


TMIP Webinar Series



Activity-Based Modeling

Session 5: Population Synthesis and Household Evolution

The **Travel** Model
Improvement
Program

Speakers: John Gliebe & Peter Vovsha

April 26, 2012

Acknowledgments

This presentation was prepared through the collaborative efforts of Resource Systems Group, Inc. and Parsons Brinckerhoff.

- Presenters
 - John Gliebe, Peter Vovsha
- Moderator
 - Stephen Lawe
- Content Development, Review and Editing
 - Bhargava Sana, John Gliebe, Peter Vovsha, John Bowman, Mark Bradley, Joel Freedman, Maren Outwater
- Media Production
 - Brian Grady




Resource Systems Group and Parsons Brinckerhoff have developed these webinars collaboratively, and we will be presenting each webinar together. Here is a list of the persons involved in producing today's session.

- John Gliebe and Peter Vovsha are co-presenters. They were also primarily responsible for preparing the material presented in this session.
- Stephen Lawe is the session moderator.
- Content development was also provided by Bhargava Sana, Joel Freedman, and Maren Outwater. John Bowman and Mark Bradley provided a review of the material.
- Brian Grady was responsible for media production, including setting up and managing the webinar presentation.

2012 Activity-Based Modeling Webinar Series

Executive and Management Sessions	
Executive Perspective	February 2
Institutional Topics for Managers	February 23
Technical Issues for Managers	March 15
Technical Sessions	
Activity-Based Model Frameworks and Techniques	April 5
Population Synthesis and Household Evolution	April 26
Accessibility and Treatment of Space	May 17
Long-Term and Medium Term Mobility Models	June 7
Activity Pattern Generation	June 28
Scheduling and Time of Day Choice	July 19
Tour and Trip Mode, Intermediate Stop Location	August 9
Network Integration	August 30
Forecasting, Performance Measures and Software	September 20

Activity-Based Modeling: Activity-Based Model Basics



3

For your reference, here is a list of all of the webinars topics and dates that have been planned. As you can see, we are presenting a different webinar every three weeks. Three weeks ago, we covered the fourth topic in the series—Activity-Based Model Frameworks and Techniques. This session provided some of the concepts fundamental to activity-based modeling, including how activities and tours are represented in data; and how various choices structures may be used to model particular dimensions of activities, tours, and travel patterns. We also covered important aspects of discrete choice modeling and model implementation using Monte Carlo simulation methods. In the second half of that webinar, we discussed activity-based model design and highlighted some of the key trade-offs that model developers consider. Today's session is the second of nine technical webinars, in which we will cover population synthesis and household evolution models.

Learning Outcomes

By the end of this session, you will be able to:

- Describe a synthetic population
- Describe the methods used to synthesize a population
- Describe the process of household evolution

In today's session, we will be covering the basics of activity-based modeling. At the end of this session you should be able to:

- Describe a synthetic population
- Describe the methods used to synthesize a population
- Describe the process of household evolution

Session Outline

- Role of synthetic populations in activity modeling
- Specifying a synthetic population
- Methods used to create a synthetic population
- Synthetic populations in practice
- Household evolution models

In this webinar, we will describe a synthetic population and how its place in the overall structure of an activity-based model system. This will be followed by a discussion of considerations for specifying a synthetic population generator. Next, we will cover the methods used to create a synthetic population. We'll then look at some examples of synthetic population generators that have been implemented. Finally, the last part of today's webinar will introduce a class of emerging methods in population synthesis known as household evolution models.

Terminology

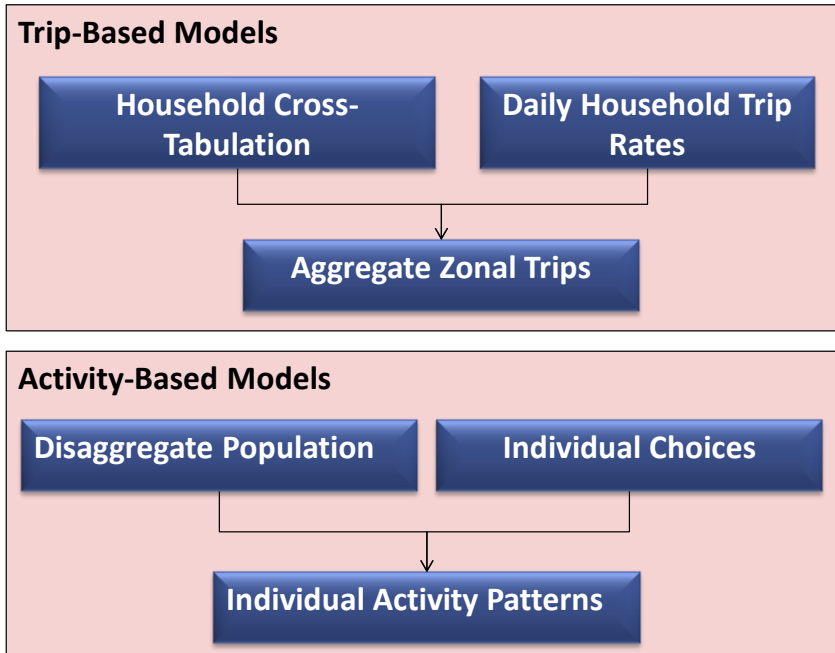
- Socio-demographic attribute
- Controlled attributes
- Disaggregate household sample
- Seed data
- Iterative Proportional Fitting (IPF)
- Drawing a sample
- Uncontrolled attributes
- Household evolution

In today's webinar, we will discuss some of the essential methods in population synthesis. First, we'd like to define some of the terms that we will be using frequently throughout the presentation.

- **Socio-demographic attribute:** A household or person characteristic that is maintained for each household. Examples of household attributes would include the number of persons (household size), number of workers, presence of children, age of householders, and household income. Examples of person attributes might be age, gender, worker status and students status. Of course, there may be others.
- **Controlled attributes:** Attributes that are chosen as segmentation targets. We will refer to the total number of households or persons that fall into each control attribute category as the marginal control totals.
- **Disaggregate household sample:** Observations of individual household and person attributes.

- **Seed data:** Sample of correlation structure between household or person attributes, usually based on the disaggregated household sample.
- **Iterative proportional fitting:** Algorithm to balance to produce the final aggregate distribution of households in the synthetic population to match the marginal control totals, while maintaining the correlation between attributes that is represented by the seed data.
- **Drawing a sample:** Selecting households from a disaggregate sample to match the final aggregate distribution of households in the synthetic population.
- **Uncontrolled attributes:** Other attributes of households and persons, which are of interest, but which we do not explicitly control in terms of their distribution. They are present in the disaggregate sample and obtained through the drawing process.
- **Household evolution:** A general class of advanced methods of forecasting a future population by aging a household through time by predicting household births, deaths, marriages, household formation and household dissolution.

Aggregate vs. Disaggregate Travel Representation



First, it might be instructive to review how populations are represented in trip-based models and contrast them with disaggregate activity-based models

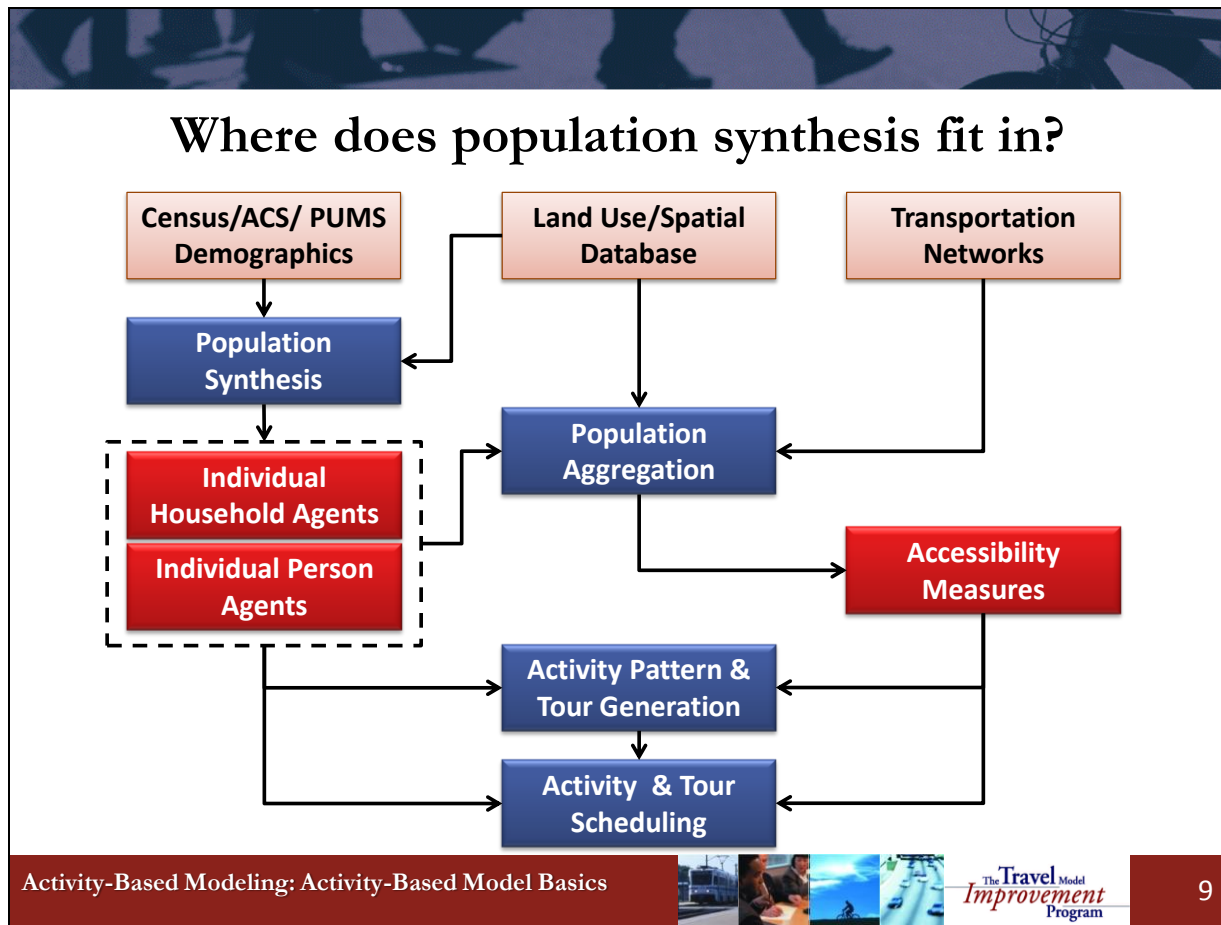
In trip-based models, we are dealing with aggregate households grouped in transportation analysis zones (TAZ). We apply trip production rates to groups of households, and trips are generated and aggregated by TAZ.

In activity-based models, we maintain households in disaggregate form. Individual choices are simulated, and individual activities, trips, tours are generated and scheduled.

Role of Population Synthesis

- Households and persons are represented individually in the activity-based model through micro-simulation
- Population synthesis “creates” these households and persons for use in the activity-based model
 - Synthetic households and persons should possess all of the demographic attributes needed for model inputs
 - Variables that will be used to explain variation in ...
 - Daily activity patterns and tour generation
 - Levels of participation in various activities
 - Preferences for time of day, mode and destination attributes
 - Value of time (willingness to pay)
 - Coordination between household members

Synthetic populations are essential to simulating individual activity-travel patterns. The design of the synthetic population should support the design of the activity-based model and provide the variables it needs. In addition, the activity-based model should only rely on information that can be realistically provided in the synthetic population.

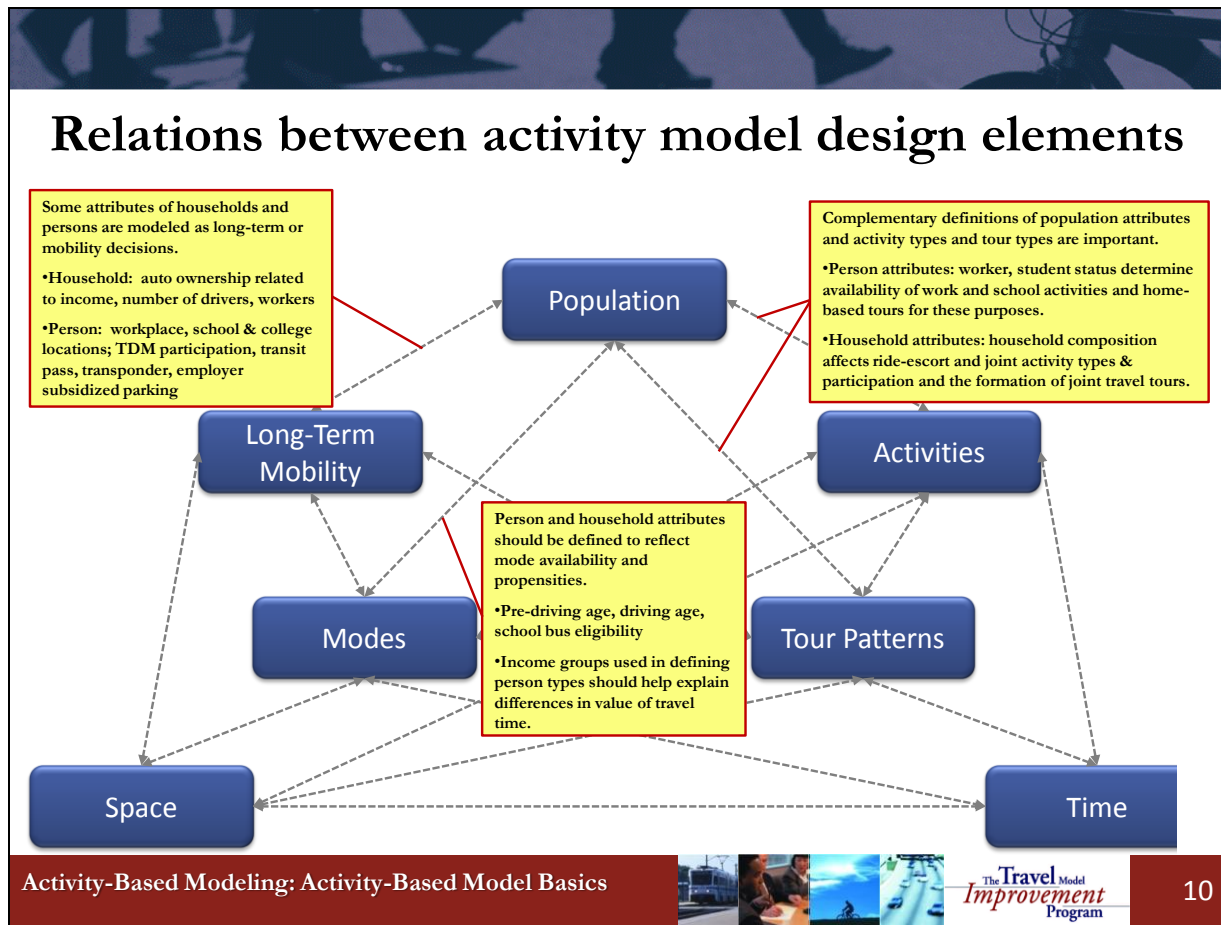


In an activity-based model, population synthesis is a modeling step, much like activity pattern and tour generation and activity and tour scheduling steps. Census-related data is combined with geographic data to create synthetic households, with specific locations throughout the region of interest.

