

Introduction .....	2
<i>Census Boundaries</i> .....	2
<i>Database Overview</i> .....	3
<i>Variable Summary</i> .....	5
Estimates and Projections .....	9
Consumer Expenditures .....	14
Retail Potential.....	16
BusinessCounts.....	18
CrimeRisk .....	20
Environmental.....	22
<i>WeatherRisk and QuakeRisk</i> .....	22
<i>Climate</i> .....	23
MOSAIC™ Segmentation .....	24
Consumer Behavior .....	26
<i>INSOURCE (Experian) Profiles</i> .....	26
<i>Simmons Profiles</i> .....	27
Assets, Debts, and Net Worth.....	29
Demographic Dimensions.....	30
2000 Census Database Releases .....	34
<i>Census 2000 SF3 Balancing</i> .....	34
<i>The Census Sampling Issues (For Background Information only)</i> .....	35
<i>Historical Census Issues</i> .....	35
<i>Census 2000 Geographic Areas</i> .....	36
<i>Standard AGS Geographic Areas</i> .....	37
Geographic Definition Changes.....	38
Geographic Area Reference .....	39
About Metropolitan and Micropolitan Statistical Areas.....	41

## Introduction

AGS has an extensive range of market analysis databases available for all standard geographic levels, including block groups, Census Tracts and ZIP codes. With over 18,000 individual variables available for each block group in the country, AGS provides a vast range of data to satisfy most requirements. In order to provide the very best quality data, AGS has chosen its suppliers and partners very carefully. Foremost among these partnerships is a long-term relationship between Experian and AGS, in which Experian provides AGS with its household level INSOURCE™ database that serves as the basis for AGS' current year demographic estimates. AGS in turn provides Experian with access to its demographics in order to update and enhance its MOSAIC segmentation product.

For 2008, the most significant changes occur as a result of some substantial enhancements to the low level allocation of postal data, which was possible by utilizing updated ZIP+4 rosters which were geographically linked to census blocks, some of which were "created" for this purpose. This affects areas which were unpopulated in 2000 but now contains significant concentrations of households (e.g. the old Denver airport).

## Census Boundaries

All boundaries are based on Census 2000 boundaries, as modified in 2002 to reflect post-Census adjustments.

### *Census 2000 Boundaries:*

- Block Group (BG)
- Census Tract (TR)
- Place (PL)
- County Subdivision (CS)
- County (CO)
- State (ST)

### *Geographic Definition Changes:*

There were several changes made to the metropolitan area definitions as of November 2007 and as reported by the Office of Management and Budget in OMB Bulletin 08-01 (which is available at <http://www.whitehouse.gov/omb/bulletins/fy2008/b08-01.pdf>).

These changes are made on an annual basis as a result of a review of the annual Census Bureau place and county estimates. In some cases, the changes do not affect the geographic definitions, but rather only the order and list of city names included in the metropolitan area name.

## Metropolitan Statistical Areas

There were several changes in the Micropolitan areas this year, including the addition of Show Low, AZ (43320), consisting of Navajo County, AZ.

There were several name changes, and three that required new codes -- Helena-West Helena (25760), Washington Court House (47920), and Bradenton-Sarasota-Venice (14600). Please note that in none of these cases were there actual changes in the geographic definitions of these areas.

The distinction between Metropolitan and Micropolitan areas is as given below by the OMB:

"Metropolitan Statistical Areas have at least one urbanized area of 50,000 or more population, plus adjacent territory that has a high degree of social and economic integration with the core as measured by commuting ties. Micropolitan Statistical Areas -- a new set of statistical areas -- have at least one urban cluster of at least 10,000 but less than 50,000 population, plus adjacent territory that has a high degree of social and economic integration with the core as measured by commuting ties. Metropolitan and Micropolitan Statistical Areas are defined in terms of whole counties (or equivalent entities), including in the six New England States. If the specified criteria are met, a Metropolitan Statistical Area containing a single core with a population of 2.5 million or more may be subdivided to form smaller groupings of counties referred to as Metropolitan Divisions." (OMB Bulletin 06-01, Page 2)

The Metropolitan Areas are identified within the AGS data environment with the acronym MA while the combined metropolitan/micropolitan areas are identified as CB (County-Based Statistical Areas).

