A Framework for Modeling and Forecasting Population Age Distribution in Metropolitan Areas at Transportation Analysis Zone Level

By

Xiaoyu Zhu[1*]; Sabyasachee Mishra[1]; Timothy F. Welch[1]; Birat Pandey[2]; Charles M. Baber[2]

[1] National Center for Smart Growth Research and Education, University of Maryland, College Park, MD 20742

[2] Transportation Planning Division, Baltimore Metropolitan Council, 1500 Whetstone Way, Suite 300, Baltimore, MD 21230

* Corresponding Author

Xiaoyu Zhu, PhD., E-mail: xyzhu@umd.edu, Phone: 301-405-9386, Fax: 301-314-5639

1226 C Preinkert Field House, University of Maryland, College Park, MD 20742

Total Word Count: Words (4,628) + Number of Tables and Figures (9 x250) =6,878

(Submitted August 1, 2012)

Submitted for Consideration of Presentation at the 92nd Annual Meeting of the Transportation Research Board (TRB) in January 2013 and Publication at Transportation Research Record (TRR)

Abstract

Recent travel demand modeling practices focus on micro, disaggregate, and activity level travel behavior and patterns. The application of such practices requires detailed population information in socio-economic and demographic data. For example, in a four-step travel demand model total household and employment at Traffic Analysis Zone (TAZ) level are sufficient for trip generation. However, in an activity based model more detailed information in the small area (TAZ), such as population by different age categories and employment type, is required to produce trip chaining and other details in the population synthesis step. Conventionally many studies have used Iterative Proportional Fitting (IPF) to generate such detailed information. But, IPF suffers from severe drawbacks and is blind to detailed synthesis of variables. In this paper, a novel approach is presented where population by age category evolves over time period using logistic regression technique. The methodology is presented in three steps: coefficient estimation, forecast and validation. First, the 1990 census data is used to model population by age group in 2000 at the TAZ level. The model result is applied to forecast 2010 data for validation. The methodology is applied to Baltimore Metropolitan Council (BMC) region and the results show that the proposed model produces and forecasts reasonably well. The experiences gained from this study are: (1) population evolution pattern in city area should be treated separately from other, e.g., Baltimore City has a special population structure from other surrounding counties; (2) this model provides a good estimation and prediction for the age group 0-24 and 35-64 and the problems occurs in 25-34 and 65+ groups, whose migration trend is not consistent over time and cannot be captured by the current parameters alone. Though in this paper population by age is considered for demonstration, the proposed methodology can be used for other variables of interest such as household type, householder's age, employment type, occupation, etc. The proposed tool can be adapted by small and large scale planning agencies for preparing detailed socio economic and demographic input data for travel demand modeling practices.

Keywords: population forecast, distribution of personal age, logistic regression

1. Introduction

Over the last few decades many parts of the world have seen rapid urbanization, growing urban boundaries and increasing congestion. Answering various questions raised by urbanization with added demand poses a challenge to policymakers, planners and researchers. Adequate understanding of travel behavior and traveler demographics is a critical component in devising policies to tackle this problem. Currently, there is a trend of focus shift from macro-level to micro, disaggregate and activity-oriented travel behavior and travel demand modeling. The application of these studies in travel forecasting and land use policy requires more detailed population information on socioeconomic and demographic data. Currently, synthesis methods, such as Iterative Proportional Fitting (IPF), are greatly used to generate detailed socio-demographic characteristics of every resident household in the study area. There are limitations of controlled attributes used as input to these synthesis models: (1) The population projection models (e.g., Cohort-component method) to derive control attributes are commonly used for larger level of geography (county, state, national); (2) Limited set of variables, such as household size, income, are currently projected but not enough.

Meanwhile, there is a growing concern in the small area (TAZ, community) population projections because it is highly related to community service, transportation level of service and other social wellness. The population size by socio-demographic in each TAZ or community is an important indicator to predict the trip generation and distribution, intra-zonal linkage and housing growth. Because of the limitations in current projection methods, there are several attempts to build a framework for the small area population projection. Issues of lacking historical and current trend and developing reasonable migration assumptions are the critical problems.

In this paper, we are facing the issue to provide the supplemental input to Baltimore Metropolitan Council (BMC) population synthesizer, which is widely used to generate a detailed socio-demographic characteristic of every resident household in the model area. The BMC synthesizer (called as PopGen-BMC) produces future population based on the observed year data. Therefore, estimations of socio-demographic characteristics that change over time such as aging of population are less dependable. At present, limited set of variables (Number of Household by Size; Number of Household by Income; Number of Household by Worker and Total Population; Group Quarters Population) are used as controlled inputs to the synthesizer to generate other detail variables of interest.

Within this context, BMC desires to establish an aggregated sub-model that will allow estimating supplemental control variables required in population synthesis such as housing type, householder age group, personal age group, employment type, and workers by occupation at the Transportation Analysis Zone (TAZ) level. Among the variables of interest, county level control estimates for population by age, gender, race and age of householder are available through Maryland Department of Planning (MDP). These county totals need to be allocated at the TAZ level for input to synthesis. These evolving socio-demographic trends can be confirmed in the synthetic population estimates only if they are controlled as the inputs of the synthesis.

Therefore, we seek for a population projection approach applicable to small areas (such as TAZ) capturing historical and current trend. Population distributions by various household and personal socio-demographic characteristics need to be estimated and forecasted, such as housing type, householder age group, person age group, employment type, and worker by occupation. In this paper, the focus is on persons by age group. However, the presented methodology can be used for all the aforementioned variables. In the next section literature review encompasses research on disaggregated socio-economic and demographic evolution processes. The methodology section discusses steps for coefficient estimation, forecasting and validation. The input data collection step is presented next. Results section shows the performance of the model. Finally, summary and conclusion of the paper is discussed.

2. Literature Review

The demographic and socioeconomic updating methods within the travel demand forecasting community and quantitative analysis and forecast at household and person level are relatively limited (Miller, [1]). Traditional four-step modeling technique has been used by most of the planning agencies to forecast travel demand. Transition to a disaggregate model requires much more intensive data processes and faster computing abilities. For example, simulating the evolution of households and firms requires disaggregate data to estimate various life-cycle transition models. In the absence of disaggregate data, many practices have used growth factors or past experiences to forecast socio-economic data. In this section, different socio-economic and demographic evolution processes are outlined.

