# Validation of the Atlanta (ARC) Population Synthesizer (PopSyn)

Prepared for the TRB Conference on Innovations in Travel Modeling, Austin, Texas, May 21-23, 2006
John L. Bowman (John_L_Bowman@alum.mit.edu)
Guy Rousseau, Modeling Manager, Atlanta Regional Commission

## INTRODUCTION

This paper presents the results of initial base year and backcast validation of the new ARC Population Synthesizer (PopSyn). A population synthesizer acts as the conduit of land use information to the travel demand model. It takes information from the census and the land use model, and creates a detailed synthetic population consistent with the land use forecasts. A travel demand model can then predict travel for this population. The synthetic population includes a record for each household in the region, and a record for each person in the household, so it is well suited for use by travel demand models employing disaggregate microsimulation.

Although a population synthesizer constitutes a powerful tool, it should be used with caution. By design it provides misleadingly precise details about every person in the population. Because of limitations of its inputs and its synthesizing procedures, at best only some of the person and household characteristics accurately represent the population at the regional level of geographic aggregation, and many of those characteristics can be imprecise and inaccurate for very small geographic areas such as census tracts. Because of this, a fundamental goal in the development of a population synthesizer is to synthesize as accurately and precisely as possible, at as disaggregate geography as possible, as many variables as possible that determine travel behavior. And a fundamental requirement in the use of a population synthesizer should be to rely on it only for the characteristics it accurately represents, and to aggregate results to a level at which the synthetic population is precise and accurate.

## THE ARC POPULATION SYNTHESIZER

From the beginning ARC took seriously the need to use a population synthesizer properly, and insisted on being able to validate the synthetic population used for travel demand forecasts. Implicit in this is the need to be able to adjust the synthesizer if the validation results aren't as expected. With a flexible validatable population synthesizer, validation then can become more of an iterative tuning procedure.

The ARC population synthesizer works in the following basic steps, common to many population synthesizers. First it starts with an estimate of the number of households in each zone, with details (in the cells of the matrix) for each of several demographic categories. It also has population forecasts for some aggregate categories. These control totals are more accurate but less detailed than the initial estimates. An iterative proportional fitting (IPF) procedure adjusts the detailed distribution to match the control totals. Then the adjusted numbers of households of each type are drawn from PUMS.

For the base year population, ARC defined controls per Transportation Analysis Zone (TAZ), with all the control values coming from census tables (sf1, sf3 and CTPP). The synthesizer's design allows flexibility in the definition of the matrix cells and control categories, so that a variety of 1-, 2- and 3-dimensional census tables can be used to supply controls, enabling the capture of valuable joint distribution information available in the census tables. For families, the controls distinguish "with children" from "without". For non-families, the controls distinguish "single" from "2+" person per household. For families with householder over 65 years of age, the distinction by presence of children is ignored.

For the forecast year we define fewer controls for each TAZ. These capture ARC TAZ-level forecasts of household income and household size. However, ARC also forecasts some elements at a regional level that can be used for regional controls, such as the average number of workers within household, and the size of age cohorts.

The Population Synthesizer creates a synthetic population for a base year and for each forecast year. There are two key differences between the base year and the forecast year. The initial distribution for the base year comes from PUMS, whereas for the forecast year it comes from the base year distribution. The controls for the base year come from census tables, but for the forecast years, they come from the land use forecasts. In both cases, it produces a synthetic population, and it also produces a validation report that compares synthetic population characteristics to known characteristics.

To validate the synthesizer's ability to generate a forecast population, ARC uses year 2000 as base year and validates a back-cast to 1990. The initial distribution comes from the base year PopSyn. The controls then emulate a 1990 forecast dataset, and synthesize a 1990 population, which is then compared against the 1990 census, testing the ability to generate a synthetic population with limited forecast information. In this process, it is assumed that the forecast input, though limited in amount and detail, is correct. In other words, the procedure does validate the synthesizer, but does not validate the land use model forecasts.

