakpoints and zero becoming a category of its own. Multinomial logistic regression models are then fitted for every stratum k separately, as previously described in order to simulate the continuous variable. Finally the values of the variable are generated by random draws from uniform distributions within the corresponding categories.

```
basic <- c("ageCat","sex","hsize", "nation", "ethnic", "religion", "highest_degree", "occupation")

gamaP_Cont <- simContinuous(gamal, gamaP, w = "weight", strata = "cluster", basic = basic, additional = c
("annual_income"))
```

Evaluation of the Simulated Synthetic Population of Greater Accra

In this section the relationship between categorical variables, including variables defining the household structure are evaluated using contingency coefficients. Pearson's coefficient of contingency is a measure of association for categorical data as obtained from the sample and the synthetic Greater Accra population. The relative differences are negligible in all instances with the correlation structure of the simulated population being very close to that found in the GLSS5 after application of kNN imputation. The result is the synthetic generation of 3,111,779 persons being described by the variables for household size, age, sex, religion, educational attainment and occupation, while the variables for nationality, ethnicity and household income will be subsequently included. This synthetic population is ready to be used as the base year data set in an urban simulation system or transportation model.

Pairwise Contingency Coefficients from the GLSS5 after Imputation

	sex	hsize	religion	highest_degree	occupation
age	0.07684275	0.3885837	0.2888835	0.7238775	0.7198963
sex	NA	0.1370375	0.1008813	0.1912632	0.3971899
hsize	NA	NA	0.5395080	0.3009451	0.5537457
religion	NA	NA	NA	0.3287186	0.5677181
highest_degree	NA	NA	NA	NA	0.8378192

Pairwise Contingency Coefficients from the Synthetically Generated Greater Accra Population

	sex	hsize	religion	highest_degree	occupation
age	0.07684068	0.3889738	0.2905272	0.7201150	0.7146574
sex	NA	0.1374701	0.1011180	0.1932895	0.3778930
hsize	NA	NA	0.5396357	0.2947910	0.5464265
religion	NA	NA	NA	0.3288038	0.5583367
highest_degree	NA	NA	NA	NA	0.8280430

Residential Mobility

Coefficients of residential mobility have been estimated from the work of Monique Bertrand and Daniel Delaunay, Residential Mobility in the Greater Accra Region: Individual and Geographical Differentiations. Bertrand et al. provide a general synthetic model of the factors of variation involved in the residential mobility in Greater Accra as well as a mean rate of mobility for a number of different neighborhoods. Bertrand et al. indicate Old Teshie, Lagos Town and New Fadema have mobility rates of 3.1%, 4.5% and 9.3% respectively, while Greater Accra as a whole has a probability of 8.6%. The areas of Lagos Town (New Town) and New Fadema are the closest in terms of proximity to Korle Bu, but are relatively small communities when compared to Korle Bu which is a district of nearly 200,000 persons. In fact Korle Bu incorporates a diversity of distinct communities which all exhibit somewhat different characteristics. The areas of James Town and Ussher Town are comparable to Old Teshie while Agbogbloshie exhibits characteristics typical of very low residential mobility, while other areas may be comparable to New Town or New Fadema. Presented in terms of the odds or hazard ratio in a Cox regression model, these parameters were transformed with and applied to the GLSS5. For the purposes of this study, the mean household mobility rate was set at 7.8%. It may seem counterintuitive that an increase in annual income (in GHCs) positively correlates to increased residential mobility.

$$\mu = \beta_0 + .077x_{\text{sex}} - 0.11x_{\text{age}} - 0.08x_{\text{cohort}} + 0.31x_{\text{edu1}} \text{ or } 0.49x_{\text{edu2}} + 0.07x_{\text{dwelltype}} - 0.34x_{\text{occupystat}}$$

Where:

μ = number of incidents

β_0 = constant

x_{sex} = sex

x_{age} = age

x_{edu1} = primary or junior secondary school versus non − educated

x_{edu2} = senior secondary school or University versus non − educated

$x_{\text{dwelltype}}$ = compound houshing versus flat or self − contained house

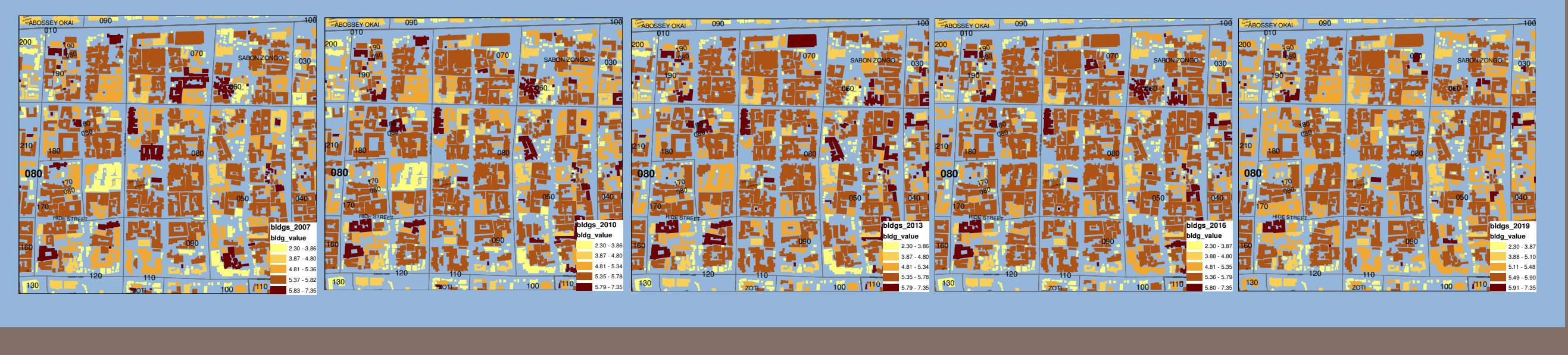
$x_{\text{occupystat}}$ = freeholder versus owner or tenant

Age/Income	0-35	36-249	250-599	600-959	960-1301	1302-1919	1920-3159	3160-100000	Mean
15-19	0.133013	0.131223	0.182112	0.126377					0.143181
20-24	0.123688		0.130773	0.133227	0.124518	0.128205	0.153337	0.180859	0.13923
25-29	0.102521	0.111597	0.114808	0.126122	0.124093	0.118106	0.130143	0.123715	0.118888
30-34	0.096336	0.087733	0.104155	0.089789	0.094839	0.099655	0.113142	0.111723	0.099672
35-39	0.081382	0.071712	0.074156	0.077072	0.085478	0.088004	0.097852	0.098497	0.084269
40-49	0.059396	0.070479	0.064743	0.067391	0.066962	0.070236	0.077083	0.080031	0.06954
50-59	0.055671	0.055082	0.053964	0.054074	0.05247	0.056731	0.062332	0.070263	0.057573
60-100	0.047353	0.049653	0.055031	0.050041	0.048988	0.054145	0.045664	0.058313	0.051149
Mean	0.08742	0.082497	0.09746775	0.090511625	0.085335429	0.087868857	0.097079	0.103343	