The popular approaches to forecast the demographic characteristics of future population are mostly used for the larger levels of geography, e.g., US Census Bureau uses the cohort-component method to produce the national and state population projections. Information of birth, death and migration are necessary in the forecast and the accuracy is relatively high at state level. As the growing need in small area studies, researchers from various fields (social science, statistics, urban planning) have adapted various methods for small areas analysis. Rees et al. [2] discussed a framework for small area population estimation, which is constructed by four stages. Estimation methods, such as apportionment, ratio, IPF, Cohort-component and enhancements (hybrid method, district level constraints) were compared in the research. Kanaroglou et al. [3] studied the spatial distribution of population at the census tract level using Cohort-component and aggregate spatial multinomial logit (ASMNL) model. A recent application of multinomial logistic model for Transportation Analysis Zone (TAZ) level population projection is proposed by Choi and Ryu [4]. Beyond the traditional methods, this is a new approach to forecast demographic distribution by capturing the historical and current trend.

Over the last few decades, a number of demographic and socioeconomic updating modules have been developed over multiple disciplines including DYNAMOD (King et al., [5]), DYNACAN (Dussault, [6]), NEDYMAS (Nelissen, [7]), and LIFEPATHS (Gribble, [8]). These modules explicitly model demographic processes at a high level of detail. However, they are not well suited for applications in the context of an activity-based travel microsimulation system because generating the necessary land-use and transportation system characteristics with these models is not straightforward. Sundararajan and Goulias [9] studied simulation of demographic evolution for the purposes of travel forecasting in a tool called as DEMOgraphic (Micro) Simulation (DEMOS) system. Other population updating systems have been developed in the travel demand forecasting community with varying levels of detail and sophistication, including the Micro-analytic Integrated Demographic Accounting System (MIDAS) proposed by Goulias and Kitamura [10] and the Micro-Analytical Simulation of Transport Employment and Residences (MASTER) recommended by Mackett [11]. Certain aspects of the population evolution processes, such as residential relocations and automobile ownership are focused by land-use transportation modeling systems, including TRANUS (Barra, [12]), MEPLAN (Hunt, [13]), URBANSIM (Waddell, [14]), STEP2 (Caliper Corporation, [15]), ILUTE (Miller et al., [16]), PECAS (Hunt et al., [17]), and POPGEN (Pendyala et al., [18]).

Models of life-cycle transitions require special panel surveys to track changes in the demographics of a household. Since such surveys are rare, there have been very few models which track household evolution in great detail. MIDAS by Goulias and Kitamura's [10] is one of such models, which combines models of travel behavior with a microsimulation model of household demographics. MIDAS was calibrated using the Dutch National Mobility Panel dataset. Another study of interest is STEP2 model for Nevada's Clark County (Caliper Corporation, [15]), which is closely mimicked by this study's' rules of household evolution.

In this study, the supplemental data needed for POPGEN is studied. IPF procedure used in POPFGEN only matches the control totals in the disaggregation process, but is blind to the temporal evolution. The disadvantages of IPF are (1) only controls for household attributes but not personal attributes, (2) fails to synthesize populations to match distributions of target person characteristics, and (3) ignores differences in household composition among households within a TAZ (Pendyala and Konduri, [19]). In the next section, methodology framework used to prepare supplemental data is discussed.

3. Methodology Framework and Forecasting Process

The modeling framework in this research is shown in Figure 1. The framework consists of three steps: estimation, forecast and validation. The methodology in each step is discussed in this section.

3.1 Coefficient Estimation

In this step of coefficient estimation, we have six designed target variables in our framework: Household type, householder's age, personal age, employment type, school child year and worker by occupation. Variables corresponding to each target can be grouped as major variables. All the other variables are secondary variables, such as household size, income, workers, and zone characteristics. The methodology in this process is baseline-category logit model or multi-category logit model, one of the logistic regression models. To predict the future population distribution by various socio-economic and demographic in each TAZ, the population distribution data for two base years, 1990 and 2000, in these zones are required. The impact of historical population (1990) on the population ten years later (2000) is examined and the evolution trend is captured skipping the detailed birth, death and migration. The formulation is explained taking person by age group as an example.

Let probability of population in each age group defined as $\pi_j = P(Y = j), j = 1,2, \dots, 8$. $j = 1$ for age *0-4;* $j = 2$ for *5-14;* $j = 3$ for *15-37;* $j = 4$ for *18-24;* $j = 5$ for *25-34;* $j = 6$ for *35-44;* $j = 7$ for *45-64;* $j = 8$ for over *65*. The age group $j = 7$ is chosen to be the baseline (reference) category, because the population in this group is generally more than other categories and less likely to be zero. The formulation of the baseline category logit model is

$$ln\left(\frac{\pi_{j\_00}}{\pi_{7\_00}}\right) = X\beta_j \quad or \quad \frac{\pi_{j\_00}}{\pi_{7\_00}} = exp\left(X\beta_j\right),\ j = 1, 2, \dots, 6, 8. \qquad (1)$$

Where, π_{j_00} is the $(n \times 1)$ vector of probabilities of age group j in year 2000. n is the number of TAZs. X is the input explanatory variables, which contain the major variables (1990 population by age group), and secondary variables like median income. $\beta_j, j = 1, 2, \dots, 6, 8$ are the parameters to be estimated. $\frac{\pi_{j\_00}}{\pi_{7\_00}}$ is the odds ratio of group j to group 7.