## Combined Statistical Areas

There were no changes to the CA level of geography this year.

The combined areas are defined by the OMB (Page 2) as:

"If specified criteria are met, adjacent Metropolitan and Micropolitan Statistical Areas, in various combinations, may become the components of a new set of complementary areas called Combined Statistical Areas. For instance, a Combined Statistical Area may comprise two or more Metropolitan Statistical

Areas, a Metropolitan Statistical Area and a Micropolitan Statistical Area, two or more Micropolitan Statistical Areas, or multiple Metropolitan and Micropolitan Statistical Areas that have social and economic ties as measured by commuting, but at lower levels than are found among counties within Metropolitan and Micropolitan Statistical Areas.”

The scale and population of these areas varies considerably, and a useful role for these areas in analysis is doubtful.

#### *Current Year:*

- ZIP Codes

The ZIP code roster is 2nd quarter 2008 and has been constructed from the cartography released by TeleAtlas. Data can be constructed on an ad-hoc basis to correspond to a particular TeleAtlas release as necessary.

## **Database Overview**

### **Census Data**

AGS has processed and makes available the vast majority of data from the 2000 Census and has retained a core variable set from the 1990, 1980 and 1970 Census.

The 2000 Census SF1 and SF3 databases including base population counts, income information, educational achievement and employment information have been incorporated into the current year estimates

### **Estimates and Projections**

The estimates and projections database includes a wide range of core demographic variables for the current year and 5- year projections, covering five broad topic areas: population, households, income, labor force, and dwellings. With a foundation of the Experian household-level databases and over fifteen years of experience in demographic forecasting, AGS offers the highest quality demographic estimates in the marketplace today.

As of the 2005 update, additional improvements were made to the base population and household models which more accurately incorporate changes to the postal delivery counts, which will be most noticeable in new growth areas.

Last year, we continued to improve our efforts to incorporate the Census Bureau's American Community Survey (ACS) results. The ACS is an annual survey which over the course of several years will result in a national rolling estimates database which is eventually intended to replace the decennial SF3 sample database. The ACS results at the county scale are an excellent means of

tracking demographic attributes over the course of the decade.

For this release, significant efforts were undertaken to ensure that high growth areas, especially those which were unpopulated in 2000, were properly estimated.

### **BusinessCounts**

BusinessCounts is a geographic summary database of business establishments, employment, and occupation. The core BusinessCounts data, which now utilizes the industry standard InfoUSA database as its primary source data, includes data to the major SIC group with detailed establishment types.

This 2008 release of BusinessCounts includes new tabulations by NAICS (North American Industrial Classification System) to the three and four digit level for employees and establishments.

### **Consumer Expenditures (CEX)**

AGS provides current year and 5- year projected expenditures for over 390 individual categories that collectively cover almost 95% of household spending. Based on extensive modeling of the BLS Consumer Expenditure Survey, CEX provides reliable estimates of market demand and average household expenditures.

### **Retail Potential**

This tabulation utilizes the 2002 Census of Retail Trade tables which cross-tabulate store type by merchandise line. The Consumer Expenditure data was aggregated to the merchandise line classification and then distributed to each of the major store types.

### **CrimeRisk**

CrimeRisk is the result of an extensive analysis of over seven years of FBI crime statistics. Based on detailed modeling of the relationships between crime and demographics, CrimeRisk provides an accurate view of the relative risk of specific crime types at the block group level. A number of updates were made to this database to include the latest national and metropolitan trends from the UCR (Uniform Crime Reports) publications.

### **WeatherRisk and QuakeRisk**

Many businesses are subject to severe loss because of natural hazards. Using historical records of various weather and earthquake phenomena, these databases provide risk assessment staff with accurate and detailed indexes of relative risk for each hazard type.

### **Climate**

The AGS climate database includes temperature, precipitation, degree-days, and air quality measures. Unlike other databases that are to a county level only, the AGS Climate database provides details to the block group

level. Derived from an extensive analysis of historical climatology data, this database provides a detailed view of local climate, which is vital in merchandising analysis.

## **MOSAIC™ Segmentation**

MOSAIC is Experian's multi-national geodemographic segmentation system, available in over twenty countries worldwide. AGS demographics are an integral part of the MOSAIC system within the United States. MOSAIC licenses include a profiling tool authored by AGS and now available in several countries.