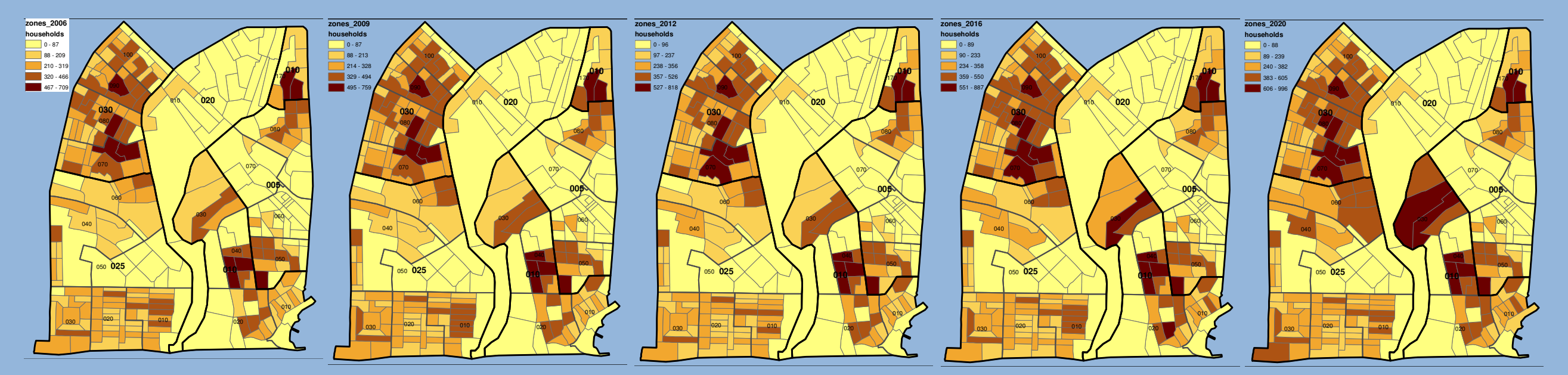
GAUSS Results

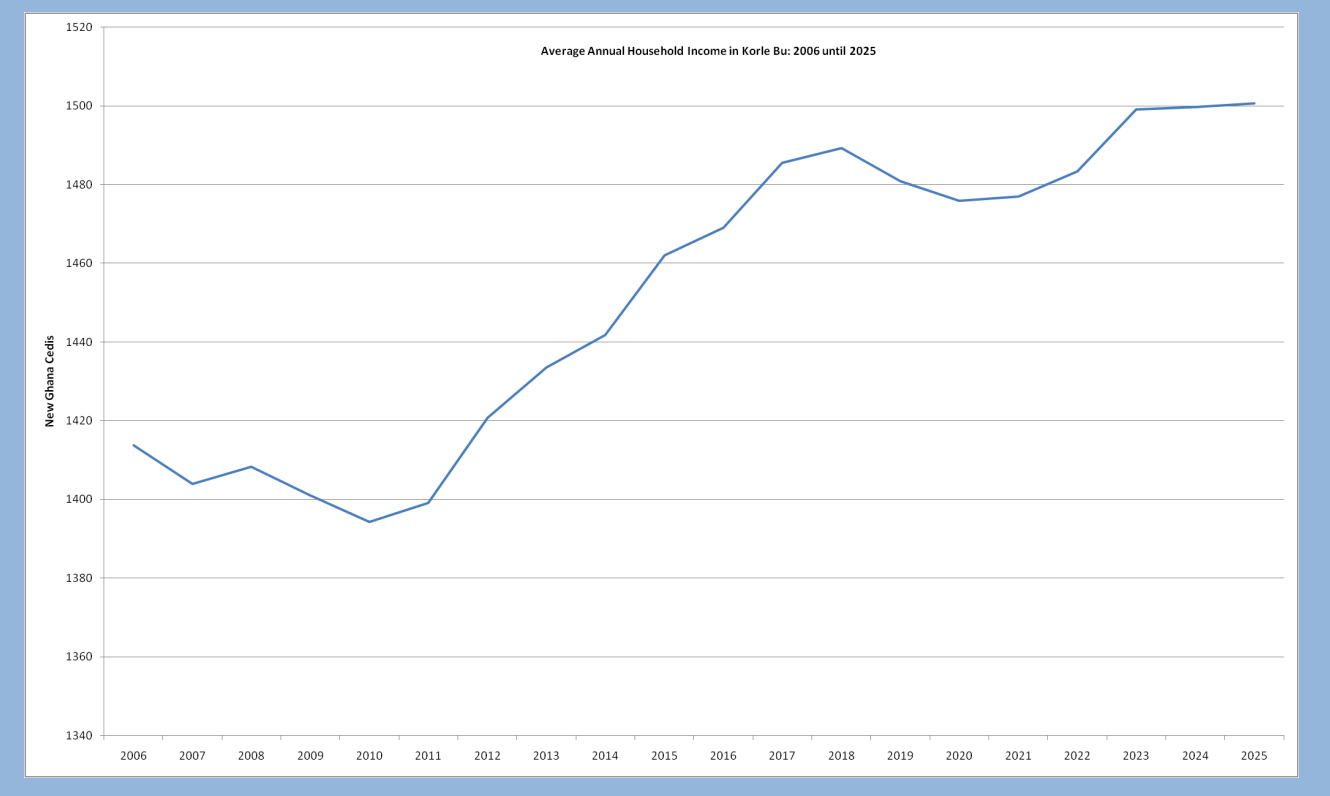
The Real Estate Price Model uses real estate prices as the indicator of the match between demand and supply of land at different locations for different land uses. The graphic presents building value results from the Real Estate Price Model (run in terms of a low population and economic growth rate scenario) in the area of Sabon Zongo at three year intervals beginning in 2007. Perhaps the most obvious result is that building values fluctuate only slightly in this area, indicating a relatively entrenched population. Considering the low household incomes and low mobility rates it can be concluded that the numerous multi-family dwellings are rentals which are income produces for landlords. Residential densities in this particular area can range upwards from 65 dwelling units/acre with average household sizes of 4 persons and including a remarkable amount of site coverage since nearly all residential structures are a single story in height (less than 20 feet).