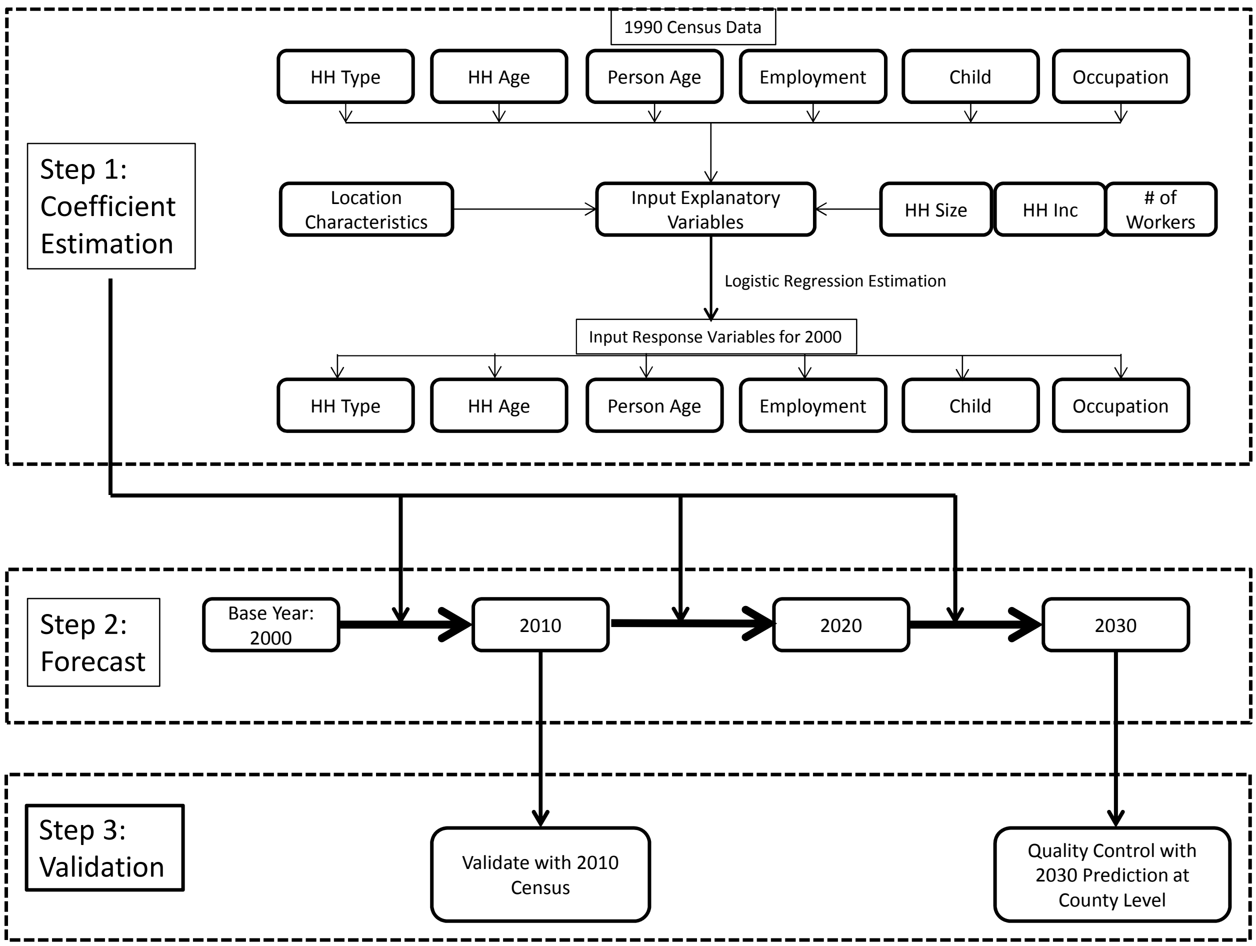


FIGURE 1 Flowchart of the Proposed Methodology

3.2 Forecast

The second step is using the estimation result $\widehat{\beta}_j, j = 1, 2, \dots, 6, 8$ from step 1 as the growth trend and 2000 census data as base year input X_{00} to forecast the population in 2030. The forecast is conducted as the following process by each decade. First, probability of 2010 population in each age group π_{j_10} will be calculated using 2000 as base year.

$$\pi_{j\_10} = \frac{\exp(X_{00}\widehat{\beta_J})}{1+\sum_i exp(X_{00}\widehat{\beta_\iota})}, \quad i,j = 1\ 2, \dots, 6, 8$$
$$\pi_{7\_10} = \frac{1}{1+\sum_i exp(X_{00}\widehat{\beta_\iota})}, \quad i = 1\ 2, \dots, 6, 8 \qquad (2)$$

Then the population by each group could be calculated based on the total population Pop_{10} in each TAZ in 2010 by the formulation $Age_{10} = Pop_{10} \times \Pi_{10}$, where $\Pi_{10} = [\pi_{1\ 10}, \pi_{2\ 10}, \dots, \pi_{8\_10}]$. Age_{10} or Π_{10} can serve as a major component of X_{10} which also includes other secondary variables as well. Similarly to the above step, we can calculate the probability of population by each age group in 2020 π_{j_20} using X_{10} as input.

$$\pi_{j\_20} = \frac{exp(X_{10}\widehat{\beta_J})}{1+\sum_i exp(X_{10}\widehat{\beta_\iota})}, \quad i,j = 1\ 2, \dots, 6, 8$$
$$\pi_{7\_20} = \frac{1}{1+\sum_i exp(X_{10}\widehat{\beta_\iota})}, \quad i = 1\ 2, \dots, 6, 8 \qquad (3)$$

Repeatedly, $\pi_{j\_30}, j = 1\ 2, \dots, 8$ can be calculated and the target population by each age group X_{30} can be achieved.

3.3 Validation

The validation is designed at two stages. First, with the 2010 census at county level, the 2010 forecast could be compared with the actual census outcome. We can compare the observation and prediction by examining the value, Mean Absolute Percentage Error (MAPE) and Median Absolute Percentage Error (MedAPE). If the validation at this step is acceptable, we can continue the forecasting for 2020 and 2030. If the validation result indicates huge deviance between the prediction and observation, we need to improve the model until it fits well. The second step is validating the final forecast of 2030, by comparing with the projected county control for the demographic distribution provided by MDP. Similarly, MAPE and MedAPE will be computed to test the fitness of prediction.