Population synthesis is the first step in the modeling process and produces individual household agents and individual person agents that are the subjects of the simulation. These household and person agents “make decisions” that through the activity generation and scheduling process.

Since synthetic households/persons are situated spatially, they may be combined and aggregated at TAZ level or some other spatial units. Some of the accessibility measures used in activity pattern and tour generation and scheduling models are derived from aggregations of households within a buffer region, usually specified to consider transportation network distances, travel times and costs. These accessibility measures, which we will discuss at length in the next

webinar, also appear as explanatory variables in activity pattern and tour generation and scheduling models.



Let's consider how decisions regarding the synthesis of a population for the purposes of simulation might affect other aspects of model system design.

This diagram serves as a backdrop for describing the relationships between key design elements in activity-based modeling. These elements include: defining the population, modeling long-term and mobility-related choices, defining activity types, defining modes, defining tour patterns and an entire day-pattern elements, as well as the treatment of space and accessibility and treatment of time. We discussed each of these design elements in the previous webinar on activity-based modeling frameworks and techniques, and we will devote an entire session to each one of these elements starting today (population), and over the next six webinars.

First, I mentioned earlier that the method of population synthesis usually provides only the basic characteristics of the population. It is usually desirable to model other attributes of households and persons as long-term or mobility decisions. This enables us to reflect policy-sensitivity into these choices. For example, household auto ownership is related to income, number of drivers,

and presence of workers. It may also be related to some notion of accessibility, which we will talk about in later webinars, that is derived from the level of service provided by the transportation network. In addition, we typically want to model certain person attributes as long-term decisions, such as workplace, school and college locations. Others may be thought of as mobility decisions, such as whether to participate in a TDM, buy a transit pass, buy a transponder, or whether an individual will benefit by employer-subsidized parking.

Second, complementary definitions of population attributes and activity types and tour types are important. For example, person attributes, such as whether a person is a worker or student, determine the availability of work and school activities and home-based tours for these purposes. In addition, household composition variables and how they are represented in the synthetic population (presence and ages of children) determine the ability to model household drop-off and pick-up events, carpooling. They also affect the propensity and structure of joint activity types and participation, and the formation of joint travel tours.

Finally, population attributes impact mode availability. It is important to identify persons of driving age. Likewise, household income attributes may be used to segment persons by their value of time... or willingness to pay for travel time savings.

Role of Person Types

No.	Person Type	Age	Work Status	School Status
1	Full-time worker	18+	Full-time	None
2	Part-time worker	18+	Part-time	None
3	Non-working adult	18 – 64	Unemployed	None
4	Non-working senior	65+	Unemployed	None
5	College student	18+	Any	College +
6	Driving age student	16-17	Any	Pre-college
7	Non-driving student	6 – 16	None	Pre-college
8	Pre-school	0-5	None	None

From San Diego ABM

- Model segmentation
- Summarize outputs
- Explanatory variables in models
- Constraints on available alternatives



We showed this slide in the last webinar on activity modeling frameworks and techniques, but it is worth reviewing again here because it provides a good summary of the role that person and household attributes play in the overall model design. Although we are modeling persons in disaggregate form, it is often useful to create person-type categories. This table shows an example of person-type categories and definitions. Person-type categories can be used for a number of purposes:

- As a basic segmentation for certain models, such as daily activity pattern models.
- To summarize and compare observed versus estimated data and calibrate models.
- As explanatory variables in models.
- As constraints on alternatives that are available; for example, work and school activities are only available to workers and student; and driving is restricted by age.

Although continuous values of age and income are available, properly grouped categorical variables often result in a better model fit and efficiency. Determining the proper cutoff points

for categorical variables related to age and income are usually derived in the model development process empirically through descriptive statistical work. For example, a community with a large population of retirees might use different age groupings than a college town. Not having proper cutoff points could result in important market segments being under-represented in model specifications and estimated models not explaining as much behavioral variation in the population as they could.

Let's look at a realistic example of a policy study and the specification of population attributes, such as income groups, might affect our analysis.

Bridge Expansion Example

- No Build Alternative
 - 4 lanes (2 in each direction, no occupancy restrictions)
 - No tolls
 - Regional transit prices do not change by time of day
- Build Alternative(s)
 - Add 1 lane in each direction (total of 6)
 - New lanes will be HOV (peak period or all day?)
 - Tolling (flat rate or time/congestion-based)
 - Regional transit fares priced higher during peak periods

Let's consider a transportation planning and policy project that might be faced by an MPO or DOT and how population synthesis fits into the picture. We may revisit this example in each of the next several sessions to come and discuss how the topic of that session relates to this particular example.

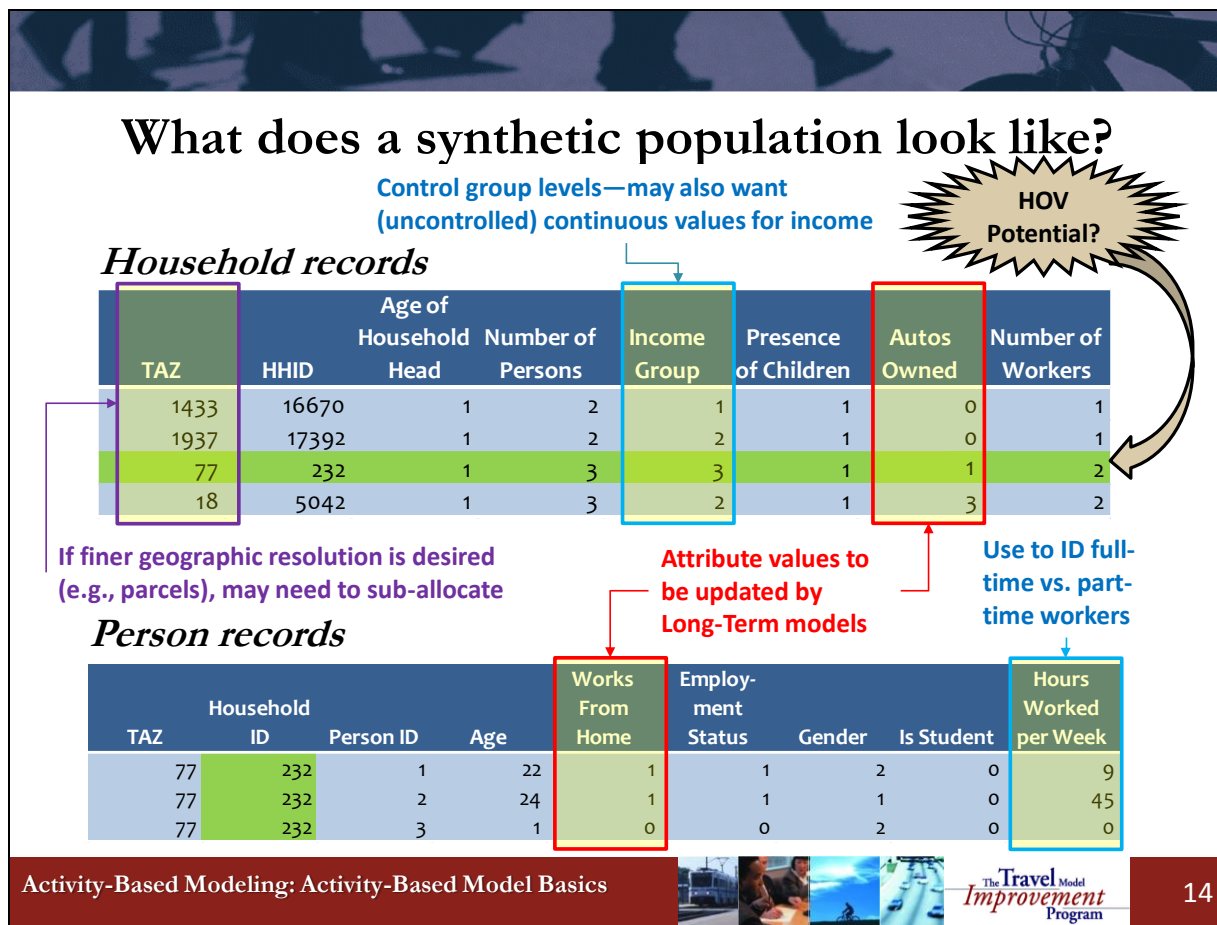
For this scenario analysis, we will be considering a number of alternatives: a no-build alternative and a various configurations of the build alternative. In the no-build alternative the bridge has 4 lanes (2 in each direction), there are no tolls, and the transit fare stays the same all day. In the various build alternatives, there are 6 lanes on the bridge. In some alternatives the two additional lanes will be HOV lanes all day, while in other alternatives the two additional lanes will be HOV lanes only during peak periods. In addition, in some build alternatives there will be a new toll that is the same across the entire day, while in other build alternatives there will be a toll that will be only applied during peak periods, or when certain levels of congestion occur. Finally, in the build alternatives regional transit fares will be higher during peak periods.

Bridge Expansion Example—Relevance to Population Synthesis

- Higher income households have a higher willingness to pay— affects distribution of project benefits
- Multiple-worker households are more likely to use HOV
- It is important that a synthetic population adequately represents all relevant sub-groups:
 - Low, medium and high income groups identified at appropriate “breakpoints”
 - The joint distribution of households by age, size, income and workers should match the real population... and by geographic sub-area
 - Household and person attributes should support prediction of relevant variables in other models within the model system:
 - Auto ownership, transit pass holders, telecommuters



For this bridge example, the synthetic population plays a major role in determining how demand will vary through the population. The set of methods that we will focus on in this webinar, generally known as “population synthesis” provide some of the answers. Of particular interest to this example, population synthesis provides direct representation of the income levels of individual travelers and household structure. Thus, it is important that the synthetic population covers the entire income range in the real population and, that those income levels are properly correlated with other important attributes, such as age of householders, household size, number of workers, and auto ownership. It is also important that households with these attributes be distributed properly in space. For example, we need to make sure that low-income households and zero-car households are placed in neighborhoods that match their occurrence in the real world.



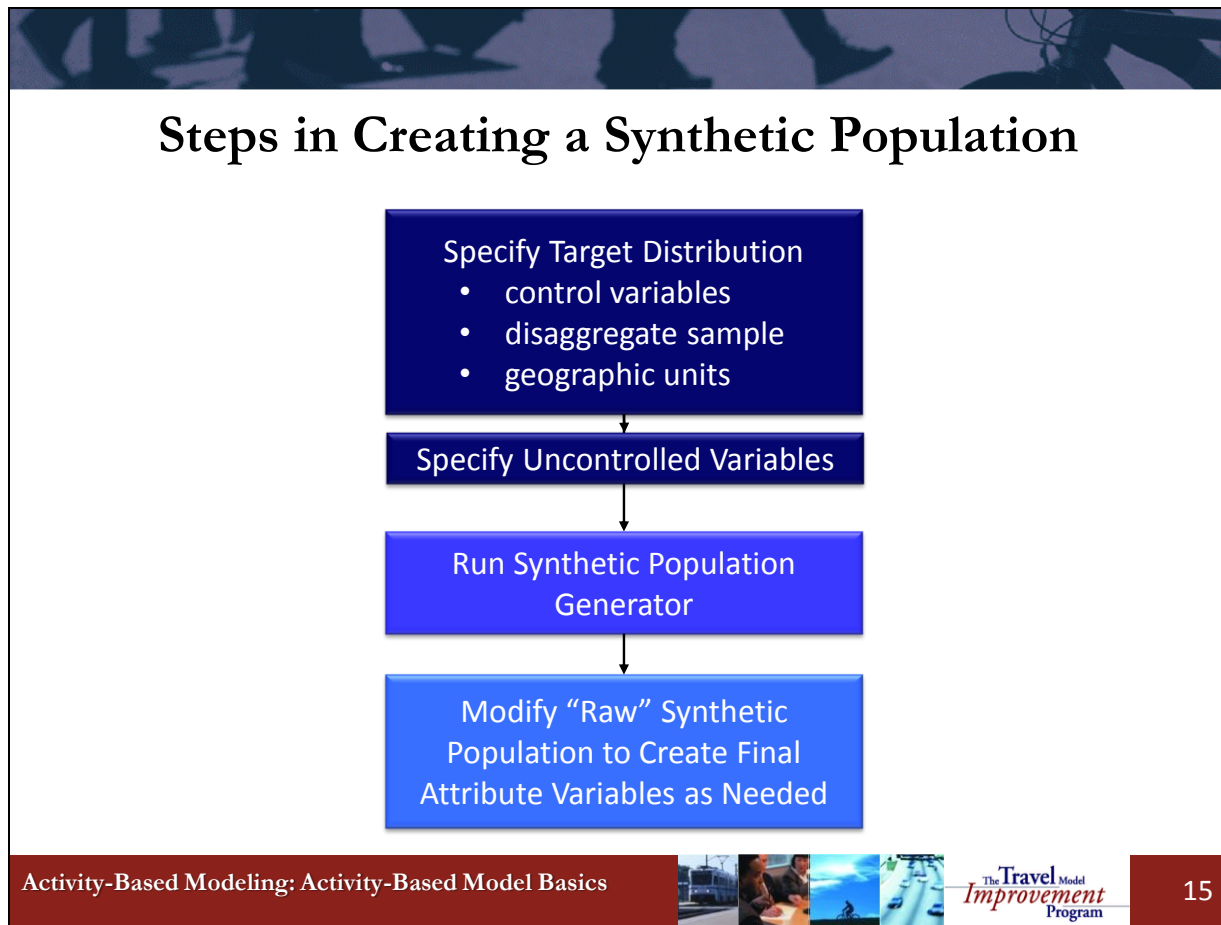
What does a synthetic population look like? This slide shows an example of output from a population synthesizer. Prior to their use in the simulation, synthetic populations are represented in data tables, often in a relational database or some equivalently structured file system. Typically there are separate tables for households and person records. The format of the data tables may be familiar to those who have worked with household travel surveys. Person records are linked to household records through ID numbers. Attributes of synthetic households and persons may be found in each record. The TAZ ID identifies the geographic location of this synthetic household. Putting the right households in the right locations is critical to model performance. The values of the other attributes are those that are typically available from the Decennial Census or American Community Survey (ACS) for control variables. For other variables, usually uncontrolled, the attribute values in the disaggregate sample are what is available. This disaggregate sample often comes from PUMS data, but could also be from a household survey. We'll talk more about data sources in a few minutes.

Some of these variables and values may need to be updated or modified in some way. For example, while it might be desirable to have household auto ownership generated by the population synthesis, and this is useful as a calibration check, we typically prefer to model household auto ownership or availability through a separate process. This is so that it can be made sensitive to transportation system accessibility and other important policy variables that might change in future scenarios. We will cover auto ownership along with other long-term household and medium-term mobility decision models in Webinar 7, two sessions from now. For our bridge expansion scenario, in addition to household auto ownership, we may also be interested in modeling whether an individual has a transit pass or a transponder, both of which are typically modeled as mobility decisions.