It validates by calculating aggregate characteristics of the synthetic population, and by calculating the same characteristics directly from the detailed census tables. It then compares them to see how well they match. There are four levels of geographic aggregation: tract, PUMA, County, super-county. Reports are then repeated for multiple synthetic populations in order to identify the variability caused by the Monte Carlo draws used in the synthesizer.

As for the software used, it is Object oriented Java, version 1.5, and consists entirely of subprograms called classes. Each class consists of member objects, that is, the information it holds, and methods, functions it can accomplish. Each class can be individually coded and tested. The PopSyn has four major groups of classes tied together by PopSyn class.

## DEVELOPMENT STATUS AND VALIDATION OBJECTIVES

The initial programming of the ARC population synthesizer is complete. Some improvements are known to be needed, including (a) improving the quality of the rounding procedure used after IPF before drawing the households from PUMS, (b) enhancing user friendliness, and (c) adjusting it to accommodate the recently expanded 20 county geographic scope. Enhancements would also be advisable to take advantage of enhanced inputs that may become available from the economic and land use models. Using the current synthesizer, base year and backcast synthesis have been tested and preliminary validation results have been produced.

The ARC population synthesizer allows the user to implement a variety of versions without reprogramming. For the initial testing and validation, we created three versions, the simplest with 52 household demographic categories, and the others with 128 and 316 categories respectively. As more categories are used, more detail can be used from the census tables (base year) or ARC demographic and land development forecasts (forecast year) to control the synthesis procedure, so more household attributes should be synthesized precisely. However, the computation takes longer, an increase in the number of sparsely populated categories causes more rounding error, and the use of regional values and averages for the additional controls might increase the noise and introduce bias. So one of the primary purposes of the validation is to choose the best version of household categories, and preliminary conclusions are reported below. The three versions are shown in Table 1, in terms of their number of categories within each of six dimensions. We will refer to them by their overall number of categories:

Table 1—Three basic PopSyn versions

| Dimension | Number of Categories | | |
|---|---|---|---|
| | Simple | Middle | Complex |
| Overall | 52 | 128 | 316 |
| Household income | 4 | 4 | 4 |
| Household size | 4 | 5 | 5 |
| Number of workers in HH | 4 | 4 | 4 |
| Family or non-family | 1 | 2 | 2 |
| Age of householder | 1 | 1 | 2 |
| Presence of own children under age 17 | 1 | 1 | 2 |

Validation allows us to better understand at which level of geographic detail the aggregate population attributes can be trusted, and which household variables are synthesized well enough to use in the travel forecasting models. Results of this analysis are reported below. We also report on the testing of other setup parameters, including the convergence criterion for iterative proportional fitting, and the aggregation level used in the seed distribution for the forecast year.

Validation can also be used to evaluate the level of variation in results caused by the stochastic nature of the simulation procedure used to generate the synthetic population. Several base year runs have been made with the same version, allowing stochastic variation. The results indicate that stochastic variation is probably not a problem, but detailed analysis of this variation has not been conducted and is therefore not reported here.

## COMPUTATIONAL PERFORMANCE

Table 2 shows computational performance for base year synthesis with the three versions, each synthesizing 3.6 million persons in 1.35 million households. Computational performance for forecast year synthesis is similar. The performance tests were run on a Pentium 4 computer with a 3 GHz processor and 2GB of memory. Regardless of version, three minutes of overhead are required to set up for synthesis, it takes over a minute to produce the validation statistics (if desired), and over two and a half minutes are required to save the synthetic population. However, version 316 requires much more time for other parts of the process, especially the iterative proportional fitting (IPF) procedure, so that overall run time of version 316 is nearly twice that of version 52.

The Table 2 results come from runs in which the IPF stopped when all cells changed less than 5%. Reducing the stop criterion to 0.5% doubled the required iterations, but increased the total run time by less than 5%.