4. Data

There are four datasets retrieved for the study. The first group is for 1990 and the second for 2000. The 1990 data is collected from the census ftp site and included summary file 1 (SF1), which is 100% data from the short form census and summary file 3 (SF3), which is sample data from the long form census. The year 2000 data is collected from the same ftp site and consisted of summary files 1 and 3. SF1 contains the information of age, gender, race, household structure, housing units, etc. SF3 contains data, such as education, occupation, and commute mode, etc. The entire collection, allocation and aggregation process is shown in figure 2, with the retrieved data at the top of the figure for each census year and summary file. The mid-section of the figure describes the data formatting and the bottom of the figure shows how the data was either allocated or aggregated to TAZs depending on the type of summary file.

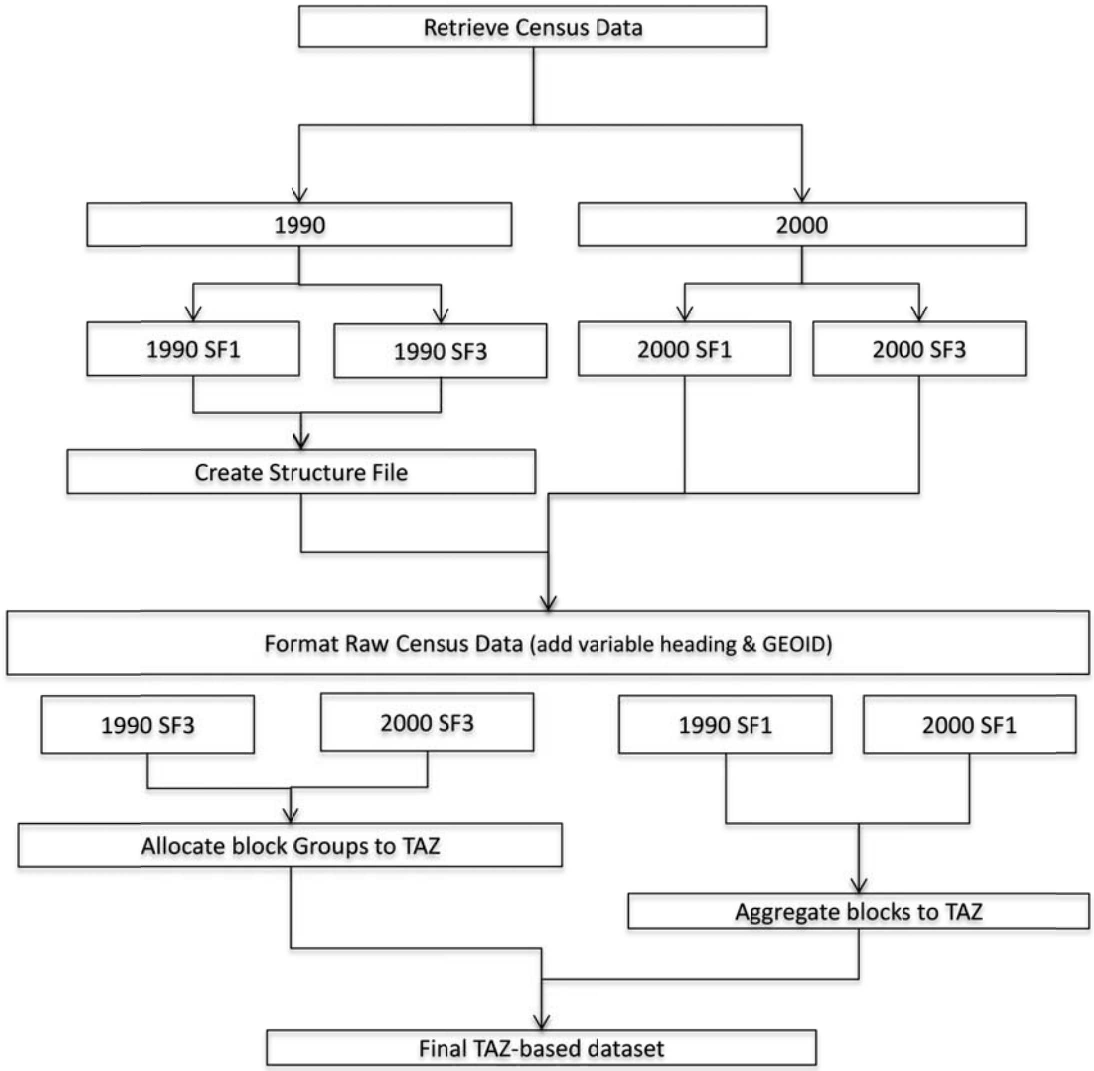


FIGURE 2 Census data collection and TAZ allocation Process

To manipulate the data to match the 2010 TAZ division, allocation and aggregation procedure are required on SF3 and SF1, correspondingly. The SF3 data is only available at the block group level for 1990, which does not always nest within TAZs. To convert the SF3 data, each block group record had to be allocated to TAZs which in some cases were larger than block groups and in other cases smaller. To properly allocate the block group data to TAZs, each census block group boundary file was imported into ArcGIS. The block group boundaries were overlaid on a 2010 TAZ shapefile. Each of the shapefiles was clipped to remove water and other non-developable features where census data likely did not exist. For the remaining area, in the absence of more detailed spatial data, it was assumed that population and households are evenly distributed across each block group. The ARCGIS creates a ratio for each block group to proportionately re-allocate each record to the 2010 TAZ. Once the ratios were established, the 1990 and 2000 formatted census data was merged with the block group geographic data, with the ratios dividing the results by TAZ. The SF1 files for both 1990 and 2000 are available at the block level, which nests very well in to BMC TAZ geography. Each census block boundary file was imported into ARCGIS. The block boundaries were overlaid on a 2010 TAZ shapefile. The ArcGIS spatial join tool was used to attach the TAZ number that each block fit into. Once this relationship was established, the final block data was aggregated to TAZ. The census data is collected at TAZ level for BMC region. The study area including county and TAZ boundaries is shown in Figure 3.