Similarly, we tend to model an individual's decision to work from home as a daily activity pattern alternative that might be influenced by not only demographic factors, but also transportation system and other policy variables. Activity pattern generation is the subject of Webinar 8, where you will see how important individual household attributes become in the prediction of daily tour patterns. As stated a few minutes ago, the bridge analysis will be greatly affected to by the number of commuters, and the potential demand for HOV usage. So, knowing whether someone is likely to telecommute or whether they live in a household with other workers, who might share rides, would be key inputs to that analysis.

Other variables may need to be augmented or transformed for our use. Here, we have used PUMS variables that indicate the number of hours a person works per week, which we might later transform to a variable that indicates whether a person works full-time or part-time. In this example, we control for household income by group and that is what shown, but we might also want to draw from our sample a continuous value of income. This might help us be a bit more precise in modeling incomes for the value-of-time calculations that are so important to the bridge study.

Finally, since some activity-based models are designed to utilize a level of geography that is finer in resolution than the spatial units used for control variables, there may need to be a separate sub-allocation process to place households on parcels.



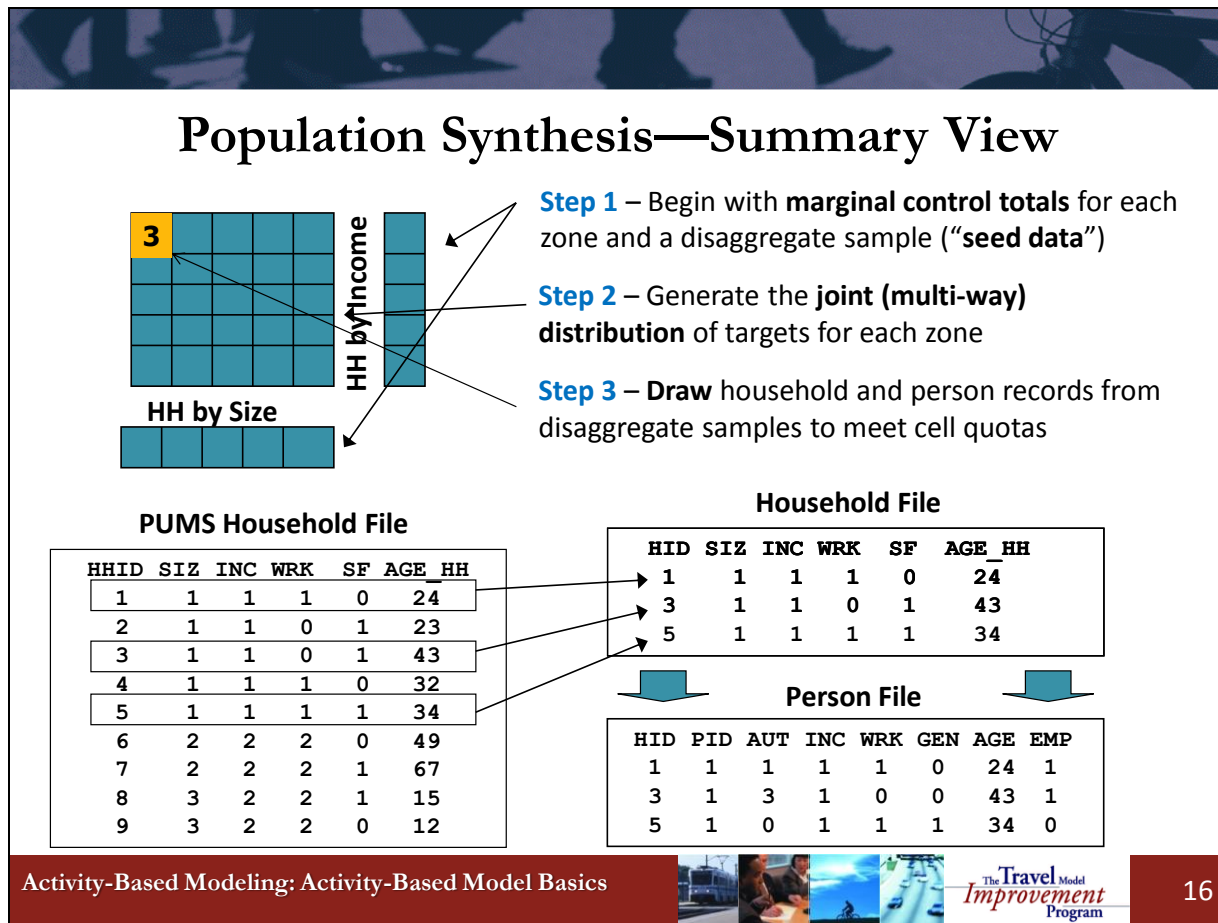
There are three major steps in creating a synthetic population, the first of which is specifying the inputs to the process—the control variables and sample households as well as the level of geographic resolution. Specifying the control variables is essential. In addition, there is often an additional step of specifying additional, uncontrolled variables to be added to the synthetic population.

The second major step is actually running a program that produces the synthetic households. Synthetic population generators may be packaged with an activity-based travel model. In addition, some standalone population generators are available, and we will talk about these a bit later.

The third major step would be transforming the model-generated outputs into characteristics of the population that will be used throughout the rest of the model system. This would be doing some of the things we just discussed, such as creating categorical variables out of continuous

variables, reformulating income, or allocating households from the zonal level to a finer level of geographic resolution, such as a parcel.

Next, we will talk about each one of these steps in more detail. In order for you to better understand the first step shown here (specifying a target distribution), you need to know what is coming downstream—how the synthetic population generator works.



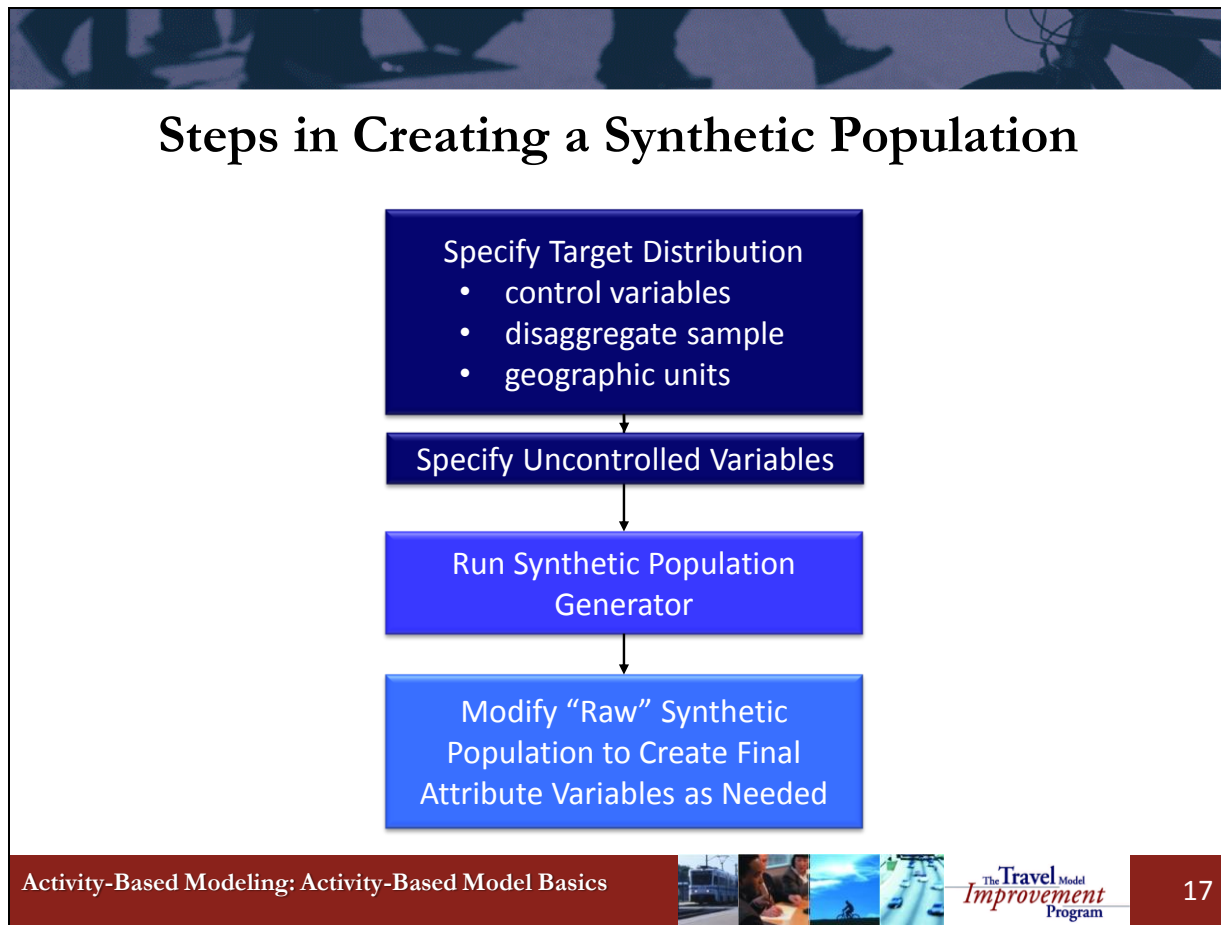
So, let’s briefly review how a population synthesizer works. We showed this slide during Webinar 4. It provides a good conceptual view of a general process.

In the first step, marginal control totals for each zone are created along with a sample of the correlation between attributes, usually from a disaggregate sample. It is common to refer to these control totals as “marginal control totals” because they represent the margins of the joint distribution. They are represented here as row and column target values. It is also common to refer to this sample data as “seed data” or the “seed matrix” if structured as a matrix, as in this example.

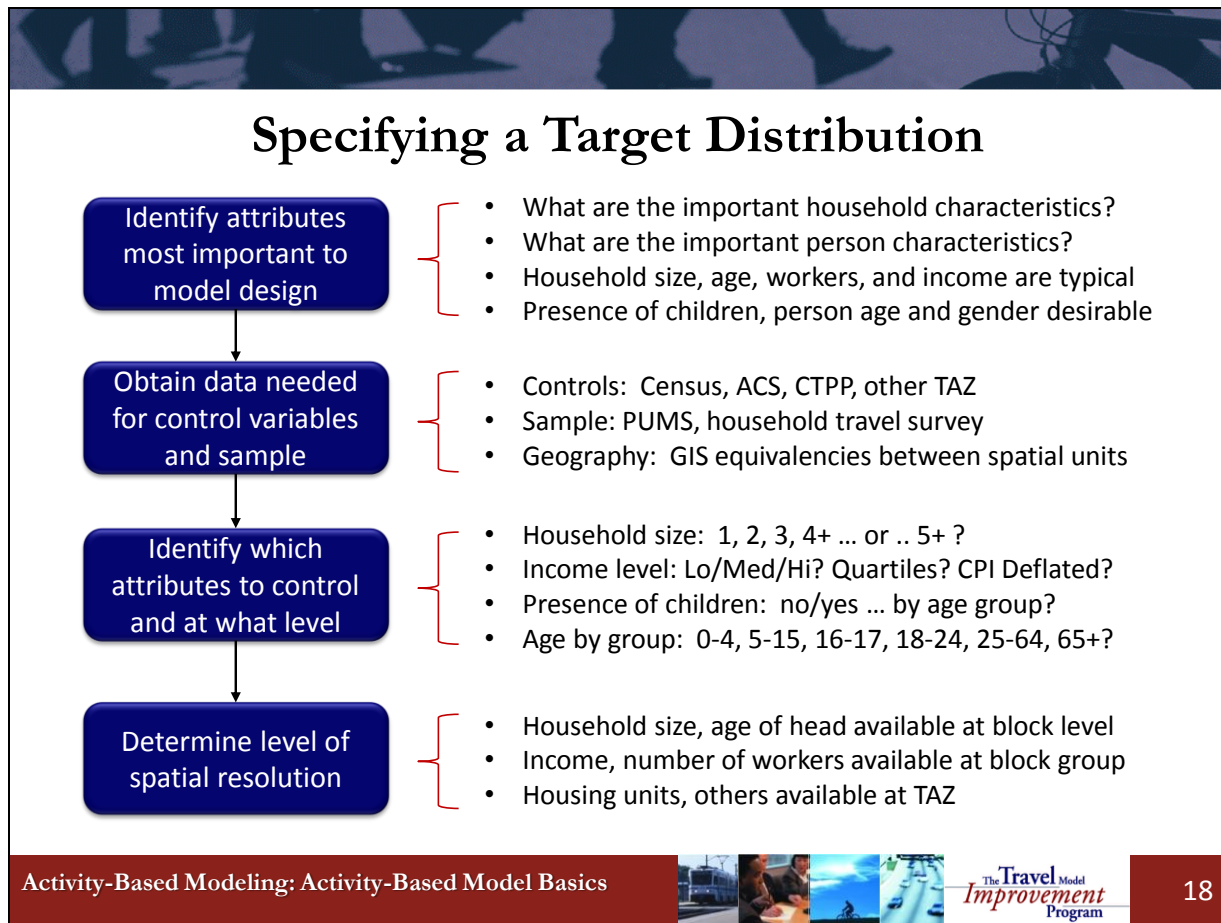
In the second step, the aggregate characteristics of the synthetic population are defined, in the form of a joint multi-way distribution of household attributes for each zone.

In the third step, individual person and household records are drawn from a sample of households, such that when we sum them up, they meet the joint distribution defined for the zone. So, the construction of joint multi-way distribution of households provides a target

distribution. Within each cell in the matrix are the number of sample households that match those control attributes that we need to draw from the disaggregate sample.



We will discuss more about the technical details of population synthesizers in few moments. First, we need to return to the top of this flow chart and cover the essential steps involved in creating the inputs to the population generator—specifying the target distribution through the selection of control variables. We will also talk a little about the use of uncontrolled variables.



The first step in population synthesis is to determine which attributes are most important to the model design and then to specify those as the variables to be “controlled.” This includes both important household and person attributes and should be policy-driven; however, over the years a fairly standard set of variables has emerged, namely household size, age of householder, number of workers, and income. The presence of children and person-attributes for age and gender are also desirable and fairly common.

The next step is to obtain the data needed to summarize the distribution of these attributes in the population. This might include Census, American Community Survey (ACS), CTPP, or other TAZ-level control totals.

It is important to determine how to group the data, so that the resulting population covers the sample space. What is shown here are just examples. For instance, we might specify age group control values for person attributes, so that we make sure we draw samples that include pre-school age children (0-4), school-age non-driving children (5-15), driving-age high-school

students (16-17), college-age adults (18-24), and senior citizens (65+). For some regions, other groupings might make more sense. In addition, specifying income levels can be tricky, because of the need to ensure representation from low and high-income households and to adjust for inflation. In our bridge example and other studies involving equity analysis, income specification is important to estimating not only willingness to pay for travel time savings, but also for identifying how different income groups benefit from a proposed project or policy. Since the model will be used for multiple purposes, it is important not to be too narrow in focus to avoid specifying variables that will limit other analyses.