As of 2005, MOSAIC was fully rebuilt and updated using Census 2000, Estimates and Projections and Experian data resources.

The new MOSAIC contains 60 segments (as opposed to the previous 62). The nomenclature has been changed somewhat, in that segments are now numbered within each group starting at one (e.g. B01 instead of B08).

As of 2005, Global MOSAIC was updated. Global MOSAIC is a system which links the segmentation models of a number of countries covering Europe, North America, and the Asia-Pacific region. Consisting of 10 segments (as opposed to the previous 12), the system allows for cross-national analysis and program execution. Further details are available in the on-line documentation at <http://www.appliedgeographic.com/mosaic.html>.

## **Consumer Behavior**

### **INSOURCE Consumer Profiles**

This series of variables are presented as relative indexes, derived from an analysis of the Experian BehaviorBank and Automotive databases using the MOSAIC classification system. Included in this database are a number of variables related to direct marketing

responsiveness.

### **Simmons Consumer Profiles**

Based on Simmons Market Research Bureau (SMRB) surveys, this consumer behavior database offers insight into the consumption patterns and preferences of consumers. A total of 2679 variables have been loaded from the Simmons survey. This is the latest 'doublebase' survey from 2005. Additional variables may be obtained from AGS, as Simmons has provided to us the Choices software which enables extraction of additional variables.

### **Assets & Debts and Net Worth**

This database provides an important look at the financial health of households – including information on the nature and value of both the assets and debts of households, and of the net worth of households. This database is based upon recent surveys of consumer finances undertaken by the Census Bureau, supplemented by statistical modeling in order to provide geographic estimates.

This database was completely remodeled in 2007, which resulted in some changes to the variable list – especially the net worth by value which was eliminated as a result. The database no longer uses a median value as a result of difficulties in working with medians within geographic aggregations.

### **Demographic Dimensions**

This innovative database consists of sixteen core "dimensions" of neighborhoods, such as "Affluence" and "Growth and Stability" which together account for the primary differences between neighborhoods. Based on an extensive statistical analysis of over seven hundred separate demographic attributes, this database is highly useful for undertaking statistical modeling, as each of the variables is essentially uncorrelated to the others.

# Database Methodology Guide

2008 update

Released July 2008

## Variable Summary

Variable Group	1970 Census	1980 Census	1990 Census	2000 Census	Current Year	5-Year Proj	10-Year Proj
<b>Population</b>	X	X	X	X	X	X	X
Population by Household Type		X	X	X	X	X	
Population by Age		X	X	X	X	X	
Population by Age and Sex			X	X	X	X	
Population by Race		X	X	X	X	X	
Population by detailed Race			X	X			
Population by Hispanic origin		X	X	X	X	X	
Population by Race and Hispanic origin			X	X	X	X	
Population by Race and Age			X	X	X	X	
Population by Age and Hispanic origin			X	X	X	X	
Group Quarters population by type			X	X			
Hispanic origin by nationality			X	X	X	X	
Hispanic origin by detailed nationality			X	X			
Marital status			X	X	X	X	
Educational attainment			X	X	X	X	
<b>Households</b>	X	X	X	X	X	X	X
Households by type		X	X	X	X	X	
Household structure/presence children			X	X	X	X	
Population by household type		X	X	X	X	X	
Households by size		X	X	X	X	X	
Age of head of household		X	X	X	X	X	
Length of Residence					X	X	
Vehicles available			X	X	X	X	
Households by race			X	X	X	X	
Households by Hispanic origin			X	X	X	X	
<b>Dwellings</b>			X	X	X	X	
Dwellings by tenure		X	X	X	X	X	
Vacant dwellings by reason			X	X			
Year structure built			X	X			
Value of owner occupied dwellings			X	X			
Rent			X	X			
Units in structure			X	X			
Year moved in			X	X			
<b>Income</b>							
Aggregate income		X	X	X	X	X	
Aggregate income by household type		X	X	X	X	X	
Median/avg income by household type	X	X	X	X	X	X	
Household income distributions		X	X	X	X	X	
Income distributions by household type			X	X	X	X	
Income by age of head of household			X	X	X	X	
Income by race			X	X	X	X	
Income by Hispanic origin			X	X	X	X	
Median disposable income					X	X	
Households by disposable income					X	X	
Households by Asset ownership					X		
Households by Debt holding					X		