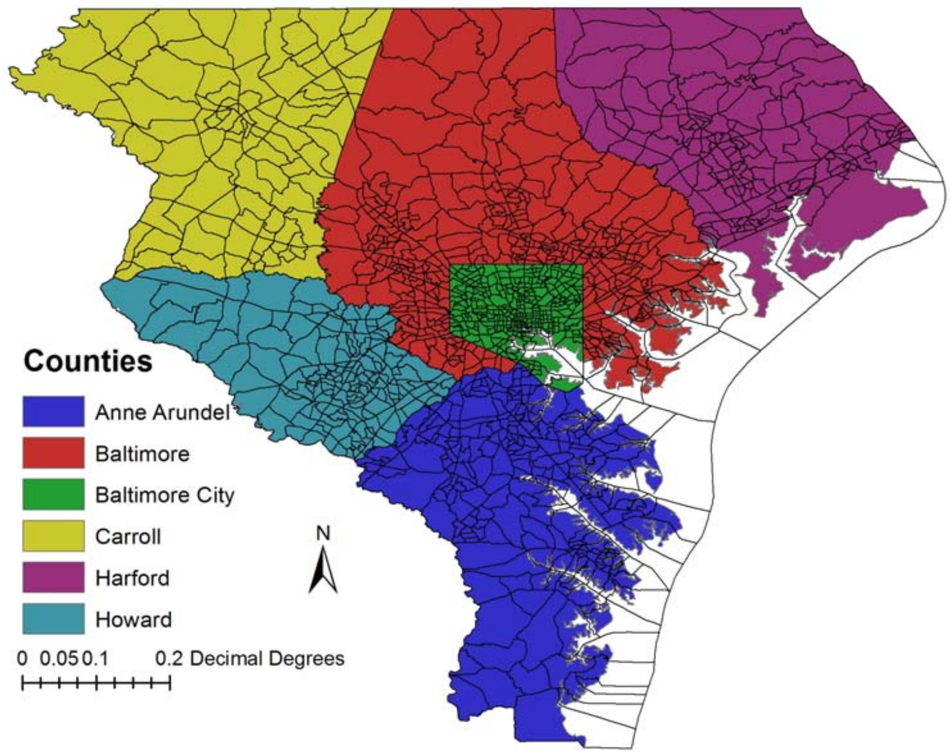


FIGURE 3 TAZ and County Boundary of the Study Area

5. Estimation and Forecasting Results

In this section, we present the model framework and discuss result using one of the targets population age group as an example. We use the same example as the methodology section to apply the framework to estimate and forecast population by age group.

The data cleaning step is to remove the outliers and invalid data. Special TAZs in the sample are not included in the estimation, such as empty zones, TAZs exclusive for group quarters or with high percentage of group quarter populations. At the beginning, we worked on the TAZs in six counties, but the validation did not fit well because the Baltimore City is quite different from others. The result presented in the section is for the model applied on five counties: Anne Arundel, Baltimore County, Carroll, Harford, and Howard, totally 763 TAZs. The data description for the variables is displayed in Table 1.

TABLE 1 Description of explanatory variables in the age sample

Variables	Label	mean	min	max	Std-deviation
PAge0_4	Percentage of Age 0-4 in 1990	7.42%	0.00%	16.67%	2.31%
PAge5_14	Percentage of Age 5-14 in 1990	13.35%	0.00%	24.45%	3.50%
PAge15_17	Percentage of Age 15-17 in 1990	3.75%	0.00%	19.62%	1.33%
PAge18_24	Percentage of Age 18-24 in 1990	8.97%	0.00%	29.15%	2.72%
PAge25_34	Percentage of Age 25-34 in 1990	17.93%	3.64%	50.00%	6.65%
PAge35_44	Percentage of Age 35-44 in 1990	17.23%	0.00%	36.54%	3.73%
Page45_64	Percentage of Age 45-64 in 1990	20.90%	5.23%	41.67%	6.03%
PAge65	Percentage of Age over 65 in 1990	10.45%	0.00%	51.17%	6.21%
medinc (10K)	2000 Median income in TAZ (in unit 10,000)	6.3966	1.1035	13.5460	2.0925
HHDEN00	2000 Household density in TAZ (per acre)	1.7364	0.0357	9.5340	1.7273
EMPDEN00	2000 Employment density (per acre)	2.2654	0.0464	9.7926	2.0881
GQDEN00	2000 Group quarter density (per acre)	0.0503	0.0000	2.4783	0.1742

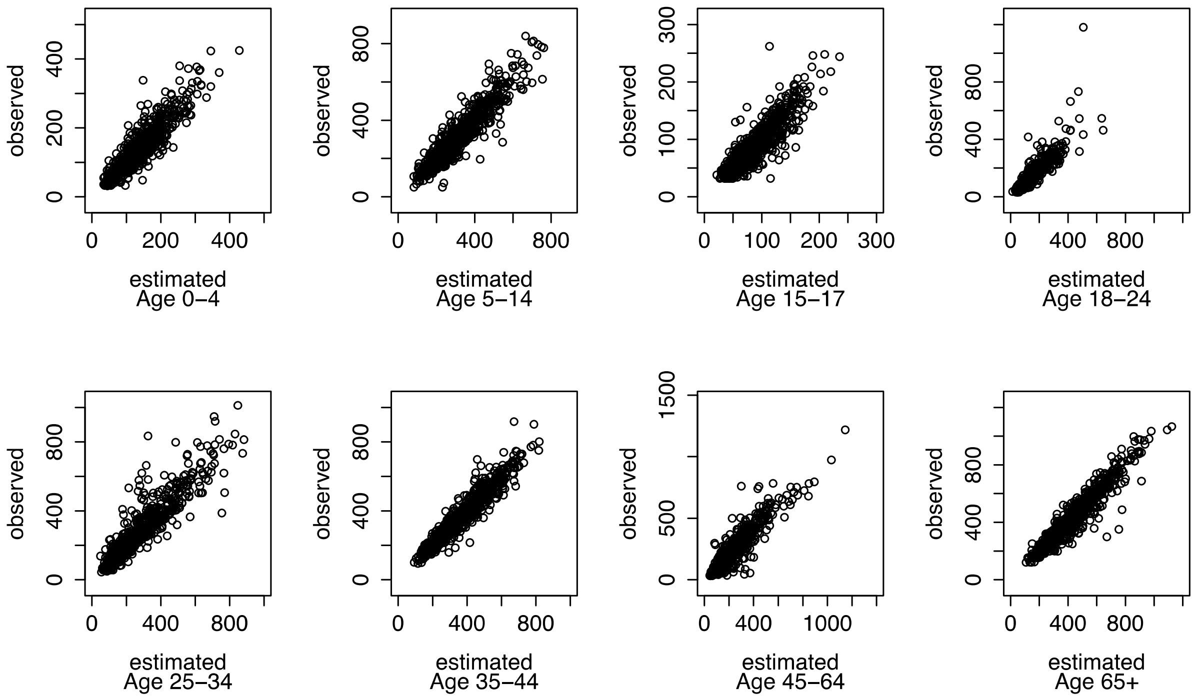
The explanatory variables displayed in the Table 1 include the historical age distribution ten years ago, current median income, population density, employment density and group quarter density. We also examined variables, such as the distribution of household size, income and number of workers. But these variables are proved to be not highly correlated with age distribution. The estimation result is shown in Table 2.