Table 2—Base year computational performance of versions 52, 128 and 316

| Household Categories | 52 | 128 | 316 |
|---|---|---|---|
| Balancer IPF iterations | 7 | 11 | 14 |
| Total running time (minutes) | 9.9 | 11.9 | 17.4 |

## VALIDATION RESULTS

In this section we examine the precision and accuracy of household and person variables included in the synthetic population, for both the base year and backcast. As used here, the word accuracy refers to statistical bias; a variable with a non-zero mean percentage difference between the synthetic population and the census validation value is considered inaccurate. The ‘percentage difference’ referred to is the percentage difference between synthetic value and census value for a single geographic unit (tract, PUMA, county or supercounty). The ‘mean percentage difference’ is the average of this difference across all the geographic units in the region. Precision refers to statistical variance; a variable with a large variance in the difference between the synthetic population and the census validation value is considered imprecise. The order in which variables are discussed below corresponds roughly to the decreasing level of detail in which forecast controls are applied. Figure 1 provides a graphical presentation of selected variables relevant to the text discussion.

**Figure 1--Selected validation results**

([c] indicates variable controlled at TAZ level; numbers after label correspond to variable number in complete validation output)

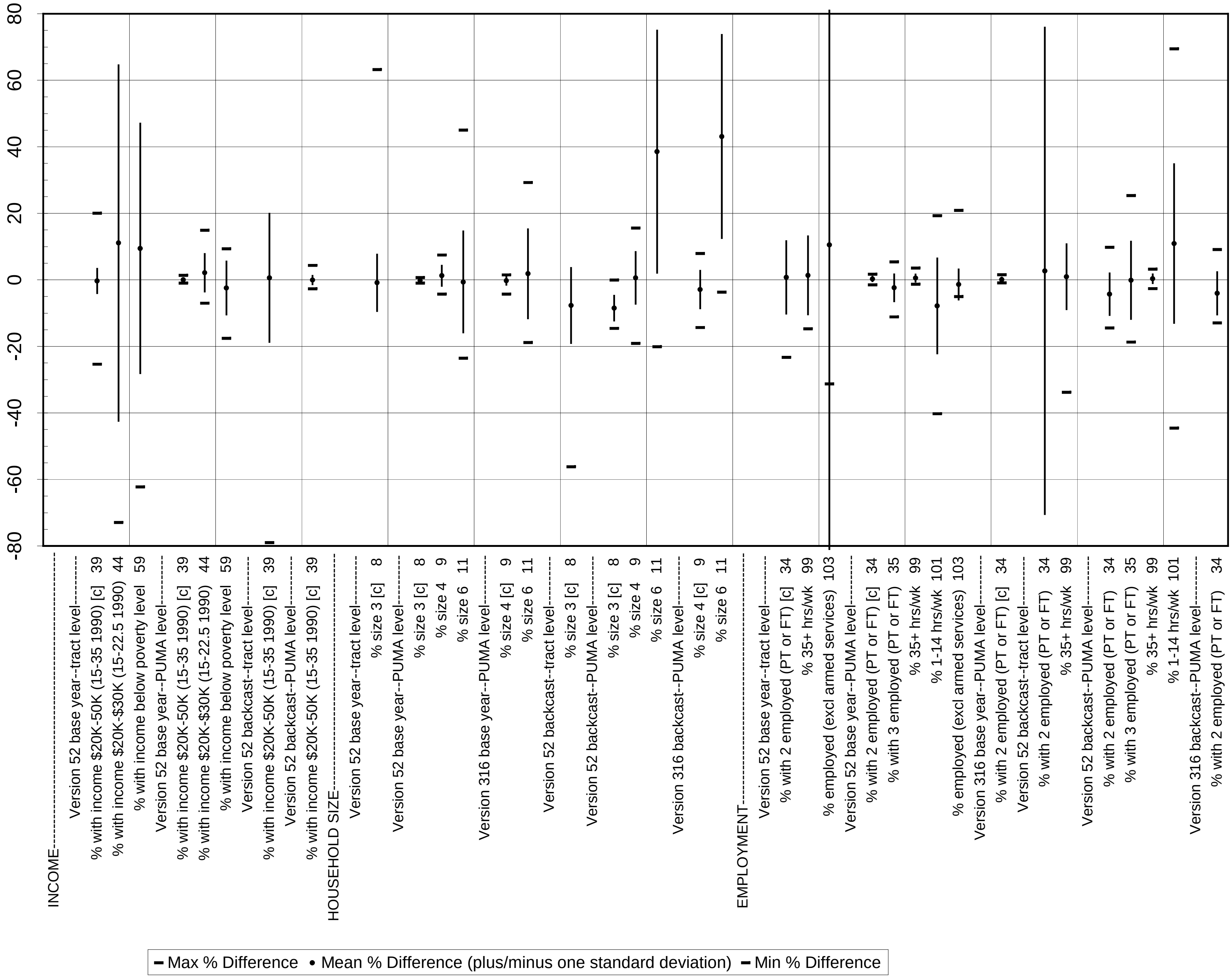
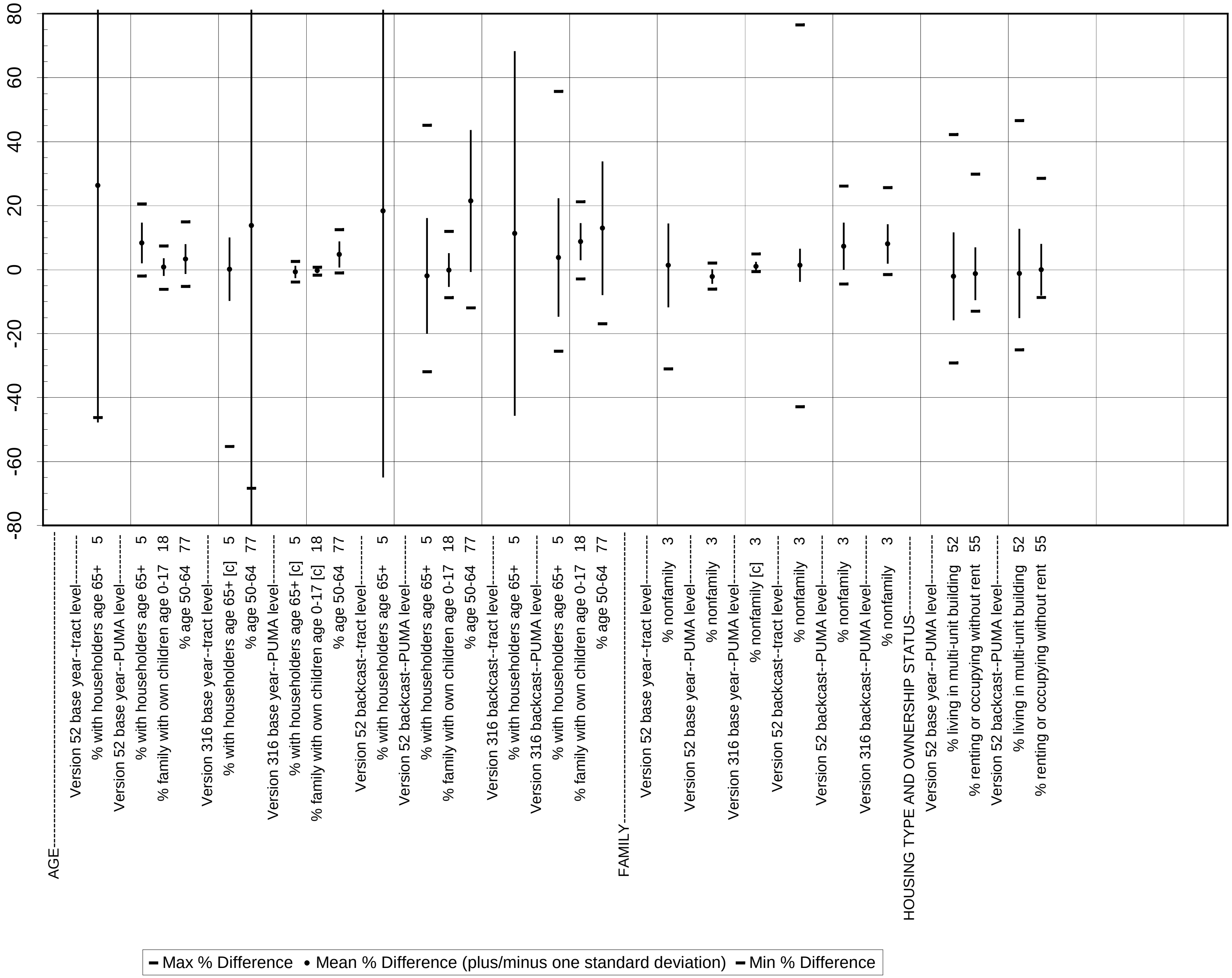