The Household Location Choice Model illustrates the number of households located in each of the 280 zones throughout Korle Bu as predicted by GAUS-KB for the years 2006 to 2025 in terms of a low demographic and economic growth rate. This spatial representation of the model projection indicates the largest number of households will be located in Sabon Zongo with round 030-080 absorbing the largest number and demonstating the most significant growth during the 20 year period with an increase of approximately 3000 new housholds. The adjacent round 030-070, which is in Zoti, trends towards having the second largest number of households, and in general the block comprising these four rounds (Zoti, Sabon Zongo and Abossey Okai) exhibit a strong trend towards very high densities. The second hotspot where household population demonstrates large numbers and growth is in the area of Adedenkpo, particulary round 010-040 in the southern half, which is adjacent to Korle Dudor where large numbers of households also trend in the plots (zones) along the shared border. The model projects that in the year 2017, population growth will extend into the northern plots comprising Adedenkpo or the area often referred to as "Sodom and Gomorrah." As with the real estate price model analysis of Sabon Zongo, the Household Mobility simulations present questions related to population growth and if any real poverty reduction can be expected if business-as-usual in Korle Bu continues. Under the low economic and demographic growth scenario, average household income increases by only 100 GHCs (1400 to 1500) over the twenty year time span from 2006 to 2025.