As in Table 2, most the coefficients are over 99% significant (shown in black) by examining the p-value and insignificant coefficients are shown in gray. Positive sign means the larger value in this row category is positively correlated with a higher odds ratio in the category by column comparing to age 45-64, vice versa. We explain the result table using coefficient of independent variable “P_Age25_34” and dependent variable ”Age35_44_00”, which equals to 3.626 (highlight in grey) as an example. This coefficient is interpreted that if there is 1 percent more population in 25-34 age group in 1990 out of the total, there would be a multiplicative effect by $exp(3.626\ \times 1\%) =\ 1.037$ on odds of Age35_44 rather than odds of Age45_64 in 2000. Similarly, this 1% more in 25-34 will also increase the odds ratio of any other age groups to 45-64, except the odds of 65+, by observing the positive coefficients in row 25-34 except the last one. Another example is the coefficient of -0.127 in row “HHDEN00” and col “Age25_34_00”. Odds ratio of Age25_34 to Age45_64 would decrease with higher household density. This indicates that comparing with 45-64 age group, younger (25-34) are less likely to leave in high density area. While positive or negative sign does not definitely imply the increase of decease in probability for a particular age group. The impact of one parameter on the probability of any age group is finally decided by all the coefficients in the row of this parameter (refer from Equation 2).

TABLE 2 Estimation results for age group

	Age0_4	Age5_14	Age15_17	Age18_24	Age25_34	Age35_44	Age65
constant	**-3.200**	**-1.769**	**-3.232**	**-4.082**	**-2.432**	**-1.494**	**1.533**
PAge0_4	**7.288**	**5.844**	**4.530**	**3.226**	**3.545**	**3.260**	**-2.764**
PAge5_14	**4.295**	**5.093**	**5.713**	**4.873**	**1.238**	**2.048**	**-1.531**
PAge15_17	-0.384	-0.389	**2.036**	**3.530**	**5.288**	**3.091**	**-5.831**
PAge18_24	**3.752**	**1.220**	**2.471**	**11.221**	**4.099**	**0.629**	**-3.686**
PAge25_34	**3.608**	**2.249**	**1.442**	**3.345**	**5.699**	**3.626**	**-1.945**
PAge35_44	**-3.761**	**-3.414**	**-1.869**	**0.976**	**-1.314**	**-2.351**	**-4.115**
PAge65	**1.838**	**1.206**	**2.058**	**4.060**	**2.171**	**1.573**	**1.433**
medinc (10K)	**0.050**	**0.039**	**0.014**	**-0.060**	**-0.038**	**0.031**	**-0.066**
HHDEN00	**-0.019**	-0.005	**-0.082**	**-0.059**	**-0.127**	**-0.067**	**0.096**
EMPDEN00	**0.030**	**0.014**	**0.063**	**0.070**	**0.132**	**0.052**	**-0.076**
GQDEN00	**-0.192**	**-0.158**	**-0.091**	-0.004	**0.065**	0.017	**0.339**

The next step is the model evaluation before using the estimated coefficients for prediction. We compare the fitted value of the estimation with the observed data in 2000 by plotting the observed against the fitted population of 763 TAZs for each age group. The validation result is displayed in Figure 4. Most of the points are homoscedastic (along the diagonal line) with acceptable deviation. The validation proves the model fits well and error is moderate. We also evaluate the model with a mean absolute percentage error (MAPE) of 15% and median absolute percentage error (MedAPE) of 10%.



(Note: MAPE = 15% and MedAPE = 10%)

FIGURE 4 Validation plot of observed population against fitted population in 2000

Then we start the prediction and validation step for 2010. With the estimated coefficient, we calculate the probability of population distributed in each age group in 2010, using the observed population by age group in 2000. With approximated total population in each TAZ in 2010, we obtain the number of population by age category in these TAZs. The prediction procedure is conducted on 1047 zones in 5 counties. We present the predicted population age distribution at county level instead of TAZ level in the first row of each county in Table 3.

The validation is conducted at county level because currently the observed age distribution in 2010 is available at county level. To validate the 2010 forecast, we combine the age group 15-17 and 18-24. The county level population by age group in 2010 is achieved from Maryland Department of Planning (MDP) and is used to examine the prediction accuracy. The absolute percentage error of the validation is shown in the second row of each county in Table 3. The average error (MAPE) at county level is 10.2% and median error (MedAPE) is 6.2%. We observe a larger error in Age 25-44 and over 65. Age 25-34 is the population with huge migration potentials, such as marriage, graduate and new employment opportunity. The migration pattern for 25-34 from 1990 to 2000 is not consistent with the pattern from 2000 to 2010. Also the error of 65+ means that the aging pattern for the older is not well captured in this model. The population evolving trend is stable over the last two decades in age 15-24, 35-44, and 45-64. Overall, the validation results appear reasonable and trustable.