# Database Methodology Guide

2008 update

Released July 2008

Variable Group	1970 Census	1980 Census	1990 Census	2000 Census	Current Year	5-Year Proj	10-Year Proj
<b>Labor Force</b>							
Labor force status			X	X	X	X	
Employment by occupation			X	X			
Employment by industry			X	X			
Time leaving for work			X	X			
Means of transportation to work			X	X			
Workers in family			X	X			
Class of worker			X	X			
Employment by disability status			X	X			
<b>Consumer Expenditures</b>							
Apparel					X	X	
Entertainment					X	X	
Food and Beverage					X	X	
Gifts and Contributions					X	X	
Health Care					X	X	
Household equipment					X	X	
Household furnishings					X	X	
Household Operations					X	X	
Insurance					X	X	
Personal Care					X	X	
Reading					X	X	
Retail and non-retail					X	X	
Shelter					X	X	
Transportation					X	X	
Utilities					X	X	
<b>Risk Indices</b>							
Crime					X		
Earthquake					X		
Weather Risk					X		
<b>Climate</b>							
Temperature (Jan/Jul/annual)					X		
Precipitation (snow, rain)					X		
Heating/Cooling Degree Days					X		
Air Quality					X		
<b>MOSAIC Segmentation</b>							
MOSAIC					X		
MOSAIC Groups					X		
Global MOSAIC					X		
<b>INSOURCE Consumer Behavior Profiles</b>							
Automobile Ownership					X		
Contributions to Organizations					X		
Direct Marketing Responsiveness					X		
Direct Marketing Purchasing Types					X		
Household Member Characteristics					X		
Lifestyle Preferences					X		
Mail Order Purchases					X		
Memberships					X		
Musical Preferences					X		

# Database Methodology Guide

2008 update

Released July 2008

Variable Group	1970 Census	1980 Census	1990 Census	2000 Census	Current Year	5-Year Proj	10-Year Proj
Upscale Item Ownership					X		
Technology					X		
Telephone Usage					X		
<b>Simmons Consumer Behavior Profiles</b>							
Alcoholic Beverages					X		
Apparel					X		
Attitudes					X		
Fashion/Apparel					X		
Automobiles					X		
Food					X		
Finance					X		
General					X		
Health and dieting					X		
Internet					X		
Media					X		
Pharmaceuticals					X		
Product Placement					X		
Travel					X		
Technology					X		
Automotive					X		
Beverages					X		
Cable Television					X		
Collectibles					X		
Computers					X		
Contributions					X		
Demographics (Of Sample)					X		
Family Restaurants					X		
Fast Food Restaurants					X		
Financial					X		
Fitness and Sports					X		
Food					X		
Gambling					X		
Games and Toys					X		
Grocery Shopping					X		
Home Improvement					X		
Health and Medical					X		
Home Furnishings and Equipment					X		
Health Products					X		
Lawn and Garden					X		
Leisure					X		
Medical					X		
Media Quintiles					X		
Movies					X		
Music					X		
Parks					X		
Pets					X		
Radio Dayparts					X		
Reading					X		
Sports					X		
Telephone					X		
Theater					X		
Travel					X		

# Database Methodology Guide

2008 update

Released July 2008

Variable Group	1970 Census	1980 Census	1990 Census	2000 Census	Current Year	5-Year Proj	10-Year Proj
Television Dayparts					X		
<b>Demographic Dimensions</b>					X		
<b>BusinessCounts</b>							
Employees					X		
Employees by occupation					X		
Employees by SIC (major group)					X		
Employees by Establishment type					X		
Employees by Land Use					X		
Establishments					X		
Establishments by Size					X		
Establishments by SIC (major group)					X		
Establishments by Detailed Type					X		
Establishments by Land Use					X		
NAICS 3 digit employees					X		
NAICS 3 digit establishments					X		
NAICS 4 digit employees					X		
NAICS 4 digit establishments					X		

Note: Additional 2000 Census tables are not listed in this table. Typically, these additional tables consist of detailed cross-tabulations and additional information related to the housing stock.