Typically, the smallest level of spatial resolution that can be feasibly and reliably used to control attributes is used. This is usually a function of available data (e.g., Census blocks, block groups, TAZs). In recent application, control totals are grouped at the TAZ level, which makes things easier for forecasting. In addition to the data itself, it is typically necessary to use GIS layers representing blocks, block groups, TAZs and potentially other spatial units to develop equivalency tables.

Data Sources for Control Data

- Base year
 - Decennial Census: ~100% sample
 - American Community Survey (ACS) summary files:
 - 3% sample, rolling 5-year sample, yields an estimate of ~15% of pop.
 - Other zonal data developed locally (TAZs)
 - Census Transportation Planning Products (CTPP)
 - TAZ-Census Block/Group geographic equivalencies
- Forecast year
 - Regional socio-economic forecasts, growth factor models, allocation processes
 - Output from a land-use model

A bit more needs to be said about data sources. Here again we've listed some of the typical sources used in the U.S.

They fall into two general categories based on how they will be used in the population generator:

Sources of control data, which include full-population estimates/counts/forecasts at the desired level of geographic precision

Sources of disaggregate household/person sample data, which provide a basis for estimating the correlation between attributes

First, we will talk about data for specifying the target distribution, the control data. Base year data might include: the decennial Census (2010) is now available, which provide a 100% sample of the population and is available in summary form as fine as the block level for certain attributes of interest, such as household size, age of householder, presence of children; and person

attributes, such as age and gender. The smaller the geographic units, the closer the synthetic population should be to the true population.

Other attributes, such as households by income group and number of workers may only be available in summary form from the American Community Survey (ACS), which has become an annual 3% sample, that is available in increments of 1, 3 and 5 years.

In addition, some MPOs may develop their own socioeconomic data locally, using some combination of sources. For example, this could include households living in single- and multi-family housing unit types, as well as seasonal households.

The Census Transportation Planning Products (CTPP) can be used and is under preparation for 2010. It typically includes ACS level data and TAZ-Census geographic equivalencies.

TAZ-Census geographic equivalencies may be developed through GIS overlay analysis if they have not been created through CTPP or some other source.

Forecast year control data may be derived in any number of ways. The Census is not available today for future years, so official regional socio-economic forecasts may be used or simple growth factor models, in combination with some type of allocation process to produce TAZ-level forecasts. In some regions, a land use model may be used to generate these forecasts.

Data Sources for Disaggregate Sample

- Census/American Community Survey (ACS) Public Use Microdata Sample (PUMS) data
 - Rolling 3-years (e.g., 2007-2009) or 5-years (2005-2009)
 - Geographically sampled/grouped by public use micro data areas (PUMAs) ~ 100,000 population -/+
- A representative regional household survey could be used for disaggregate sampling
 - Insufficient quantities and sampling biases could be an issue
 - This may be the only source for certain sub-groups
 - e.g., National Household Transportation Survey (NHTS) add-on for Florida includes seasonal households

In most cases, the primary source of disaggregate sample data will be PUMS data, which is now part of the ACS, and follows the same sampling framework, but provides disaggregate records for households and persons across numerous different attributes. PUMS is sampled and grouped according to geographic units, better known as PUMAs. PUMAs cover contiguous areas of roughly 100,000 population, including persons living in group quarters. For example, a metro area of 850,000 might be covered by 8 or more likely 9 PUMAs.

In some cases, it might be desirable to use a representative regional household survey for the disaggregate sample. This is generally not done, because of concerns over the insufficient quantities and potential sampling bias. It may make sense, however, for generating populations of certain sub-groups. For example, the NHTS add-on survey for Florida, included seasonal households, the characteristics of which could be used to generate a synthetic population of these part-year residents who do not show up in the in the Census of residents.

In general, ACS-PUMS provides good representative coverage of most regions and is rigorously tested and monitored, so it is generally preferred over household surveys for developing characteristics of the population.

For forecasting, there are particular challenges associated with areas that are likely to change their land use composition significantly. This may require special analysis to predict how such areas are likely to change in terms of the types of households expected to live there. Again, this is where a land use model may be helpful.

Specifying Control Attributes(1 of 3)

- Control attributes could be single or multi-dimensional
 - Multi-dimensional attributes can be treated as single dimensional attributes with number of categories equal to the product of the numbers of categories in individual attributes, minus infeasible cells.
 - Advantage: more precise regional control over the correlation between attributes
 - Disadvantages: sparse combinations
 - difficulty in finding samples to meet cell quota
 - may over-represent certain sample households

Size1 Workers0 Income1
 Size1 Workers0 Income2
 Size1 Workers0 Income3
 Size1 Workers0 Income4
 Size1 Workers1 Income1
 Size1 Workers1 Income2
 Size1 Workers1 Income3
 Size1 Workers1 Income4
 Size2 Workers0 Income1
 Size2 Workers0 Income2
 Size2 Workers0 Income3
 Size2 Workers0 Income4
 Size2 Workers1 Income1
 Size2 Workers1 Income2
 Size2 Workers1 Income3
 Size2 Workers1 Income4
 Size2 Workers2 Income1
 ... etc.

Size4 Workers2 Income1
 Size4 Workers2 Income2
 Size4 Workers2 Income3
 Size4 Workers2 Income4



Since the target distribution of households is so important, it is worth discussing some of the nuances of specifying control attributes.

Control attributes may be single or multi-dimensional. At right, we've depicted an example of a 3-way distribution of household size (4 levels), workers (3 levels), and income (4 levels). Multi-dimensional attributes can be treated as single dimensional attributes with number of categories equal to the product of the numbers of categories in individual attributes. So, in this example, which was developed for the SHRP2 C10A Jacksonville model, there would be 48 categories if all cells were feasible. However, certain combinations are definitely infeasible, because we cannot have more workers than household members, there only 44 feasible combinations.

The primary advantage is more precise regional control over the correlation between attributes. The disadvantages are that some of the combinations of attribute levels are rare and may be non-existent within certain sub-geographies. These sparse combinations may make difficult to find samples to meet the cell quota during the drawing process. In addition, if a certain cell

combination is represented by only one or two households in the sample, those households may be over-represented in the final population.

Specifying Control Attributes(2 of 3)

- It is desirable that all control attributes are somewhat “orthogonal” to each other
 - i.e., their variance in the population is largely independent
- Controlling for two attributes that are highly correlated is no better than controlling for just one
 - Example: In a region, if certain income categories are correlated with race, then it may not be efficient to include both income and race of head of household in the control attribute set

It is not possible to control many attribute levels well simultaneously. There are tradeoffs to be made that can lead to a more efficient design. In this sense, efficiency means explaining as much variation as possible with the fewest number of variables. The best choices of variables, will be meaningful attributes that are somewhat orthogonal to each other, which means that their variance in the population is largely independent.

Conversely, if you have two attributes that are highly correlated, then controlling for both may not achieve much more than controlling for just one. For example, if in a particular region certain income categories are highly correlated with the race of the household head, then it may not be efficient to control for both household income and race at the same time. In our bridge tolling and transit pricing example, this is something we’d have to look at more closely, particularly if we’re interested in equity analysis.

Specifying Control Attributes(3 of 3)

- Number of control attributes
 - Too few control attributes:
 - Synthetic population may not accurately reflect the true population
 - Too many control attributes leads to sparse cells in seed data:
 - May distort true distribution
- Control attributes may use different geographic units
 - If control attribute totals are not accurate at a particular spatial unit, they could be specified at a lower resolution
 - Best if spatial units “nest” (Census blocks & block-groups)
- Different sets of control attributes may be used for base and forecast years, limited by forecasting accuracy

The number of control attributes is also important. On one hand, if there are too few control attributes, the synthetic population may not accurately reflect the true population. In our bridge example, we'd like to control for as many income levels as possible, as well as number of workers, and race in order to be able to better address concerns of driver willingness to pay, estimating HOV and transit demand, and assessing equity impacts.

On the other hand, too many control attributes leads to sparse cells in seed data, which may distort the true joint distribution, or even make it impossible to compute using certain methods. And, as mentioned earlier, this sparseness may make it difficult to find suitable sample households in the disaggregate data, or perhaps just one or two sample households, which get replicated more than they should.

Control attributes also may use different geographic units. So, if control attribute totals are not accurate at a particular spatial unit, they could be specified at a lower resolution. Of course, it is best if spatial units “nest” (Census blocks & block-groups).

Finally, different sets of control attributes may be used for base and forecast years, if limited by forecasting accuracy. This is not necessarily desirable, though. The ability to forecast marginal control totals should be a consideration when specifying control attributes for this base year.

Uncontrolled Attributes

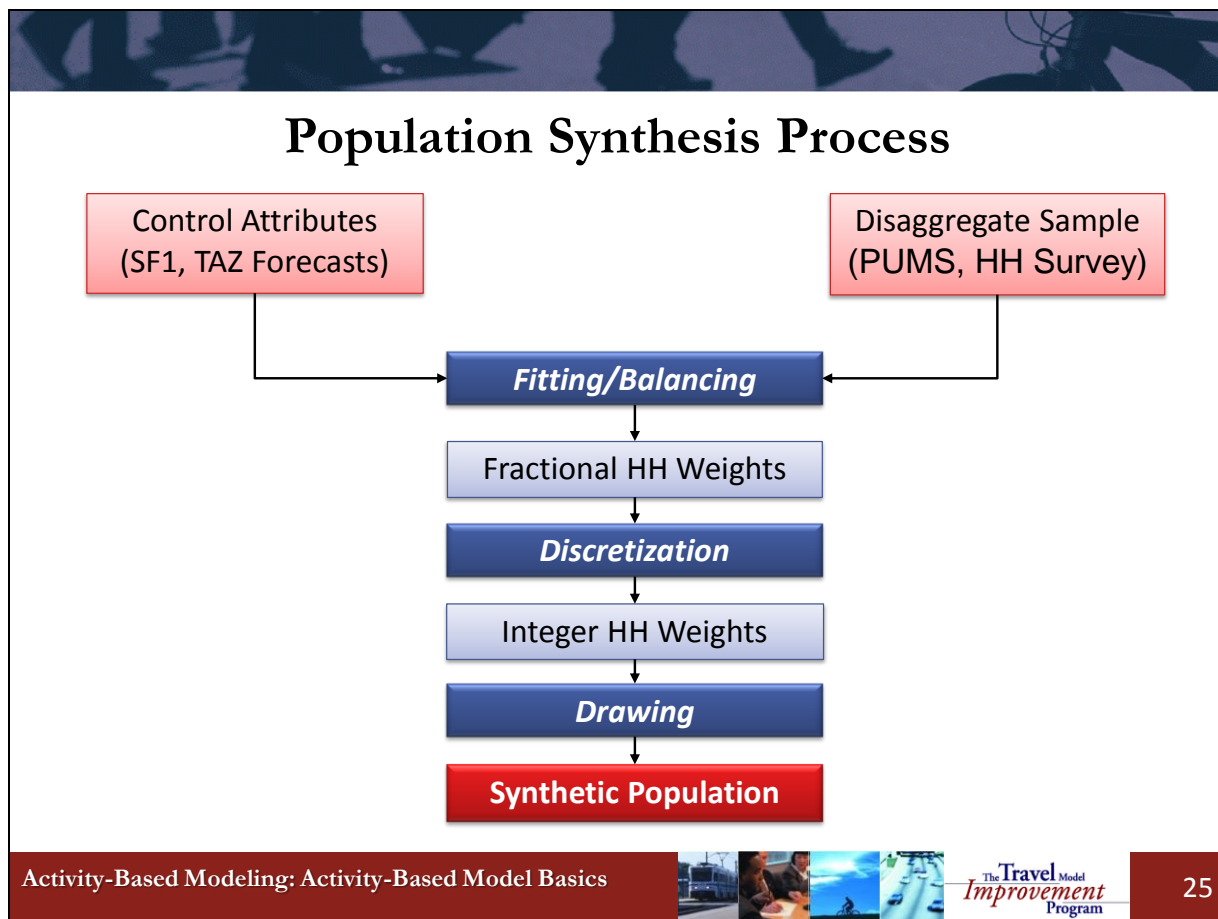
- Uncontrolled attributes are directly transferred to the synthetic population when the sample is drawn
 - All the attributes present in the sample may not be controlled for since it could affect the match negatively
 - Such variables are often needed by activity-based models in addition to the control variables
 - Should be well-correlated with controlled variables
- Examples
 - Person worker status, student status, race, occupation
- Be wary of potential bias due to loose correlation with control variables!

Although not every desired variable is available at the population level, it is possible to add them to the synthetic population by drawing them from samples, such as PUMS or a household survey. To the extent that uncontrolled attributes are correlated with controlled attributes, this should be a realistic estimate of their presence in the population. In addition, one may also want to select additional “uncontrolled” attributes to support analysis needs, such as work and student status for individuals, race and occupation.

It is important to identify the risk of biased results if uncontrolled attributes that are not correlated with controlled attributes are used as explanatory variables in the models, or used to aggregate model results. A real-world example of this is Atlanta Regional Commission, where race was not controlled, and a backcast revealed that it was not well synthesized. If race had been used in forecasting to aggregate model results, they would have been biased.

So for our bridge example, we might want to consider controlling for race, and not leaving it up to “the luck of the draw” so to speak. We’d also want to conduct a back-casting exercise to make

sure our projections for key sub-groups—low-income households, minority households, and potentially other groups—wind up in the right geographic locations.



This diagram summarizes the population synthesis process. The first step is balancing to create a population distribution, using the disaggregate sample and the control variables as inputs, as we just discussed.

Here, the algorithm fits the disaggregate sample of households to aggregate constraints (control totals) at prescribed level of spatial unit control. The most commonly used method for the fitting, or matrix balancing, is the Iterative Proportional Fitting (IPF) procedure, which we will illustrate with an example in a few minutes. There are other balancing procedures as well, including list balancing. The outcome of the balancing step is a set of fractional household weights that describe the multidimensional distribution of households in the region, a defined by these control variables.

The next step is to draw a sample to create individual household and person records to match the balanced distribution. Before we can do this, we need to first convert fractional values in the resulting multidimensional distributions to integers. This is easy enough to do with a bucket rounding procedure.

After this, we can select individual households from the sample to match the fitted, “integerized” distribution. Any specified uncontrolled attributes will also be included in the population attributes due to their presence in household and person sample records.

Iterative Proportional Fitting (IPF)

- Widely used and statistically robust procedure to control the joint distribution of attributes in a synthetic population. (See Beckman et al, 1996)

Basic Steps:

1. Choose a set of control attributes and their levels
2. Obtain control totals for the control attributes by level
3. Obtain a sample of households with the relevant attributes to be the “seed matrix”
4. Create a joint distribution of control attributes from the disaggregate sample data
5. Iteratively factor the cells in the matrix, based on marginal control attribute target values for each attribute level, until all target values are matched or nearly matched



The iterative proportional fitting (IPF) method, or some variant of it, lies at the heart of most population synthesizers. It is used to produce an estimate of the joint distribution of attributes at the specified levels and spatial units. It is also a bit easier to visualize than the list balancing method and employs the same principles of adjusting the sample to meet target distributions, so we will use it as an example to illustrate this process.