**Figure 1--Selected validation results (continued)**



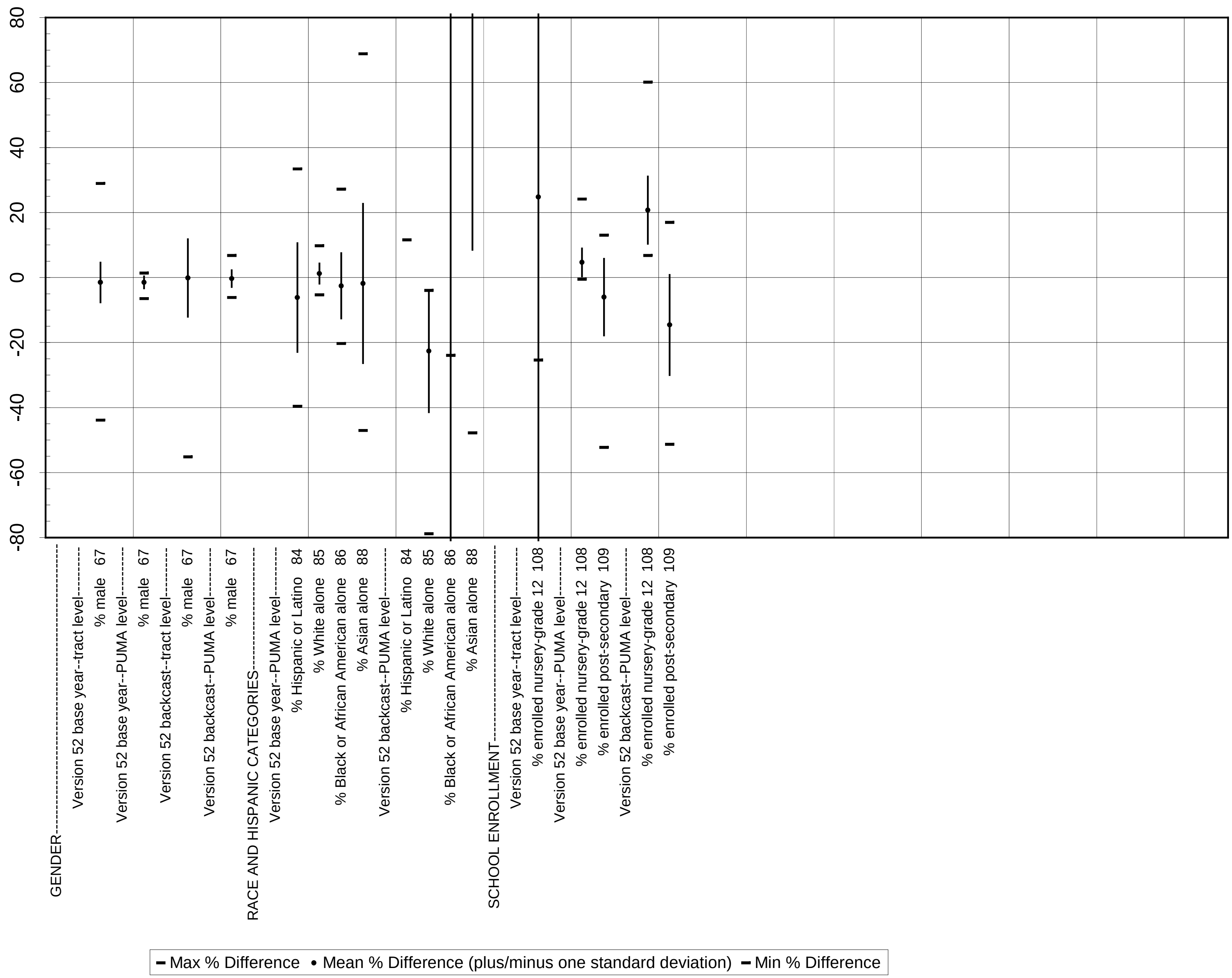



**Figure 1--Selected validation results (continued)**

**Income**. Because household income is controlled at the TAZ level in four categories for the base year and forecast year, for all three versions, it should be the most precise and accurate of all the variables, and indeed it is. Precision is slightly higher in the base year. Version 128 is over-synthesizing low income households; this probably indicates a minor bug in the setup inputs that should be found and corrected if that version is chosen for use. The precision and accuracy of uncontrolled income subcategories are noticeably worse, but could be judged as good at the PUMA level of aggregation. The backcast results in the uncontrolled subcategories cannot be correctly evaluated because of inconsistencies in the subcategory definitions between the 1990 and 2000 census years.

The census PUMS data also includes a personal variable comparing personal income to the official poverty level. The percentage of persons below the poverty level is synthesized imprecisely at the tract level, but is otherwise reasonably accurate and precise in the base year. The results cannot be validated in the forecast year because of changing poverty level definitions and dollar values between census years.

**Household size**. Household size is controlled at the TAZ level. In the base year it is controlled in 5 categories for versions 316 and 128, and in 4 categories for version 52. Household size is controlled at the TAZ level in the forecast year, but only average household size is available, and the base year distribution is used to translate this into the controlled categories. In the base year, the controlled sizes are very precise and accurate; the uncontrolled household size 4 in version 52 is noticeably less precise, but quite accurate. The uncontrolled size categories with very few households, such as size 6, achieve much less accuracy and precision, although accuracy is better in the versions that control 5 categories. The backcast validation procedure yields important results. First, noticeable inaccuracy