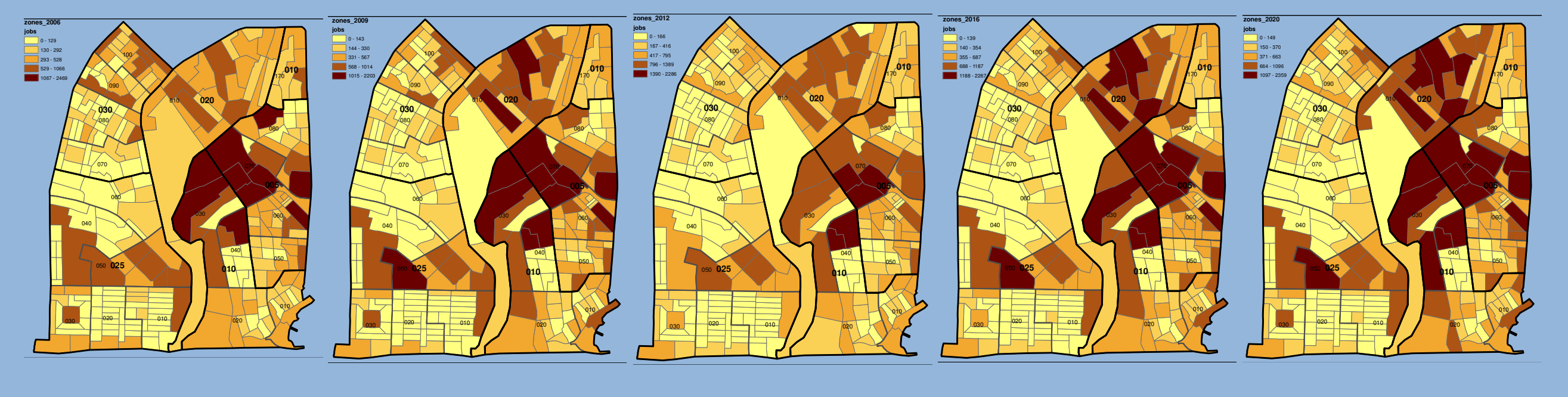


RESULTS

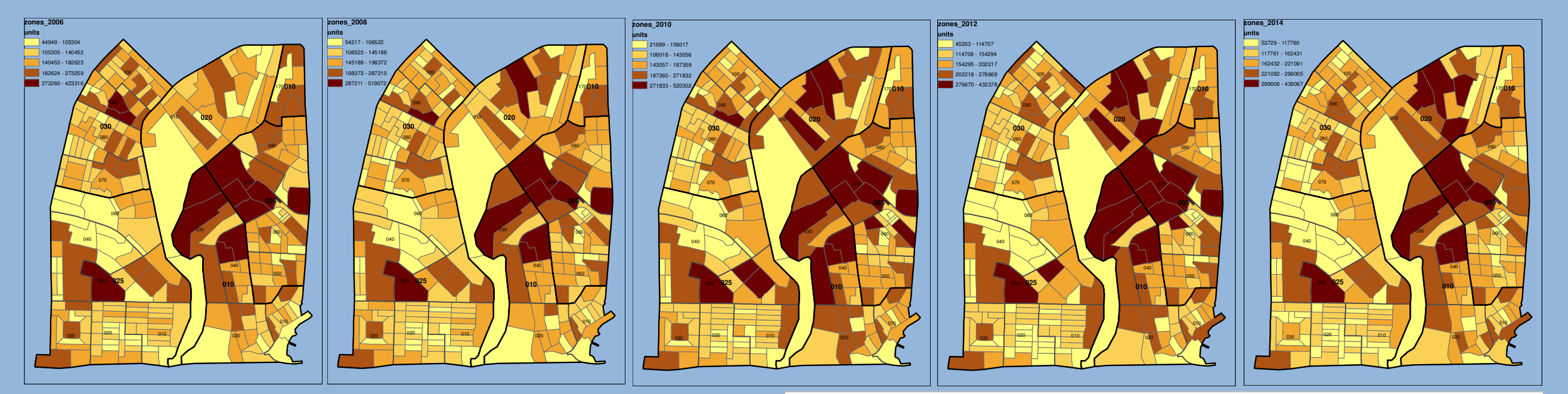


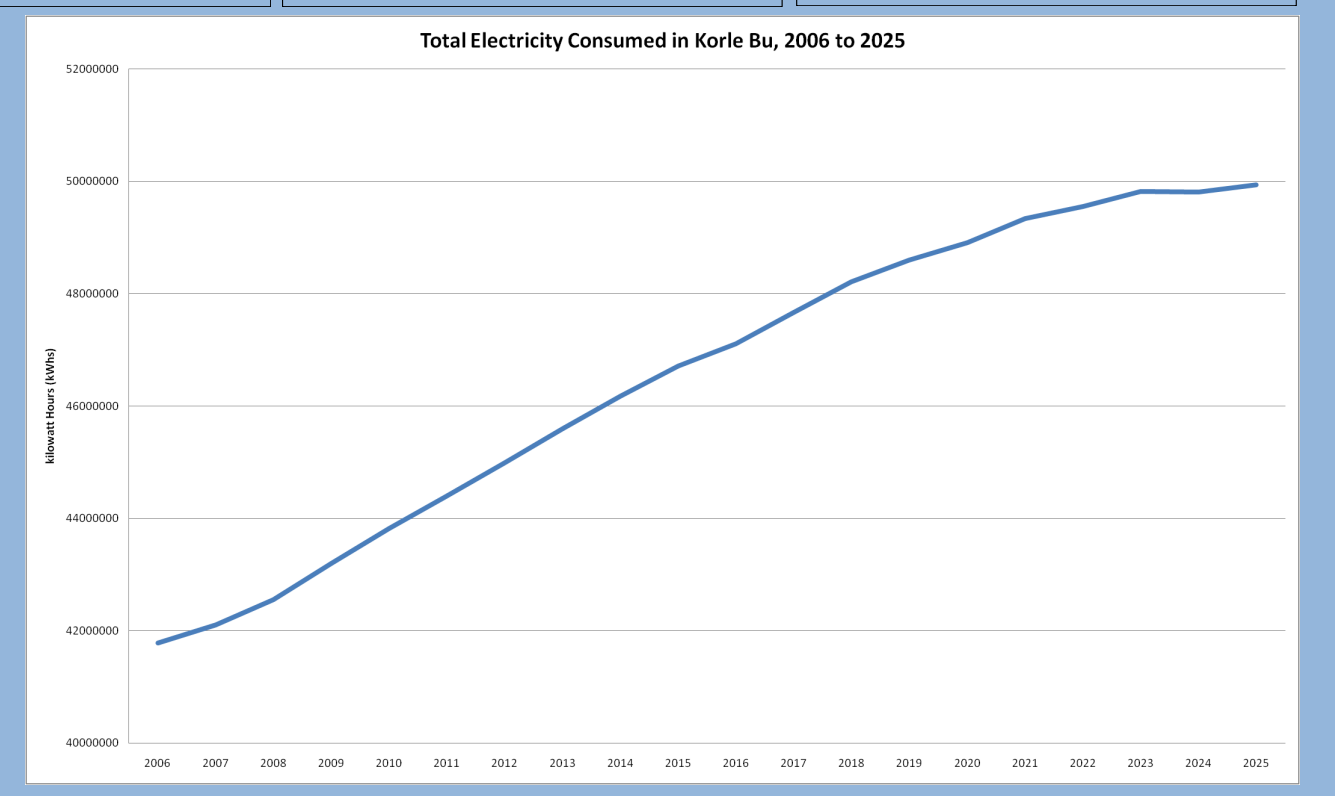


The graphic illustrates the number of jobs located in each of the 280 zones throughout Korle Bu as predicted by GAUS-KB for the years 2006 to 2025 in terms of a low demographic and economic growth rate. This spatial representation of the model projection indicates the largest single concentration of jobs will be in round 020-010, which primarily represent the South Industrial Park. This particular round is projected to outpace all other rounds over the twenty year period by more than double. The adjacent round 005-070, Agbogbloshie, also exhibits strong growth in number of jobs, which interestingly appears to be connected to the part of Adedenkpo known as "Sodom and Gomorrah." A third much smaller hotspot is located within round 025-050 and appears to be associated with a projected expansion of the Korle Bu Teaching Hospital.



The graphic presents the results of a regression model which has been used to project electricity demand based on the composition of demographic and economic characteristics comprising each of the 280 zones in Korle Bu for each year from 2006 to 2025. The dependent variable, which is measured as units of electricity consumed, projects total number of kilowatt-hours demanded, as determined by the number of households and their composition in terms of persons, workers, children and average annual income as well as the total number of jobs and number of jobs per sector. The graphic provides a spatial illustration of projected electricity demand by zone (plot) from 2006 until 2025 in terms of a low population and economic growth rate. In its aggregate, GAUS-KB projects total electricity demand for the entire district to increase by nearly 17% during the twenty year period, with total number of kWhs increasing from almost 42,000,000 to 