The basic steps of the IPF procedure are as follows:

- Choose a set of control attributes and their levels
- Obtain control totals for the control attributes by level
- Obtain a sample of households with the relevant attributes to be the “seed matrix”
- Create a joint distribution of control attributes from the disaggregate sample data
- Iteratively factor the cells in the matrix, based on marginal control attribute target values for each attribute level, until all target values are matched or nearly matched

IPF Illustration – Setup					
		Household Income			
		High	Low	Total	Household Size Control
Household Size	Adjustment	-	-		
1	-	2	3	5	20
2	-	4	1	5	50
3 or more	-	1	3	4	30
Total		7	7		
Household Income Control		40	60		

Here is an example of an IPF procedure that uses a matrix approach, with only two control variables—household size (3 levels) and household income (2 levels). It is an unrealistically simple example.

The disaggregate sample is tabulated and summed to fill in the number of observations that correspond to each cell within the 3-by-2 matrix.

The marginal controls totals are represented row and column target values for household size and income, respectively. As you can see, the control totals for each dimension are significantly larger than the initial sums from the sample data.

IPF works when the joint attributes are defined as cells, the universe of the joint distribution being the entire set of cells, and each control total is associated with some subset of the cells. An iteration consists of adjusting for all controls, one-by-one in sequence. In the extreme case, each household in a representative sample could serve as a cell, which is essentially how PopGen works.

IPF Illustration – Row Adjustments					
		Household Income			
		High	Low	Total	Household Size Control
Household Size	Adjustment	-	-		
1	$20/5 = 4$	2	3	5	20
2	$50/5=10$	4	1	5	50
3 or more	$30/4=7.5$	1	3	4	30
Total		7	7		
Household Income Control		40	60		

To begin the procedure, we use the ratio of the control total divided by the summed total for each row to create row factors.

IPF Illustration – Iteration 1

		Household Income			
		High	Low	Total	Household Size Control
Household Size	Adjustment				
1	4	$2 \times 4 = 8$	12	20	20
2	10	$4 \times 10 = 40$	10	50	50
3 or more	7.5	$1 \times 7.5 = 7.5$	22.5	30	30
Total		7	7		
Household Income Control		40	60		



Applying these row factors, we get new cell values and new row sums that exactly match row control totals for household size.

IPF Illustration – Column Adjustments

		Household Income			
		High	Low	Total	Household Size Control
Household Size	Adjustment	$40/55.5=0.72$	$60/44.5=1.35$		
1		8	12	20	20
2		40	10	50	50
3 or more		7.5	22.5	30	30
Total		55.5	44.5		
Household Income Control		40	60		

The row adjustments also changed the column totals for the two income levels, but they still do not match their control totals. So, we take the ratio of control total to column sum to create adjustment factors.

IPF Illustration – Iteration 2

		Household Income			
		High	Low	Total	Household Size Control
Household Size	Adjustment	0.72	1.35		
1		$8 \times 0.72 = 5.8$	$12 \times 1.35 = 16.2$	22	20
2		28.8	13.5	42.3	50
3 or more		5.4	30.3	35.7	30
Total		40	60		
Household Income Control		40	60		



Applying these column factors to the cells in the matrix, we obtain column sums that match income target values exactly. However, we can see that our row sums are now a bit off from the household size target values for each row.

IPF Illustration – Iteration 13 (Final)

		Household Income			
		High	Low	Total	Household Size Control
Household Size	Adjustment				
1		4.51	15.49	20	20
2		31.70	18.30	50	50
3 or more		3.79	26.21	30	30
Total		40	60		
Household Income Control		40	60		

This iterative row and column factoring continues until, at Iteration 13, we have a final set of values in the 3-by-2 table that produce row and column sums that match BOTH household income and size control totals, respectively. Although this example showed a 2-dimensional table, this same process may be extended to 3 dimensions and greater. It is guaranteed to converge, provided that there are not too many cells in the table with values of “0”.

Drawing Households from a Sample

1. Calculate selection probabilities for each household in the sample
 - Based on the attributes of the household and the number of such households in the joint distribution
2. Draw households from a sample
 - Based on the selection probabilities to match target numbers by each household type in the joint distribution
 - e.g., Monte-Carlo process, cycling methods
3. Output a synthetic population consisting of all persons belonging to the households drawn

Once we've got the multi-dimensional target distribution figured out, the next step is to draw households from our sample to match that distribution. The general steps in this process begin with calculating selection probabilities for each household in the sample, based on the attributes of the household and the number of such households in the joint distribution.

Next, we draw households from a sample, based on the selection probabilities to match target numbers by each household type in the joint distribution. The most common method for doing this is a simple random drawing based on the probabilities weights, using Monte Carlo methods. (For those of you unfamiliar with Monte Carlo selection, we discussed this in Webinar 4.) Other, potentially more efficient methods involve cycling methods in which we make sure we sample from all of the available households that meet the joint distribution criteria, taking each one in turn, rather than leaving it to chance.

Once we have drawn households to fill meet our target population distribution, we output a synthetic population consisting of all persons belonging to the households drawn.

Methods for Estimating a Joint Distribution

- Matrix balancing
 - Follows the standard IPF approach
 - Involves creation of multidimensional matrices during fitting
 - Controls either household or person attributes, but not both
- List balancing
 - Involves applying IPF on individual households in a list
 - More complex algorithm
 - Both household and person attributes can be simultaneously controlled
 - Reduces sparse matrix problems

There are two alternative approaches in developing a joint distribution of households—matrix balancing and list balancing. Matrix balancing follows the standard IPF approach in fitting a multi-dimensional matrix. It can control either households or person attributes, but not both.

List balancing involves applying IPF to individual households in a list. List balancing is a more complicated to program, but has two advantages of over matrix balancing. First, both household and person attributes can be simultaneously controlled. Second, it is easier to eliminate matrix cells that are irrelevant or combine those that have extremely low incidence rates.

Other Ways to Sample from a Distribution

- **Intelligent drawing**
 - Directly draw households from the sample to match control totals, but without a separate fitting/balancing step (no IPF)
 - Drawing informed by the extent to which a sample household would contribute to the match of both household and person controlled attributes
 - Examples: ARC (Atlanta), FSUTMS (Florida)

Two emerging methods in drawing a sample are intelligent draws and discretizing. “Intelligent” drawing eliminates the need to develop a target multi-way distribution through a separate fitting and balancing step, so no IPF process. Instead, households are drawn directly from the disaggregate sample to match control totals. Drawing is informed by the extent to which a sample household would contribute to the match of both household and person controlled attributes. The challenge is in developing selection criteria and algorithm, which is why this is considered to be an emerging method, although there have been at least two implementations of this: ARC population synthesizer and FSUTMS in Florida.

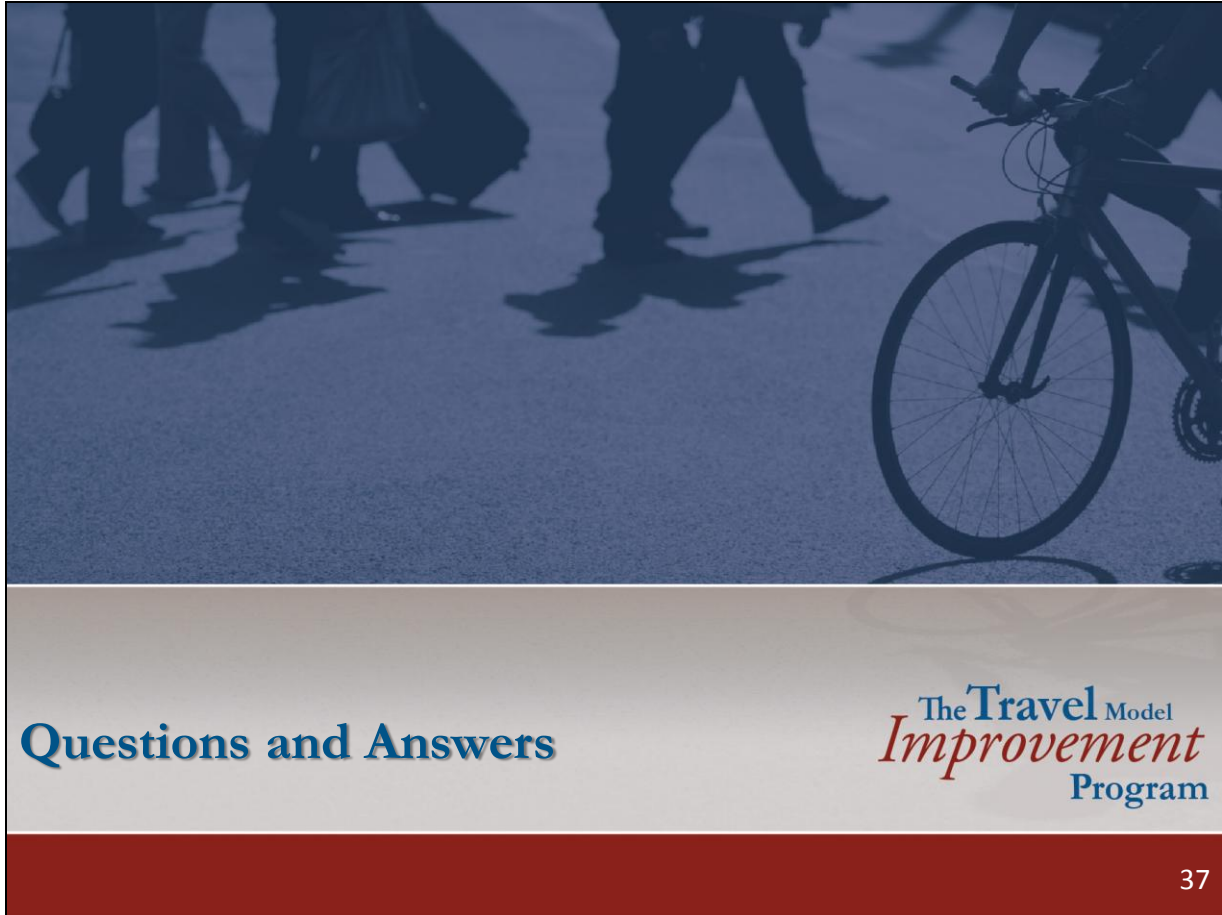
Other Ways to Sample from a Distribution

- **Discretizing**

- Developing individual household weights which are then “discretized” so that the distributions with respect to controlled attributes are preserved
 - (e.g. “intelligent bucket rounding”)
- Individual weights can be developed by IPF or entropy maximization techniques
- Similar to list balancing, but does not involve a drawing process
 - Example: SANDAG



The other method is referred to as “discretizing,” which may be thought of as “intelligent bucket rounding.” In this method, households are assigned weights, which may be developed through IPF or an entropy maximization method. The weights should be proportional to their probability of selection, if Monte Carlo draws were used. However, Monte Carlo draws are not used. Rather the fractional weights are “integerized” using methods similar to bucket rounding, which seeks to preserve totals. Each integer instance of a household can then be used directly in the synthetic population. This method is similar to list balancing without the drawing step. One example of this has been implemented in San Diego. This approach has the advantage of ensuring full use of the sample space when selecting households, which is not guaranteed in Monte Carlo sampling.



Implementations of Population Synthesizers

Type of Synthesizer	Fitting and Drawing	Fitting only	Drawing only
Matrix & table balancing	<ul style="list-style-type: none"> • PopSynWin (U. Illinois-Chicago) • CEMSELTS (U. Texas-Austin) • TRANSIMS (FHWA) • ALBATROSS (TU-Eindhoven) • MORPC (PB) 		<ul style="list-style-type: none"> • ARC PopSyn (PB)
List balancing	<ul style="list-style-type: none"> • PopGen (Arizona St.) • ILUTE (U. Toronto) 	<ul style="list-style-type: none"> • SANDAG PopSyn (PB) 	<ul style="list-style-type: none"> • FSUTMS (U. Florida)




This table is a list of known population synthesizers, or the modeling packages of which they are a part. They are organized by those that use matrix balancing, which is the most common, and those that use list balancing. They are also further organized by whether they utilize both fitting and drawing procedures, which seems to be the most common, or whether they only do fitting or only do balancing. Note that most of these were developed at universities, with two exceptions developed by consultants in Atlanta and San Diego. Other activity-based models now in practice have used some version of these population synthesizers. Let's look at a few examples.


Balancing Procedures

Generalization

Type	Controls	A priori weights (seed distribution)	Contribution coefficients
Multidimensional Matrix (MORPC)	Row/column totals	Initial matrix	Cell-row/column incidence (0,1)
Table of categories (ARC)	Column totals	Initial weight for category (row)	Row/column incidence (0,1)
List of individual records (SANDAG)	Column totals	Initial individual weight (row)	Row/column coefficient (≥ 0)

- Each subsequent method includes the previous one as a particular case and guarantees the same result
- Not every table of categories can be reduced to a matrix form!
- Not every table of individual records can be reduced to table of categories!

Activity-Based Modeling: Activity-Based Model Basics



39

It is important to understand the technical differences between 3 types of balancing procedures that represent the core of any population synthesizer. Each more advanced procedure is a generalization of a simpler procedure. Thus, we are not talking about 3 different methods but rather about 3 levels of complexity & flexibility. Each method has controls set in a specific way, the seed sample distribution set in a specific way, and a matrix of contribution coefficients that link between the controls and seed distribution.

- Each subsequent method includes the previous one as a particular case and guarantees the same result
- Not every table of categories can be reduced to a matrix form!
- Not every table of individual records can be reduced to table of categories!

Example of List of Individuals for Balancing

HH ID	HH size				Person age				HH initial weight
	1	2	3	4+	0-15	16-35	36-64	65+	
	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	
$n = 1$	1							1	20
$n = 2$		1			1	1			20
$n = 3$			1			1	2		20
$n = 4$				1		2	2		20
$n = 5$				1	1	3	2		20
....									...
Control	100	200	250	300	400	400	650	250	



The list balancing procedure can be illustrated in the following way. The rows in the list are individual HHs from the sample. Each HHs has an initial a priori weight that is normally equal to all HHs but can be differentiated. The columns of the list table correspond to controls. HH-level controls correspond to the total number of HH of certain type while person-level controls correspond to total number of persons of certain type that we need to synthesize in the given TAZ. The matrix contains contribution factors for each household to each control. Contribution factors are 0,1 for household-level controls but it can be any number 0,1,2,3,4... for person-level controls. The balancing procedures is applied to calculate HH weights that would match all controls.

Program Formulation (Fixed Controls)

$$\min_{\{x_n\}} \sum_n x_n \ln \frac{x_n}{w_n} \quad - \text{Preserve initial weights as much as possible}$$

$$\sum_n a_n^i x_n = A^i, (\alpha^i) \quad - \text{Meet all controls}$$

$$x_n \geq 0$$

- Convex mathematical program with linear constraints
- Solution can be found by forming the Lagrangian and equating partial derivatives to zero (necessary conditions)
- Conventional matrix balancing or table balancing are particular cases



This problem has a closed form formulation as the maximum entropy problem. We have developed very efficient methods to find the unique solution. The essence of this formulation (and that is what a good synthetic population is) is to calculate weights that would meet all controls but also preserve the initial weights as much as possible (i.e. would use all HHs in the sample in the most uniform way). Conventional matrix balancing (many of you are familiar) or table balancing are particular cases of this general method.