arises from the use of average household size to generate the forecast control. Second, for version 52, the precision and accuracy of the uncontrolled household size 4 category are not noticeably worse than the four controlled sizes. Third, the precision and accuracy of version 52 are not worse than versions 128 and 316. So, given that the forecasts are only available as averages, controlling 5 size categories instead of 4 yields little or no improvement in the forecast year. The benefits in the forecast year of more size categories would probably be much greater if ARC could forecast a household size distribution rather than an average household size.

**Employment**. Employment is controlled as number of workers in household, in four categories, with TAZ-level control in the base year. In the forecast year, the control is enforced only for the region overall, and only average number of workers per household is supplied, so the forecast year control is really quite weak. In the base year, the controlled worker categories are very accurate and precise, and uncontrolled subcategories are noticeably less accurate and precise. In the backcast, we see again that inaccuracy is induced by deriving controls from averages, and quality of the largest uncontrolled category (3 workers) is not worse than the controlled categories. Also, using the regional control degrades the tract-level precision. There would possibly be much to gain from trying to estimate a distribution of households by number of workers rather than only an average number of workers per household, and the benefits would probably be much greater if this could be done for geographic units smaller than the region.

The employment status of each person is also in the synthetic sample. In the base year, the categories of ‘employed civilian’ and ‘not in labor force’ are precise and accurate at the PUMA level, apparently because of the household-based employment controls, although they are very imprecise at the tract level. The backcast validation values taken from the census are incorrect (for a yet undetermined reason), so the validity of personal employment in the backcast cannot be determined.