TABLE 3 Estimated county level population by age group in 2010

County	0-4	5-14	15-17	18-24	25-34	35-44	45-64	65plus
Anna Arundel	32,925	87,364	25,929	37,356	58,769	95,090	128,765	66,390
	8.5%	*13.7%*	*2.4%*		*25.2%*	*1.0%*	*1.2%*	*25.0%*
Baltimore County	46,822	122,218	37,365	65,990	93,122	135,910	185,999	129,054
	4.4%	*9.7%*	*0.9%*		*15.1%*	*1.7%*	*2.5%*	*8.1%*
Carroll	10,317	29,839	9,571	11,991	16,644	31,281	44,161	21,719
	12.2%	*3.6%*	*6.2%*		*20.0%*	*4.9%*	*5.1%*	*14.8%*
Harford	15,207	42,376	13,088	15,372	25,269	44,054	61,827	31,508
	15.2%	*5.4%*	*1.6%*		*22.5%*	*3.9%*	*5.0%*	*25.0%*
Howard	18,976	52,662	15,204	15,024	29,474	53,870	70,195	28,155
	9.1%	*13.5%*	*0.5%*		*29.5%*	*3.2%*	*3.9%*	*33.2%*

We also display the error by age group in Figure 5 for 5 counties. Predictions for Age 0-4 and 35-44 are matched with observation quite well, with the points along the diagonal line. The percentage error for Age 0-4 in Carroll and Harford are above 10% in Table 3 but along the diagonal in Figure 5, because the population in this group is small and the percentage error is enlarged relatively. From Figure 5, we also observed an underestimation in age group 25-34, and 45-64. In addition, age 5-14 and over 65 are overestimated. Based on the percentage error and comparison between observed value and prediction at county level, this model provides a good estimation and prediction for the age group 0-24 and 35-64 and the problems occurs in 25-34 and 65+ groups, whose migration trend is not consistent over time and cannot be captured by the parameters in Table 1 alone.

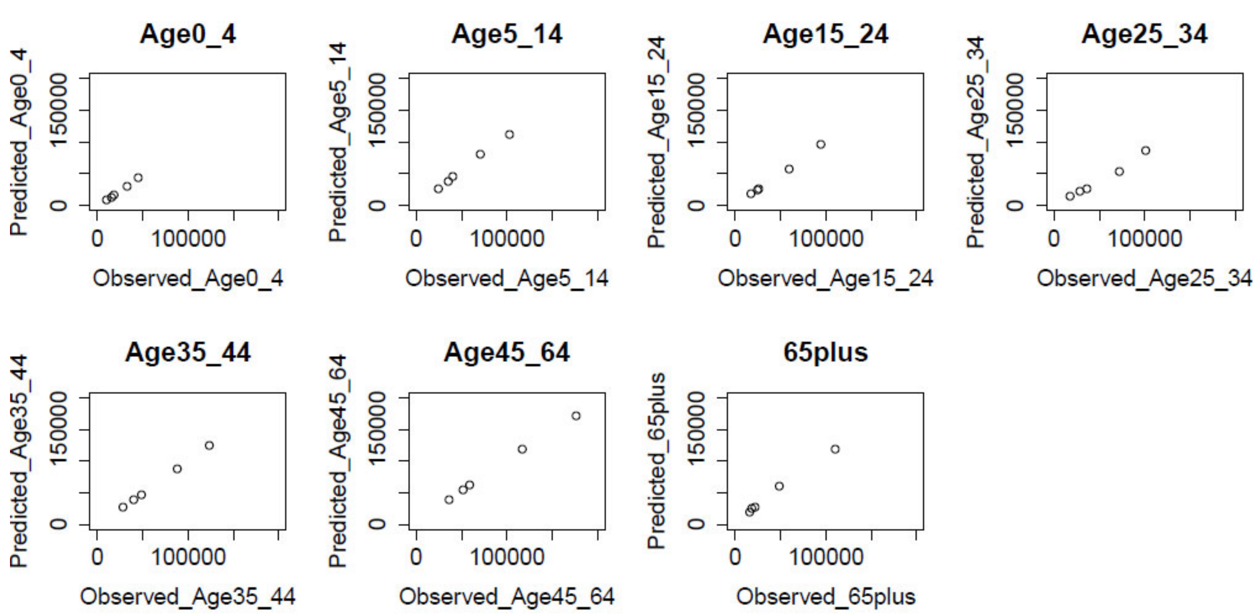


FIGURE 5 Validation plot of predicted population and observed population at county level by age group

After the above validation, we continue the forecast step designed in the framework and achieve the forecasting result in 2030. The approximate total population for each TAZ in 2030 is provided by BMC. The estimated county level population by age group is presented in Table 4.

TABLE 4 Estimated county level population by age group in 2030

County	0-4	5-14	15-17	18-24	25-34	35-44	45-64	65plus
Anna Arundel	37,041	97,821	33,488	49,993	56,169	104,099	125,462	69,953
	8%	*29%*			*12%*		*-8%*	*-39%*
Baltimore County	51,171	131,340	44,414	83,906	92,933	144,533	181,729	132,110
	5%	*10%*			*17%*		*-7%*	*-28%*
Carroll	12,804	34,624	12,492	17,762	19,407	38,415	47,133	24,674
	8%	*21%*			*21%*		*4%*	*-49%*
Harford	17,847	47,749	16,901	22,962	28,070	52,122	64,761	37,420
	1%	*20%*			*13%*		*0%*	*-38%*
Howard	23,392	64,162	21,472	22,951	28,924	62,553	70,429	33,573
	17%	*41%*			*11%*		*-8%*	*-50%*

Table 4 also shows a comparison of county level prediction of 2030 with demographic projection provided by MDP. The age categories provided by the MDP projection are 0-4, 5-19, 20-44, 45-64 and over 65. We could not compare exactly using our prediction of population age category, for example, the second column result in Table 4 is comparing the prediction of age 5-17 with the MDP projection of age 5-19. The prediction in our model is more than the current projection. The age group of 65+ still has the largest error, which cannot be predicted very well in this current model. Generally, we obtained that our model has an underestimation for older age and an overestimation for teenage comparing with the projection data.

6. Conclusions

In conclusion, this paper provides a framework of forecasting future demographic and socio-economic distribution in a small area (TAZ level). The framework is applied to forecast age group distribution and the modeling results, model evaluation, forecasting and validation process are presented in this paper. The model evaluation and validation of prediction results prove that the baseline category logit model is a reasonable approach and the prediction is acceptable. The final prediction for 2030 in our model has an underestimation for population over 65+ (consistent with synthesis outcome) and an overestimation for teenage than the projection data.