Advantages of Balancing List of HHs

- No reason to fight zero cells in joint distributions, they cannot be utilized anyway
- Can incorporate HH-level and person-level controls naturally
- Prepares background for discretizing; no need in drawing

There are many advantages of the List Balancing procedure vs. matrix balancing applied in many earlier population synthesizers. No reason to fight zero cells in joint distributions, they cannot be utilized anyway. Can incorporate any HH-level and person-level controls naturally. Prepares background for discretizing; no need in drawing HHs from the sample.

Relaxation of Controls (MAG, BMC)

- Objective function:
$$\min \sum_n x_n \ln \frac{x_n}{w_n} + \sum_i \alpha_i Y_i \ln Y_i$$
- Match relaxed controls:
$$\sum_n a_{ni} x_n = A_i Y_i (\lambda_i)$$
- HH weights and relaxation factors: $x_n \geq 0, \quad Y_i \geq 0$
- Importance factors for controls: $\alpha_i \geq 0$
 - Set to a large value of 1,000 to ensure match if feasible
 - Set to 1,000,000 for total number of HHs

Several useful extensions of this method have been recently introduced (MAG, BMC). One of them allows for differential relaxation of controls. It solves a very frequent problem of non-convergence if the controls are not fully consistent between themselves (which is a usual case in practice for at least some TAZs in the region). In this formulation the balancing procedure would find the best compromise solution. It can also naturally incorporate differential importance of controls. Controls with high weight will be fully satisfied. Less important controls will be satisfied to the extent possible.

Importance of Entropy-Maximizing Balancing vs. Simple IPF

Household type	Initial distribution	Linear IPF	Entropy-max balancing
0 workers	20%	4%	12%
1 worker	30%	36%	25%
2 workers	40%	48%	47%
3 workers	10%	12%	16%
Control average	1.4	1.8	1.8

- That's how entropy-maximization works!



This is not just a theoretical achievement. Balancing methods have a very important practical impact on the results especially if structural changes in the population are expected. Consider a zone with the initial seed (base year) distribution of HHs by number of workers as shown in the second column. It corresponds to the average number of workers per HH of 1.4. Now, let's say we need to restructure the distribution to meet a constraint of 1.8 labor force participation. A naïve IPF would result in an unrealistic distorted distribution shown in the third column. The entropy-maximizing method yields a much more reasonable structural shift shown in column 4.

Implementation Example 1: Baltimore Metropolitan Council (BMC)

- Implemented using PopGen (ASU) at the TAZ level
- Only household attributes controlled
 - Household size (5 categories)
 - Income (4 categories)
 - Workers (4 categories)
- Joint totals of size-income and income-workers from the synthetic population used in trip generation models (4-step)

This first example is a somewhat simple specification of the PopGen population synthesizer, which has seen wide use in a number of locations. PopGen was developed at Arizona State. In this implementation at the Baltimore Metropolitan Council (BMC), only three household attributes are controlled—five levels of household size, four income levels, and four worker categories. The spatial unit of control is the TAZ.

PopGen also has the ability to control person-level attributes, although this was not done here. This is primarily because the synthetic population is being used to produce a joint distribution of household size, income and workers for use in the trip generation stage of a 4-step modeling. This is a preliminary step in BMC's long-term plans to develop an activity-based modeling system.

Implementation Example 1: Baltimore Metropolitan Council (BMC)

Variable	Control	Synthetic	% Difference
Household Size			
1 person	527,210	527,266	0.01%
2 persons	561,788	562,293	0.09%
3 persons	333,499	333,607	0.03%
4 persons	261,710	261,534	-0.07%
5 or more persons	207,021	206,528	-0.24%
Total	1,891,228	1,891,228	0.00%
Household Income			
< \$11,800	190,133	189,005	-0.59%
≥ \$11,800 and < \$26,000	284,592	283,968	-0.22%
≥ \$26,000 and < \$44,200	378,935	378,845	-0.02%
≥ \$44,200	1,037,528	1,039,410	0.18%
Total	1,891,188	1,891,228	0.00%

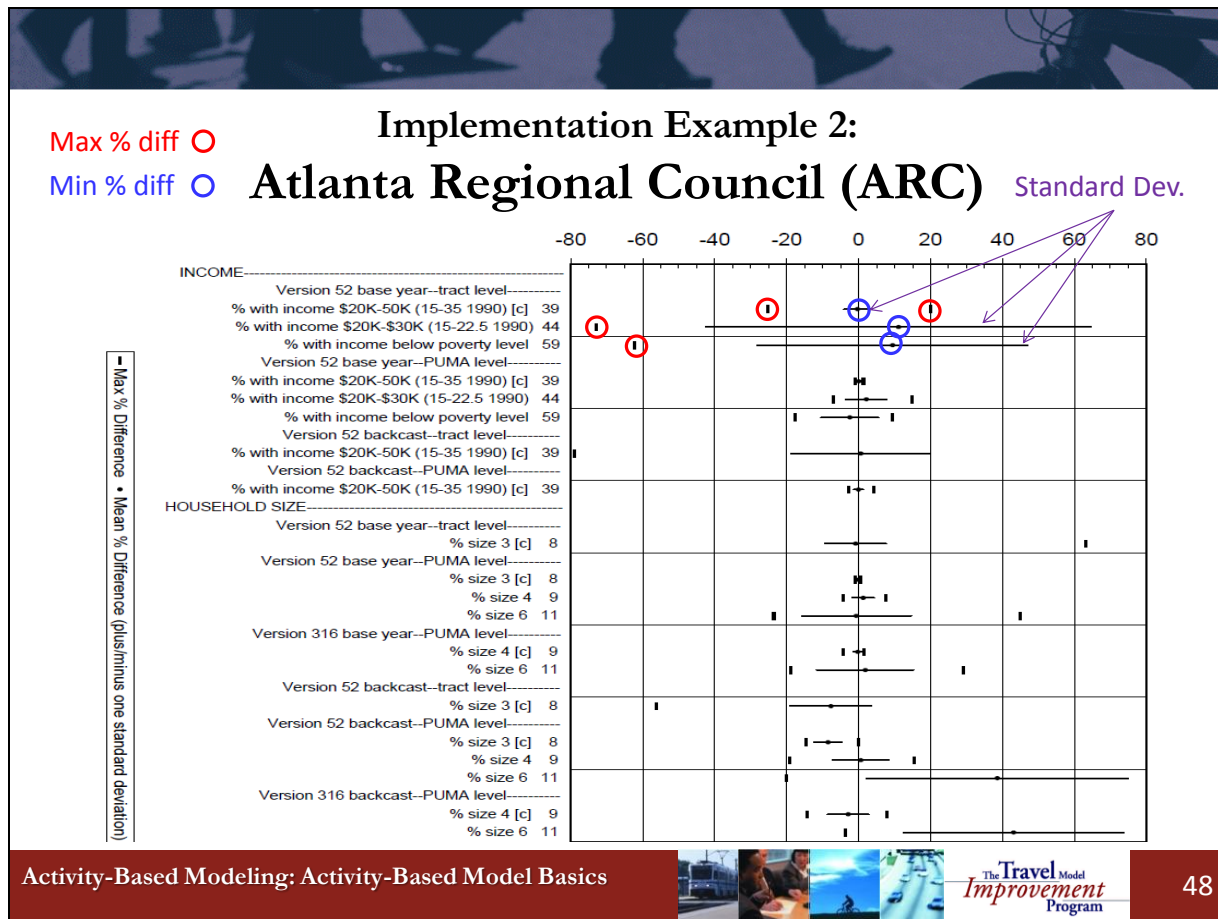


This table shows the fit of the synthetic households to the control data, which as expected is quite good. One thing to notice here is that the model is specified to be sensitive to lower-income ranges, hence the first three income groupings are all below the median household income for the region, with one category representing more than half of the households in the region. You may also notice that categories which have the smallest total numbers (lowest income group, five-plus person households) are also the most difficult to fit.

Implementation Example 2: Atlanta Regional Council (ARC)

- Implemented using PopSyn at the TAZ level
- Household attributes controlled
 - Household size (5 categories)
 - Income (4 categories)
 - Workers (4 categories)
 - Child presence (2 categories)
 - Age of head of household (2 categories)
 - Family/Non-family (2 categories)

In contrast to the Baltimore example, this example from Atlanta is a little more complex. It uses six household control variables, having added the presence of children, two age of head categories, and a family/non-family household indicator. There are no person control variables.



Here are selected validation results from the ARC model. Although the model variables shown are controlled at the TAZ level, validation statistics are shown at the Census tract level for household income and at the PUMA level for household size. The comparison here was with a backcasting exercise to see how well the model projected to a different year than the baseline. This is a very useful approach to model validation.

The plots show maximum differences between synthesized and actual populations by the vertical dashes, examples of which are circled in red for the first three categories.

Dots symbolize the minimum difference, and these are circled in blue for the first three categories.

The longer horizontal lines, pointed out here in purple for the first three categories represent one standard deviation about the mean. As you can see, there are some quite large deviations.

Implementation Example 2: Atlanta Regional Council (ARC)

- Lessons learned from “backcasting”
 - Bias arising from uncontrolled variables
 - Race was uncontrolled and backcasts revealed gross inaccuracies in projections
 - Bias arising from more aggregate controls for forecast year
 - The seed preserves the distribution from the base year and the distribution of the uncontrolled variable changes in the future year within the more aggregate category
 - Practical implications:
 - If uncontrolled variables are not directly used in the model these biases are tolerable
 - Uncontrolled variable with a strong impact on travel demand should be better re-specified and additional controls introduced

With this ARC example, there are a couple important aspects. First, there is the bias that occurs with uncontrolled variables. In this model, race was uncontrolled and backcasts revealed gross inaccuracies in projecting the distribution of household by race.

Second, this example illustrates the bias that can occur in forecasts when the controls in the forecast year are much more aggregate than in the base year. What happens is that the seed preserves the distribution from the base year and the distribution of the uncontrolled variable changes in the future year within the more aggregate category.

Other ARC results point out the problems of bias with uncontrolled categories. From the practical perspective it is important to keep in mind how the problematic uncontrolled variable is going to be used in the travel model. In general, important variables that have a strong impact on travel demand have to be controlled in the population synthesis.

Implementation Example 3: **SHRP 2 C10A – Jacksonville, Florida**

- Implemented using PopGen at the TAZ level
- Includes seasonal households explicitly:
 - Have very different HH structure & travel behavior
 - Relevant for certain seasons only
- Household attributes controlled
 - Household size-income-workers (44 categories)
 - Child presence (2 categories)
 - Age of head of household (3 categories)
- Person attributes controlled
 - Age (5 categories)
 - Gender (2 categories)


The third example we have is from the SHRP2 C10A project in Jacksonville, Florida. This is the same example we showed before in which we used a 3-dimensional joint distribution of household size, income and workers as a control variable. This created 44 categories for this single control variable. Also, in this model we have added explicit person control attributes for age and gender. In addition, this project also modeled seasonal households, rather simply, by their size and age, with control data obtained from the NHTS add-on survey for Florida. Adding seasonal households is not something done in most regions, however, Florida cities these households represent an important demand segment and appear in trip-based models. Seasonal households typically do not work in the region, but exhibit travel behavior similar to other non-working households.

Implementation Example 3: SHRP 2 C10A – Jacksonville, Florida

Variable	Control	Synthetic	% Difference
Household Size			
1 person	118,841	119,901	0.89%
2 persons	161,113	161,595	0.30%
3 persons	84,235	83,394	-1.00%
4 or more persons	115,067	114,408	-0.57%
Total	479,255	479,298	0.01%
Person Age			
0-15 years	286,068	283,248	-0.99%
16-20 years	78,668	77,511	-1.47%
21-44 years	443,351	435,734	-1.72%
45-64 years	270,899	266,070	-1.78%
65+ years	123,868	122,237	-1.32%
Total	1,202,855	1,184,800	-1.50%



Here are the goodness of fit statistics for this model. This shows that even with moderately complex household and person specifications, good fit can be achieved.





Implementation Example 4:

San Diego Assoc of Governments (SANDAG)

- Implemented using PopSyn II at the TAZ level
- Household attributes controlled
 - Household size (4 categories)
 - Income (5 categories)
 - Workers (4 categories)
 - Child presence (2 categories)
 - Dwelling unit type (3 categories)
 - Group quarter status (4 categories)
- Person attributes controlled
 - Age (7 categories)
 - Gender (2 categories)
 - Race (8 categories)

Activity-Based Modeling: Activity-Based Model Basics



The Travel Model
Improvement
Program

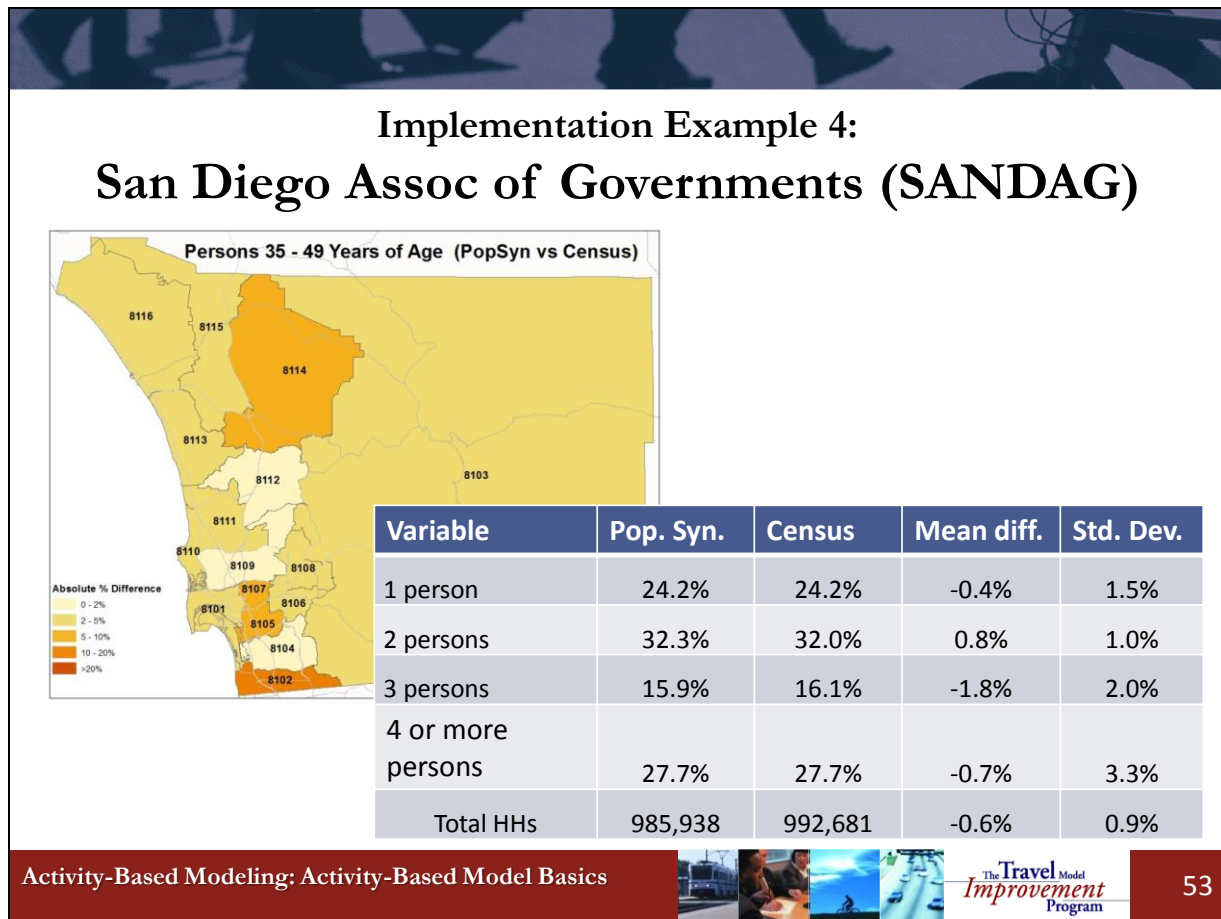
52

This example from San Diego illustrates an even more complex model specification in which we control for both household and person types and have added dwelling types. This is the most complex control structure for person attributes of the examples we have seen. By controlling for race, this model is likely to avoid some of the statistical bias issues that the ARC model revealed when race was included as an uncontrolled attribute. At the same time, this poses challenges for forecasting race for future year populations.