For employed persons, the hours worked per week is also recorded in the synthetic population. The category working 35 or more hours per week is very accurate and precise, even at the tract level, in the base year and backcast. For the categories of 15-34 hours per week and 1-14 hours per week the results are imprecise and inaccurate at the tract level, but reasonably accurate and precise at the PUMA level, in both the base year and backcast.

**Age**. Age is controlled only in version 316, with three categories, based on whether householder is over or under age 65 and, for those under age 65, whether own children under age 18 are present. For the forecast year, the control is supplied only as the regional sizes of the subpopulations aged 65+ and less than 15. The forecast control categories are sized by using relationships in the base year census PUMS data between the available values and the needed control categories. In the base year, the controls noticeably improve the accuracy and precision of the corresponding household categories. However, in the backcast, the version 316 controls provide no apparent improvement in accuracy or precision over the uncontrolled versions 52 and 128, and, for the presence of children under age 18, version 316 is less accurate. Apparently, the method of transforming the population estimate into the control category induces bias in the forecast.

Examination of base year validation statistics for age categories of persons in the synthetic population shows that controlling households by age of householder in version 316 may provide a small improvement in the accuracy and precision of person ages for the major categories of 0-17, 18-64 and 65+. For detailed age subcategories, results are unusably inaccurate and imprecise at the tract level; at the PUMA level, precision and accuracy are more acceptable. In the forecast year the version 316 results differ from those of the other versions, but with similar accuracy and precision, and for all versions the quality degrades somewhat from the base year, but not a lot.

**Family**. Family is controlled in versions 128 and 316, but only in the base year. In the base year, the controls improve precision and accuracy, but without the control, version 52 precision and accuracy are still quite good, even at the tract level. The base year controls appear to have little carryover effect in the forecast, where the uncontrolled categories are no better in versions 128 and 316 than in version 52. Precision remains fairly high, but accuracy gets considerably worse.

**Housing type and ownership status**. The accuracy and precision of these variables, which are completely uncontrolled in the base year and forecast, might nevertheless be considered good enough to be usable at the PUMA level of geography, except for the tiny category of mobile home dwellers. The tract level results are too inaccurate and imprecise to be usable.

**Gender**. Although there are no controls related to gender, the number of males and females are fairly accurate and precise, regardless of version, in both the base year and backcast.

**Race and hispanic categories**. Although there are no controls related to race and hispanic categories, in the base year these are synthesized accurately and with reasonable precision at the PUMA level for hispanic, white, black and asian categories, but inaccurately and imprecisely at the tract level. For the other smaller racial categories the results are imprecise and inaccurate at all levels. The race data definitions changed from the 1990 census to the 2000 census, making it difficult to interpret the validation results, although it appears that the accuracy and precision of the backcast population are much worse than in the base year.

**School enrollment**. School enrollment in two categories—nursery-to-grade-12 and post-secondary—although inaccurate and imprecise at the tract level, is reasonably accurate and precise at the PUMA level in the base year. In the backcast, school enrollment is quite inaccurate and imprecise at all levels of geographic aggregation, but perhaps usable at the PUMA level.

**Additional validation results**

**IPF stopping criterion**. The above validation results come from test runs in which the IPF convergence criterion was set at 5%. Changing the criterion to 1% in backcast runs caused only a slight improvement in the mean percentage difference (e.g, mean percentage difference improves from from 4% to 3.9%), on average across the usable variables, at both the PUMA and tract levels of analysis.

**Forecast seed matrix**. When synthesizing a forecast year synthetic population, PopSyn uses, as its starting matrix for IPF, the balanced matrix from the base year synthetic population. Its starting distribution for each TAZ is a combination of the TAZ-level, tract-level and PUMA-level distributions. The exact combination depends on the sizes of the TAZ and tract relative to user-assigned parameters. If the TAZ (or tract) is smaller than the user-specified minimum, then it is not trusted to provide the starting distribution; if it is larger than the user-specified maximum, then it is trusted completely. In between, its distribution is blended with those of the larger geographies. The issue at hand is whether small neighborhood peculiarities persist over time. If they do, then it would be better to use the base year TAZ-level distribution, even for small TAZ, preserving the details supplied by the base year census tables; if they don't, then it would be better to use the distribution from the tract or PUMA.