## *Estimates and Projections*

### **Content**

The Estimates and Projections (E&P) database is the most extensive update available, covering a broad range of demographic characteristics for the current year, and 5- year projections. Variables include:

- Population
- Population by household type (family, non-family, group quarters)
- Households
- Households by type (family, non-family)
- Households by size of household
- Households by age of head of household
- Household type (e.g. lone parent male family with children)
- Average Household Size
- Population by age (19 age breaks)
- Population by age and sex (38 breaks)
- Population by sex
- Population by race
- Population by Hispanic origin
- Population by race and Hispanic origin (e.g. white Hispanic, white non-Hispanic)
- Population by Marital Status
- Population by Educational Achievement
- Labor Force Employment Status
- Labor Force Employment Status by sex
- Aggregate Income (family, non-family households, group quarters)
- Household income distribution (15 breaks)
- Family income distribution (15 breaks)
- Extended Upper-Income distributions
- Median and average income (family, household)
- Households by disposable income
- Age of head of household by income
- Median income by age of head of household
- Vacant Dwellings
- Tenure
- Length of Residence
- Total Vehicles Available
- Households by number of vehicles available
- Households by number of vehicles available and tenure

### **2008 Update Considerations**

Estimates for the gulf coast area affected by Katrina continue to improve, as the post office household counts continue to better reflect current rather than former delivery statistics. Projections continue to show a rebound of this area over the coming decade. For 2008, we undertook a considerable effort to improve estimates in rapid growth areas, especially those areas which were unpopulated at the time of the 2000 census and therefore lack internal geographic coding (that is, there were few or no census blocks defining these areas specifically). This permits a more robust allocation of the growth into these specific areas rather than into neighboring areas. An example of this is the old Denver airport, in which growth was misallocated to adjacent areas in past years but is now correctly allocated to this redevelopment area.

### **2007 Update Considerations**

With respect to the estimates and projections in the Katrina affected areas, the postal and ADVOC data has finally begun to reflect post-hurricane changes and subsequently, at the block group level we have been able to undertake a much more thorough and data driven update. This, we believe will result in much better small area estimates along the Gulf Coast, but we do note that overall, the population estimates at the parish level remain relatively close to our estimates of last year.

We made changes to the Age by Income distributions. The income distributions for each age group have been extended to 15 from 12, providing better detail within some of the lower income groups. It is vital that you review any existing variable extractions and reports which might utilize this data, as the sequencing of the variables has been changed. For example, the variable HINCYT2512 is no longer "Households age 25-34 with Income >\$200,000" but is now income \$100000-\$124999. See the AGS Variable List (v2) for a complete list of variables.

## 2006 Update Considerations

AGS has been working very hard to find data on the current whereabouts of the residents of the gulf coast. We have been busy gathering intelligence on how many people left the area affected by Katrina, where they went, and how many have come back.

The data from standard sources – e.g. postal service NCOA ("National Change of Address") file, ADVO delivery counts, etc. – were suppressed or missing for the affected areas.

We obtained data on FEMA registrations by ZIP code of original residence and ZIP code of temporary or new residences. These were as of late fall of 2005 and reflect the situation at its worst. We have made some assumptions about the return percentages and adjusted the demographics accordingly.

A Census database which indicates the % of housing declared unlivable by Census Block for the Gulf States was used in conjunction with detailed elevation and flooding maps which were downloaded into a GIS system in order to identify fully the census block groups which were affected. We then excluded those block groups which were not affected. The detailed elevation and flood maps were obtained from federal government sources.

The Census has since released a preliminary Katrina report which details the expected population of the affected gulf counties and the AGS numbers agree very well with those published in the report.

When the Census final report comes out in August, we will again review our numbers in comparison and, if necessary, make appropriate adjustments.

## 2005 Content Changes

While we have not undertaken any changes to the core variable list, we have begun to incorporate the Census Bureau "American Community Survey" (ACS) data into our procedures. This data source now covers over 60% of the population and is to be updated on a rolling basis over time, eventually serving as the replacement for the decennial SF3 tables.

The effects of this inclusion are most strongly noticed in some of the detailed tables where intercensal estimates are either inconsistent or non-existent.

In addition, we adjusted our household estimates downward from the 2004 release on the basis of the ACS and other sources. The ongoing housing boom has resulted in over-construction and an increase in the number of vacant dwellings.