Introduction of person-level controls is essential since certain demographic tendencies like population aging are better described in terms of person distributions than in terms of household distributions. This also creates a better linkage between the travel model and land-use and demographic models.

Also note that there is a separate group quarters status, consisting of four categories. Modeling group quarters populations is typically done separately from the non-group quarters population.

Group quarters residents are treated as if they belong to a one-person household, so their person attributes, namely age and gender are typically the only attributes.



This slide shows another table of goodness of fit statistics for the household size distribution, which look good. To be more precise, however, it is useful to see how closely the synthetic population matches the Census (or other control source) at a finer geographic level. This map shows Census Tracts in the region and how the distribution of synthetic persons ages 35 to 49 compare to the Census. A geographic analysis of this kind is recommended for validating a population synthesizer.

Why Good Fit is Important

- The outputs of a synthetic population are the inputs to all other model components.
 - If these inputs are inaccurate, forecasts and other analyses will also be inaccurate.
- Types of Errors:
 - Under-representation of market sub-segments → model may be insensitive
 - Over-representation of market sub-segments → model may be too sensitive, or sensitive in unexpected ways
 - Mis-alignment of population with geography → inaccurate forecasts (trip lengths, mode shares, etc.)

The outputs of a synthetic population are the inputs to all other model components. If these inputs are inaccurate, forecasts and other analyses will also be inaccurate. We can generally represent these as one of three types of errors:

If we are not specific enough in specifying the distribution of households and persons in the model, then we are likely to under-represent certain market sub-segments, and the model may be insensitive to key policy levers.

If we try to slice the distribution too finely, we may distort the distribution. This can lead to over-representation of market sub-segments, and the model may be too sensitive, or sensitive in unexpected ways.

If we are inaccurate in where we place households, that can also lead to inaccurate forecasts in travel behavior, just as they would with a trip-based model.

Getting good fit... practical tips!

- Choosing appropriate control attributes
 - Controlling for certain attributes may distort the distributions of others
 - Several iterations of testing may be required before determining the final set of control attributes
 - Meta-analysis of consistency between the controls
 - Differential importance weights can be applied to controls
- Zero cells in the seed matrix created from sample data
 - Redefine or combine attribute categories (collapse attribute levels to fewer groupings)
 - Expand the geographic unit to a more aggregate level

Although the methods that have been developed are relatively robust, one cannot specify a population generator naively. There are a few challenges to goodness of fit that one might need to overcome.

First, there is the proper specification of the control attributes. Controlling for certain attributes may distort the distributions of others. Several iterations of testing may be required before determining the final set of control attributes.

Second, it is important to account for sparse data across both attribute and spatial dimensions. Too many zero-valued cells in the seed matrix created from sample data could render the balancing step infeasible. Solutions include: redefine or combine attribute categories (collapse attribute levels to fewer groupings), and to expand the geographic unit to a more aggregate level.

Getting good fit... more practical tips

- It is usually possible to maintain tight control on either household control targets or person control targets, but not both simultaneously:
 - Importance weights higher for main HH controls
 - Importance weights lower for secondary HH controls and person controls
- It is possible to indirectly control person attributes at household level
 - In practice, the only person characteristics that are usually controlled for are age and gender.
 - Gender should work out without explicit controls.
 - Age can be handled at the household level by using age of head and presence/absence of people in certain age groups (i.e., very young children, school age children, etc.)

In practice, it is usually possible to maintain tight control on either household control targets or person control targets, but not both simultaneously. Certain controls can be relaxed depending on the importance (for modeling) and reliability of the data sources.

However, it is also possible to indirectly control person attributes at household level. In practice, the only person characteristics that are usually controlled for are age and gender. Gender should work out without explicit controls. Age can be handled at the household level by using age of head and presence and absence of people in certain age groups (i.e., very young children, school age children, etc.)

Forecasting Future Synthetic Populations

- Where do we get marginal control totals and sample populations for future years?
 - Control Variables:
 - Less controls are normally set for future years
 - Trend extrapolation based on horizon-year forecasts for entire region (or sub-region if available)
 - Land Use and Demographic model outputs
 - Sample population:
 - Re-use... assume similar joint distributions of household and person attributes in the horizon year... appropriate in built-out areas
 - Consider changing demographics... enrich the sample by adding HHs from other geographic units that look more like your region in the future (primarily, ethnicity or income mix)
 - Household evolution models (emerging method)

One of the big questions that comes up when discussion population synthesis is “where do we get marginal control totals and sample populations for future years?”

Control variables can be forecast using trend extrapolation methods, such as growth factoring, or the outputs of a land use model, if available, may be used. Many regions will also have macro-level regional socio-economic forecasts for future years, which can be use to grow the population at the regional level.

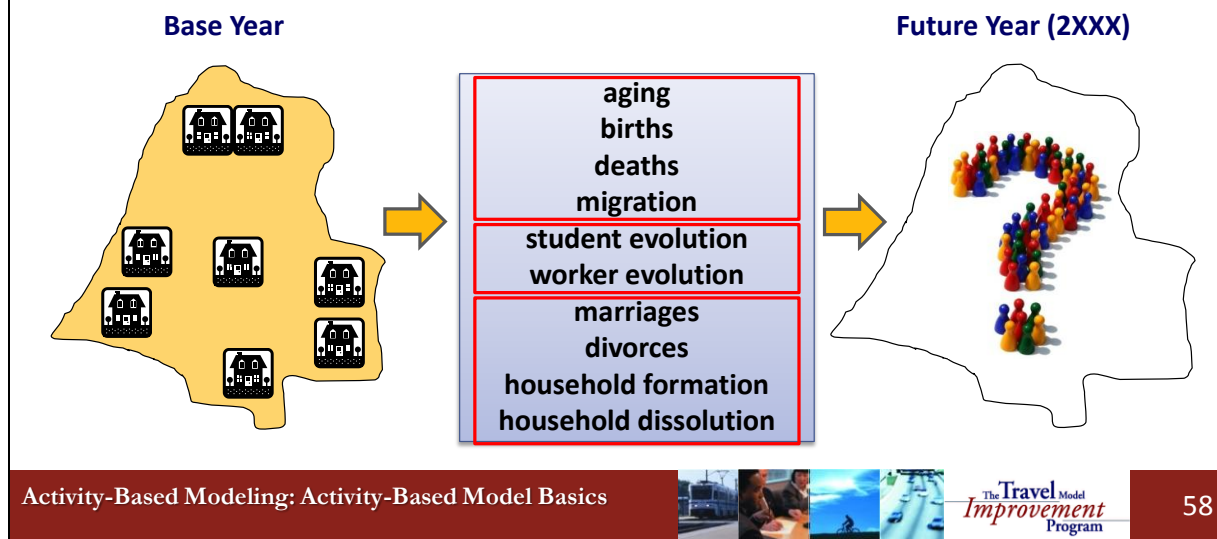
The sample population, from which we draw household characteristics may be derived by either assuming that the same types of households and correlation between control variables that exists in the base year will also be present in the future year. For some regions, or sub-regions, this may hold. For other regions, however, especially fast-growing metro area, this will most certainly not be true.

One way to reflect changing demographics in the household sample is to borrow samples from another area—for example another PUMA within the same region—that “looks” more like the way this subject area is expected to look in the future.

A third, emerging research area is “household evolution” modeling, which has the potential to replace the more mechanical methods of population synthesis by modeling processes that reflect actual human life courses. We will spend the remainder of this seminar exploring this topic.

Household Evolution

- Synthetic population for a base year is evolved to a future year by simulating certain demographic evolutionary processes



Household evolution models have been under development at least since the early 1990s, but have yet to be developed to the point where they can be used in practice. That said, they have a lot theoretical appeal and are worth considering further.

In household evolution, a synthetic population for a base year is evolved to a future year by simulating certain demographic evolutionary processes. These processes include: aging, births, deaths, marriages, divorces, household formation, household dissolution and migration. There are different methods applied for each one.

Of these, aging, births, deaths and migration are typically handled through well-accepted demographic practices and are perhaps of the most immediate relevance.

The second group of items—student and worker evolution—may also be forecast, and there are numerous good precedents available in the economics literature.

The third group of four items—marriages, divorces, household formation and dissolution—are a bit more controversial and difficult to forecast. They may be important, however, for determining the compositions of future households and their size.

Why Household Evolution?

- Demographic shifts, such as the aging of the population, are not always reflected in non-evolution forecasting processes
- Create new HHs by logical combination of features instead of locking in a relatively small sample
- Most models fail to account for in- and out-migration
- Socio-economic forecasts developed more methodically than applying naive growth rates to marginal distributions
- Potential to provide more variables for dynamic travel models – variables of change and lagged variables
- Consistent when integrated with land use models, which are also usually evolution models

Household evolution is a more “intelligent” way of forecasting the future distribution of characteristics of the population.

- The effects of demographic shifts, such as the aging of the population, are not always reflected in non-evolution forecasting processes.
- Generate new HHs by a logical combination of features instead of re-using the same small sample over and over
- Most travel models and fail to account for in- and out-migration.
- Socio-economic forecasts can be developed more methodically than applying naive growth rates to marginal distributions.
- Potential to provide more variables for dynamic travel models – variables of change and lagged variables.
- Household evolution is consistent when integrated with land use models, which are also usually evolution models.

Cohort Survival Method

- Standard demographic projection method
 - **Aging, fertility, mortality and net migration** by age cohort

		population at the beginning of the time period						population at the end of the time period	
cohort	Age group	Pop t_0	Survival Rate	Survive to $t_0 + 10$	Birth Rate	Births	Net Migrati	Pop $t_0 + 10$	
1	0 - 9	3,900	0.989		0	0	5	425	
2	10 - 19	3,200	0.999	3857	0.011	35	0	3,857	
3	20 - 29	3,300	0.998	3197	0.081	267	50	3,247	
4	30 - 39	2,800	0.998	3293	0.038	106	35	3,328	
5	40 - 49	1,700	0.996	2794	0.007	12	10	2,804	
6	50 - 59	1,800	0.991	1693	0	0	0	1,693	
7	60 - 69	1,100	0.975	1784	0	0	-20	1,764	
8	70 - 79	550	0.936	1073	0	0	0	1,073	
9	80+	200	0.88	691	0	0	0	691	
TOTAL		18,550		18,382		420	80	18,882	

to calculate births:

births by parent age cohort group:

10-19: $3200 \times .011 = 35$
 20-29: $3300 \times .081 = 267$
 30-39: $2800 \times .038 = 106$
 40-49: $1700 \times .007 = 12$
 the sum of these is 420

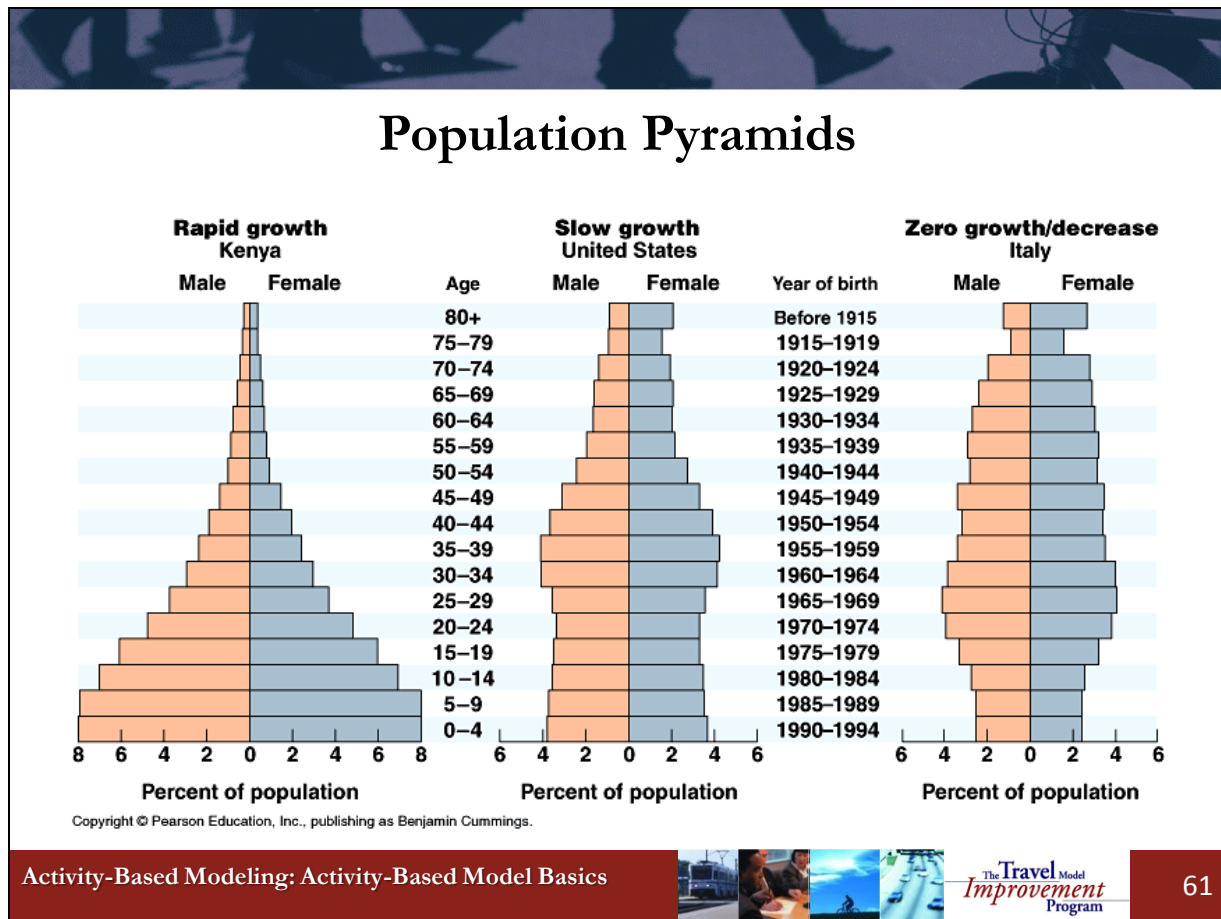
the population at the end of the time period then becomes the population at the beginning of the next time period

Source:

<http://www-personal.umich.edu/~sdcamp/up504/cohort%20survival%20examplew08.pdf>



The cohort survival method is a well-accepted method for aging a population based on fertility and mortality rates for persons within certain age cohorts. In this example, 10-year intervals are used, which is standard practice. Survival and birth rates are published by various state, federal and international organizations. The U.S. Census also maintains birth rates, death rates in migration rates for each state.