50,000,000. Considering that the approximately 200,000 persons inhabiting Korle Bu represent only about 0.008% of Ghana's total population, this is a significant amount of projected consumption compared to overall totals for the country (once demand from the Valco aluminum smelter is removed). It should be noted that the demographic and economic projection used in this simulation was very conservative, but still represented an increase of nearly 17% in expected total consumption over the twenty year period. On the contrary increases in average household income of about 5% were modest at best.





GAUSS: Conclusions

Household Relocation

- Higher Household Mobility is Essential to Poverty Reduction
- Mobility Increases as Income Increases, which is especially true at Younger Ages

Household Location Choice

- GAUSS indicates Smaller Live-Work Kiosks may be contributing as a "stepping stone" towards middle income status

Electricity Demand

- At a low growth rate of 2% (economic and demographic) Korle Bu GAUSS exhibits an aggregate increase in electricity demand from less than 42mil kWh to 50mil kWh
- A 20% increase would equate to a nearly 8% increase in local domestic product (the population of Korle Bu represents about 1% of the total population of Ghana)
- GAUSS consistently projects plots with large numbers of kiosks as high consumers of electricity
- Formal commercial activities also project large demands for electricity, especially when they are in proximity to large groupings of informal activities (Agbogbloshie / Korle Dudor / Kwame Nkrumah Boulevard)
- GAUSS projects the South Industrial Park as the most significant consumer of power as well as the largest increases in power consumption over the time period 2006 to 2025

VolkswagenStiftung

THE GREATER ACCRA URBAN SIMULATION

Tyler Frazier, AICP, PhD
Department of Transportation System Planning and Telematics
Technische Universistat Berlin