In this study, we also encounter many obstacles. The major problem is accuracy of the data for estimation and prediction. For example, the TAZ zoning system changed from 1990 to 2010. To maintain consistency in estimation and prediction, the secondary variables need to be allocated to 2010TAZ assuming the population is evenly distributed across the study area. Additionally, we use the values such as population, income, household density of each TAZ in 2020 and 2030 in the prediction procedure, which could not be evaluated how accurate they are. Also we could only compare the final forecast in 2030 with projection in 2030 provided by MDP approximately. Additionally, we wish to include variables corresponding to each TAZ but not available currently, such as number of schools, recreation centers, shopping centers, which are related with to the population residential location choice. These variables are useful for scenario planning purpose, e.g., an expanding TAZ with more schools or business area.

There are some important summaries and conclusions on population socio-demographic distribution forecast in this paper. First, population evolution pattern in city area should be treated separately from other, e.g., Baltimore City has a special population structure from other surrounding counties. Second, this model provides a good estimation and prediction for the age group 0-24 and 35-64 and the problems occurs in 25-34 and 65+ groups, whose migration trend is not consistent over time and cannot be captured by the current parameters alone. The Age 25-34 is the population with huge migration potentials, such as marriage, graduate and new employment opportunity. Also the error of 65+ means that the aging pattern for the older is not well captured in this model. More migration and aging related information are necessary to improve the model estimation.

Currently, we have applied this framework to predict age distribution. In future, we plan to apply this framework to on other demographic variables such as household type, and occupation. There are other issues to solve to fulfill the framework, such as developing a separate model for Baltimore city region and collecting more data for estimation. Meanwhile, we plan to improve this framework and build up an applicable and deliverable production in open source software and integrated into travel demand modeling practices.

Acknowledgement

This research is supported by Baltimore Metropolitan Council. Computing facilities at National Center for Smart Growth Research and Education, University of Maryland College Park are greatly acknowledged. The authors are thankful to Dr. Fred Ducca for his suggestions throughout the project. The opinions and viewpoints expressed in this paper are entirely those of the authors, and do not necessarily represent policies and programs of the aforementioned agencies.

References

[1] Miller, E. J. Microsimulation. In *Transportation Systems Planning: Methods and Applications*, Eds. K. G. Goulias, CRC Press, Boca Raton, Ch. 12, 2003.

[2] Rees, P., Norman, P., Brown, D. A framework for progressively improving small area population estimates. *Journal of The Royal Statistical Society*. Series A (Statistics In Society), 167 1 , 2004, 5-36.

[3] Kanaroglou, P.S., Maoh, H.F., Newbold, K. B., Scott, D. M., Paez, A. A Demographic Model for Small Area Population Projections: Am Application to the Census Metroplitan Area (CMA) of Hamilton in Ontario, Canada. Working paper, 2007.

[4] Choi, S., Ryu, S. Linking the Regional Demographic Process and the Small Area Housing Growth: Implications for the Small Area Demographic Projections. Presented at the 52nd Association of Collegiate Schools of Planning Conference, October 13-15, 2011.

[5] King, A., H. Baekgaard, and M. Robinson. DYNAMOD-2: An Overview. Technical Paper no. 19, National Centre for Social and Economic Modelling, University of Canberra, Australia, 1999.

[6] Dussault, B. Overview of DYNACAN - a full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes. www.actuaries.org/CTTEES_SOCSEC/Documents/dynacan.pdf, 2000

[7] Nelissen, J. H. M. Demographic Projections by Means of Microsimulation. The NEDYMAS model, part A+B, Tilburg University Press, Tilburg, 1995.

[8] Gribble, S. LifePaths: A Longitudinal Microsimulation Model Using a Synthetic Approach. In *Microsimulation in Government Policy and Forecasting*, Eds. Gupta, A., and V. Kapur, Elsevier, Amsterdam & New York, Ch. 19, 2000.

[9] Sundararajan, A., Goulias, K. G. Demographic Microsimulation with DEMOS 2000: Design, Validation, and Forecasting. In *Transportation Systems Planning: Methods and Applications*, Eds. K.G. Goulias, CRC Press, Boca Raton, Ch. 14, 2003.

[10] Goulias, K. G., Kitamura, R. A Dynamic Model System for Regional Travel Demand Forecasting. In *Panels for Transportation Planning: Methods and Applications*, Eds. Golob, T., R. Kitamura, and L. Long, Kluwer Academic Publishers, Boston, Ch. 13, 1996, pp. 321-348.

[11] Mackett, R. L. *MASTER Mode*. Report SR 237, Transport and Road Research Laboratory, Crowthorne, England, 1990.

[12] Barra, T. de la. *Integrated Land Use and Transport Modelling*. Cambridge University Press, Cambridge, 1989.

[13] Hunt, J. D. A Description of the MEPLAN Framework for Land Use and Transport Interaction Modeling. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., 1993.

[14] Waddell, P. UrbanSim, Modeling Urban Development for Land Use, Transportation, and Environmental Planning. *Journal of the American Planning Association*, Vol. 68, 2002, pp. 297-314.

[15] Caliper Corporation. STEP2 for Clark County: Household Microsimulation for Transportation Policy Analysis. Prepared for the Southern Nevada Regional Planning Coalition, 2003.

[16] Miller, E. J., J. D. Hunt, J. E. Abraham, and P. A. Salvini. Microsimulating Urban Systems. *Computers, Environment and Urban Systems*, Vol. 28, 2004, pp. 9-44.

[17] Hunt, J.D. PECAS, University of California Land Use and Transportation Center, University of California, Davis, 2011

[18] Pendyala, R.M., Christian, K.P., Konduri, K.C. *PopGen 1.1 User's Guide*. Lulu Publishers, Raleigh, North Carolina, 2011

[19] Pendyala, R. and Konduri, K. Population Synthesis for Travel Demand Modeling. Data Needs and Application Case Studies. Presented in Using Census Data for Transportation Applications Conference, Irvine, California, 2011