This slide shows graphically how differences in fertility, mortality, and net migration affect the distribution of the population by age groups. These types of graphs are called population pyramids. As you can see, there are huge differences between the three nations depicted here (in 1995).

In fast-growing Kenya a vast majority of people are children. If we were to age this population, those children will become adults and move up the population pyramid. If mortality and fertility rates maintain the same levels as the 1990s, it is likely that the shape of the pyramid, which is based on percentages, will not change.

If however fertility rates were to decline and/or mortality rates were to decline, then we might see the pyramid taking on a shape more like that of the U.S. (center). Here we can see the “baby boom” generations born between 1946 and 1964 about midway up the center of the graph. Their children, the so-called “boomlet” are represented at the bottom of the graph. A third example (right) is that of a zero-growth/declining population in Italy.

This is an example of a rapidly aging population where there are both much lower fertility and mortality rates, compared with past generations. So, in Italy, the bulk of the population is moving up the pyramid. This type of phenomenon may actually be happening in certain metro regions within the U.S. particularly where there is a net zero or negative migration pattern.

How to reflect evolutionary processes in a future synthetic population?

- Use cohort survival with net migration to forecast future controls for person-level variables
- What about households?
 - First attempts to model household formation (marriage, divorce, children, etc) explicitly
- More information needed on household formation and dissolution
 - Correlate trends in person evolution with evolutionary processes in households

Cohort survival methods seem to be most applicable for producing control variables for persons. It is not clear, however, how this should be reflected in the evolution of households. Obviously, we'd need additional information on how households are created, evolve, and dissolve, and how this relates to the characteristics of individuals. To create synthetic households it seems necessary to age households and persons at the micro level, but this requires predicting individual births, deaths, migration and a host of other evolutionary processes. Let's consider how this might play out at the micro-simulation.

Person Evolution

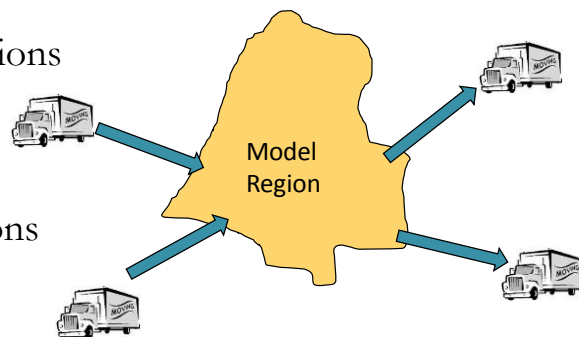
- Aging and Mortality
 - Aging simply adds one to the age of the population each year
 - Mortality rate-based probability or survival model
 - Challenge – changing mortality rates due to medical advances
- Birth
 - Determines if a female gives birth in a given year
 - Challenges:
 - Persons delaying parenthood (socio-economic and cultural tendencies)
 - Incorporating children born outside of marriage/cohabitation (adoptions, foster parents, etc.)

Person evolution would seem fairly straight forward. Aging adds year to the population and, for a given person, it seems straightforward to predict survival to the next cohort year. The challenge is that mortality rates themselves have been changing and will continue to change with advances in medicine.

Birth rates are also fairly straight forward in application; however, as we have witnessed, societal norms have changed over time such that adults increasingly wait until later in life to have their first child. In addition, predicting whether a given household will have children is not so simple due to factors such as out-of-wedlock births, adoptions and foster care.

Emigration and Immigration

- Represents migration of population into and out of the model region
- Rate-based probability models
- Challenges:
 - uncertain economic conditions
 - attributes of emigrants
 - attributes of immigrants
 - disaggregation to sub-regions




In- and out-migration to a region is somewhat straight forward to predict at an aggregate level. At the individual household level, however, there are several challenges. First, there is the macro-economy which may spur sudden in or out-migration and can be quite unpredictable. Second, there are challenges in identifying which households are likely to migrate outside of the region, without some type of longitudinal tracking, because the evidence of their move is more likely to show up in their new location than the one that they just left. Third, knowing the attributes of households and person coming into a region may be a little easier to determine, particularly from state sources; however, it is not clear how those trends will hold up in the future.

Student Evolution

- Education
 - Determines education level of a person in the population
 - Challenge – representing people who discontinue education temporarily and later return to school
- College Students' Residential Choice
 - Determines whether children starting college stay at home or move into a dormitory/rent apartment/new household:
 - MAG Population Synthesizer under development
 - Essential for modeling major universities that generate significant student population around the campus


In order to get a more complete picture of how our synthetic population will evolve, some models even go so far as to account for evolved student status and locations. There are of course challenges, such as predicting persons who take courses part-time and go in and out of student life. In addition, there are challenges associated with predicting whether certain household members will start college/university, and whether that means they will stay local and remain in the same household, or move out of the household into a group quarters situation or into a new household (rent-apartment).




Worker Evolution

- Labor Participation
 - Predicts whether a person joins the labor force
- Occupation Choice
 - Determines the occupational category of a person who is predicted to join the labor force
- Employee Income
 - This model predicts the income/earning level of an employee which can be used to estimate the household income

Activity-Based Modeling: Activity-Based Model Basics



 66

If we have student evolution, then of course we have to have worker evolution, which is even more complicated. There are several types of models associated with worker evolution. First, there are labor participation models, which simulate whether a person joins the labor force and at what participation rate (full-time, part-time).

Second, there is occupation choice modeling, which seeks to determine the occupational category of a person who is predicted to join the labor force.

Third, there are models that try to predict an individual's income, which may be aggregated up to predict household income.

All three of these model types have long been the subject of economic analysis, so there are plenty of examples to go by, some of which are probably overly complicated. The challenge is in finding the right model for simulating individual decisions, and doing it in a household context, where there are inter-individual decisions.

Household Formation and Dissolution

- Family formation/Cohabitation/Marriage
 - Determines marriage/cohabitation decision of adults in the population
 - Roommates/non-family households
- Divorce/Household dissolution
 - Determines divorce decision of adults in the population

The last set of models we will mention are models of household formation and dissolution. These are probably the least well understood of the models we've discussed here. How people come together to form new households would seem to come down to a matching problem and is complicated by the different circumstances that lead to persons living together, such as marriage, cohabitation, and roommates (non-family households).

Perhaps even more murky is modeling household dissolution. While it is somewhat easier to predict children growing up and leaving their parents' home, based on a natural life-course perspective, it is much more challenging and perhaps uncomfortable to try to predict divorces and separation between adults who have been living together.

Household Evolution--Other Challenges

- Choice of model form – discrete choice, regression, hazard-based
- Model sequencing and frequency of simulation
- Validation of the framework and models would itself pose a challenge

There are several challenges that, to date, have prevented household evolution models from making their way into common usage. These include choice of model form – discrete choice, regression, hazard-based; model sequencing and frequency of simulation; and the validation of the framework and models pose a challenge in and of themselves. Thus, this is a very promising research direction but some time is needed to bring it to the necessary level of maturity in practice.

Review: Learning Outcomes

- What a synthetic population is and how they are used in activity-based models
- The methods used to synthesize a population and the various considerations with respect to specifying attributes, including:
 - Specifying controlled and uncontrolled attributes
 - Spatial unit of analysis considerations
 - Methods used in synthetic population generators
 - Challenges to good fit
- Emerging methods in household evolution and why they might be important

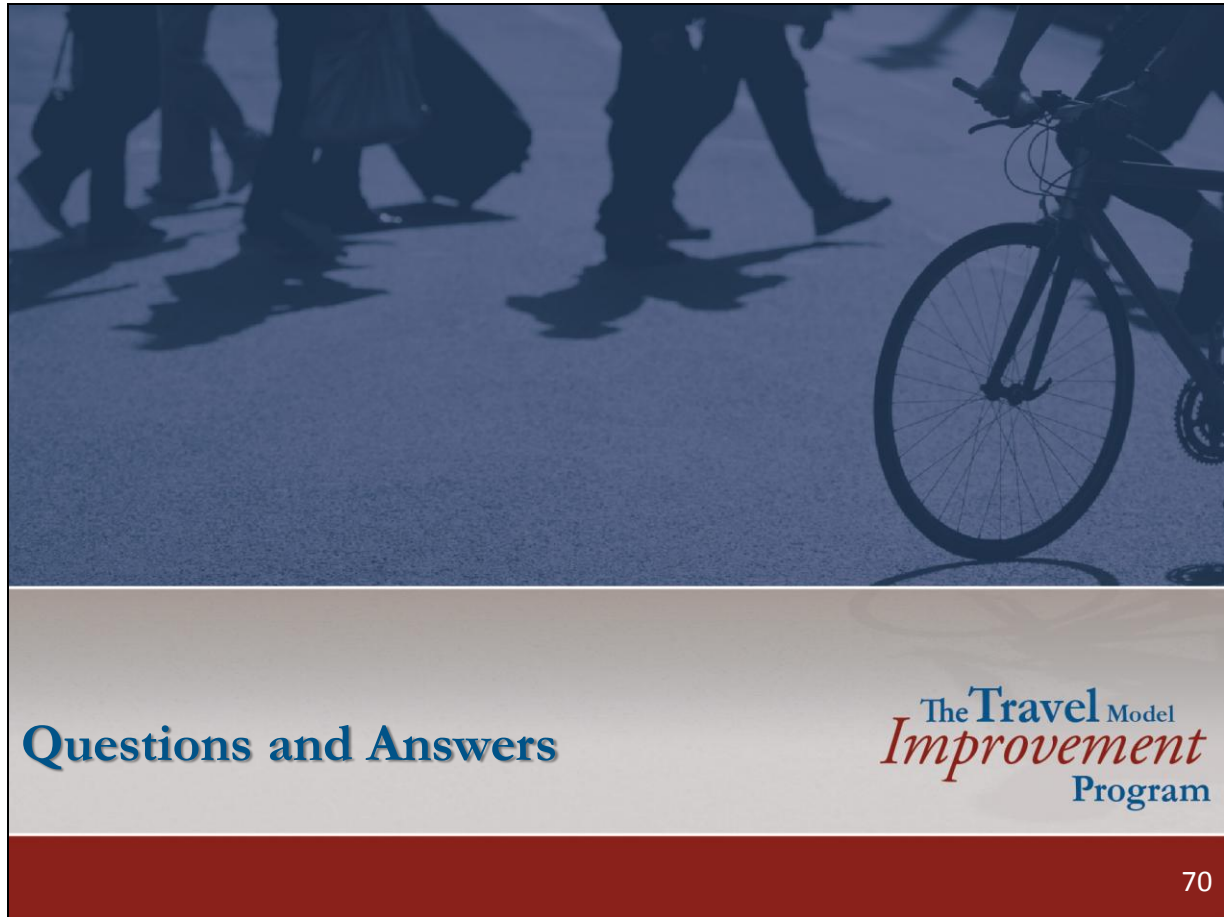


You should now be able to...

Describe a synthetic population

Describe the methods used to synthesize a population


Describe the process of household evolution




2012 Activity-Based Modeling Webinar Series

Executive and Management Sessions	
Executive Perspective	February 2
Institutional Topics for Managers	February 23
Technical Issues for Managers	March 15
Technical Sessions	
Activity-Based Model Frameworks and Techniques	April 5
Population Synthesis and Household Evolution	April 26
Accessibility and Treatment of Space	May 17
Long-Term and Medium Term Mobility Models	June 7
Activity Pattern Generation	June 28
Scheduling and Time of Day Choice	July 19
Tour and Trip Mode, Intermediate Stop Location	August 9
Network Integration	August 30
Forecasting, Performance Measures and Software	September 20

Activity-Based Modeling: Activity-Based Model Basics





71

Once again, here is the schedule for the webinar series. Our next webinar, three weeks from today, will cover accessibility and the treatment of space

Thank you!

Continue the discussion online...

The new TMIP Online Community of Practice includes a Discussion Forum where members can post messages, create forums and communicate directly with other members. Simply sign-up as a new member, navigate to <http://tmiponline.org/Community/Discussion-Forums.aspx?g=posts&t=523> and begin interacting with other participants from today's webinar session on Activity-Based Modeling.



Session 5 Questions and Answers

Is there a disadvantage to specifying too many controls?

Peter: Yes, one disadvantage is that it may not be possible to matching all the control totals that satisfy the joint distribution of all controls. This may be because the disaggregate population does not have any observations which satisfy all constraints. Additionally, it is possible to specify controls that are not consistent with other controls and convergence can be an issue. It is preferred to have a limited number of controls. Advanced methods of population synthesis can be used to mitigate some of these concerns.

Auto availability is an important characteristic in transport modeling. Why isn't auto ownership included as a control in the synthetic population?

John: Auto availability is incredibly important, but we predict auto ownership as a function of demographic variables as well as transport accessibility. We use workplace and school location choice model decisions as well as other accessibility measures in predicting auto ownership, so we don't use it as a control variable.

What methods can be used for more than two control variables?

Peter: The same Iterative Proportional Fitting method can be used for multiple variables; it is just an additional dimension (3 controls equal 3 dimensions, 4 controls equals 4 dimensions). List-based balancing can also be used. However, one must be careful of specifying too many controls and result in combinations of controls that are impossible to fulfill.

Have population models been integrated with economic models, for example REMI?

John: Outputs have been used to create controls as inputs to population synthesizers, for example in the NY region. There is room for more integration in the future. (Note: there are statewide models in Oregon and Ohio that rely upon spatially disaggregate Input\Output models for control totals, particularly for household distributions by size, number of workers, and occupation categories).

Is the universe of households in marginal controls drawn only for the TAZ or for the entire region?

Peter: Marginal controls are typically specified by TAZ. However, they can be allocated down from regional controls for a larger geography and then disaggregated. This might actually be a better approach than generating all controls at a TAZ level.

Is a land-use model required to take into account shifts in population over time, or will a population synthesizer take care of these relationships?

John: A population synthesizer is a procedure that provides a linkage between a land-use model and an activity-based model. It is not designed to shift households across zones, but rather create a disaggregate population that respects input marginal distributions. However, there are hybrid procedures possible, which can both shift distributions of households across zones and generate a synthetic population. This is not typically done; instead, the household distributions are provided by the land-use model and the population synthesizer respects those distributions.