To test this, version 316 backcasts were run with a variety of minimum and maximum size criteria. The quality of the validation results were then compared by averaging the absolute mean percentage difference and the standard deviation percentage difference across all usable variables, and comparing them across runs. The results indicate that the backcast population matches the backcast validation values best when the minimum size is between 10 and 100, and the maximum size is between 100 and 500. The results are worst when the size parameters are set so high that the PUMA distributions are used exclusively. However, except for this extreme case, difference are very minor compared to the levels of inaccuracy and imprecision in the best forecasts.

## SUMMARY OF VALIDATION RESULTS

The following summary conclusions might be drawn from the above analysis, about the preferred versions to use for base year and forecast analysis:

1. In the base year, the use of census data to control for more variables in version 316 yields a clearly superior synthetic population, especially for tract level evaluation in controlled categories. So, for base year analysis, and short term forecasts using the base year population, version 316, or perhaps even a more complex version, should be used.

2. For the forecast year, the additional controls of versions 128 and 316 provide little value, and can potentially make the population worse. The reason for this lies primarily in the reliance on averages that are translated into category distributions naively, based on base year distributions, rather than attempting to make informed forecasts of the distributions themselves. It probably also lies in relying on regional forecasts, rather than forecasts carrying information at some smaller level of geographic aggregation.

Table 3 provides a summary of the aggregation level at which various categories of variables would have a reasonably high level of precision and accuracy in the synthetic population, assuming version 316 for base year analysis and version 52 for forecast year analysis.

Table 3—Aggregation levels at which variables in synthetic population have reasonably high accuracy and precision

| Variable | Base Year Usage (version 316) | Forecast Year Usage (version 52) |
|---|---|---|
| Household income (major control categories) | tract | tract |
| Household income (subcategories) | PUMA | PUMA |
| Person poverty status | PUMA | ? |
| Household size (major control categories) | tract | PUMA |
| Household size (subcategories) | tract | PUMA |
| Household workers (major control categories) | tract | PUMA |
| Household workers (subcategories) | PUMA | PUMA |
| Person employment status | PUMA | PUMA |
| Person weekly work hours (35+ category) | tract | PUMA |
| Person weekly work hours (other categories) | PUMA | PUMA |
| Household with holder age 65+ | tract | PUMA |
| HH presence/absence of own children age 0-17 | tract | PUMA |
| Person age category | PUMA | PUMA |
| HH family status | tract | PUMA |
| HH housing type (major categories) | PUMA | PUMA |
| HH housing ownership (major categories) | PUMA | PUMA |
| Person gender | tract | tract |
| Person race and Hispanic status | PUMA | ? |
| Person school enrollment category | PUMA | ? |

These preliminary results demonstrate several important aspects of population synthesizers. First, the accuracy of synthesized characteristics depends heavily on the control variables used for population synthesis; uncontrolled variables are synthesized much less accurately, even in the base year. Second, the accuracy drops when results are examined at a more detailed level of aggregation. Third, even for controlled variables, the accuracy is not perfect; in the ARC base year case, the rounding procedures used after IPF, before the households are drawn, introduce a substantial amount of noise; and in the forecast case, the use of averages, variables that don't match the IPF categories, and regional values all degrade accuracy and/or precision. The accuracy and precision of the forecast population are less than those of the base year population, even assuming that the forecast inputs are accurate, because the quantity of forecast controls is smaller, and they are more aggregate than the base year controls. We conclude that it is indeed important to implement validation procedures that provide the user of a population synthesizer with the information needed to use it appropriately. It is also valuable to implement a flexible population synthesizer that can be adjusted and improved in response to the validation information. With this version of the PopSyn in hand, ARC is in a good position to continue validating and improving it, even as it incorporates it into the demand models and uses it for analysis.