## 2004 Content Changes

Improvements were made to the base population and household models which more accurately incorporate changes to the postal delivery counts, which will be most noticeable in new growth areas.

The following was added:

- estimates of population by detailed age and sex, by single year of age to 20

## 2003 Content Changes

There are substantial changes to this variable content for this release of the estimates and projections database, including:

- We have, with the exception of base population and household counts, eliminated the ten year projections from the database.
- The 2000 Census racial classification has been fully implemented, which includes the "multiple" category and the split of the Asian and Hawaiian/Pacific Islander groups. Variable names are consistent with past releases; however, there are differences in the meaning of each of the variables with the addition of the multiple race response and the

split of the Asian and Hawaiian/Pacific Islander fields.

- We have provided additional detail on the vehicles available table, by breaking it down by tenure (as on the census) and by providing additional detail in the 2+ vehicle categories. The VPHCY and VPHPY variables have been retained as formulas (e.g. NONE, 1, GT1)
- The high income classification (200K+) has been fully integrated into the standard income distributions, which have been reclassified to comply with the 2000 Census definitions in the lower income categories. This has resulted in the renaming of the upper income variables and the elimination of the 200+ income category (which is available as a formula in the library but should not be used within median calculations). In addition, greater detail has been added in the \$50-\$100K categories by estimating at \$5K increments.
- Aggregate and average disposable household income is now included in addition to the distributions, the disposable income distribution classes match the standard income distribution classification.
- The household structure table has been revised to eleven categories in keeping with the tables provided in the 2000 Census and these have been remodeled to agree more closely with other tables (e.g. size of household).
- We have, for the current year, not included employment by occupation or employment by industry. We will revisit these tables on a yearly basis. At present, the Bureau of Labor Statistics does not report current year estimates nor projections using the updated definitions used by the Census 2000 tabulations. These are effectively incomparable for many categories and are unsuitable as the basis for undertaking current estimates and projections. We will review the availability of source data on an annual basis with the intent of reinstating these tables when it becomes feasible.
- Additional detail has been added on labor force participation by splitting the basic labor force status data by sex. This will permit updated unemployment and labor force participation calculations for both males and females, an important aspect for many daytime based studies.
- The age distributions have been adjusted to more closely align with most users expectations of five year cohorts. This affects most age groups under age 25, although we have provided this age distribution in both the old and new formats. We encourage conversion to the new age distributions over the next year.

There are some significant methodological changes which impact the accuracy of the basic household and population data. Specifically, we have integrated delivery counts from both the USPS and commercial sources by linking the ZIP+4 delivery statistics to the census block level of geography. By using comparisons of these delivery statistics and the census data for an equivalent point in time (April 2000), the discrepancies between the census and "postal" counts can be adequately accounted for.

## Methodology and Data Sources

AGS uses a wide range of data sources in constructing its estimates and projections, including:

- Census tabulations from 1970, 1980, 1990 and most recently, the release of the 2000 Census
- USPS and commercial source ZIP+4 level delivery statistics
- Census Bureau estimates and projections of population characteristics at various levels of geographic detail, including the latest estimates of population at the city level.
- The Census Bureau's American Community Survey results, which cover over 60% of the national population and serve as an increasingly important attribute base at the county, metro, and state levels.
- Bureau of Labor Statistics estimates and projections of employment by industry and occupation at the county level
- Medicare eligible population counts at the ZIP code level, including population by sex and 5-year age cohorts, provided by the Health Care Financing Administration of Social Security. These counts provide a very accurate local

count of the population aged 65 and higher.

- Internal Revenue Service statistics on tax filers and year-to-year migration
- The Census Bureau's Current Population Survey, which provides detailed demographic breakdowns and enables a thorough longitudinal analysis of demographic trends
- Experian's INSOURCE™ database, a household level credit and demographic database which covers the vast majority of households

INSOURCE is a vast database at the household and individual level that Experian provides to AGS for use in its demographic estimates. The INSOURCE database was aggregated to the ZIP+4 and Block Group levels of geography for analysis and standardized to Census Bureau county level current estimates. A large number of demographic attributes from INSOURCE were utilized in building the current year estimates, including:

- Population
- Population by Age
- Households
- Household Size
- Household Type (presence of children)
- Marital Status
- Income
- Hispanic origin
- Population of Asian origin
- Dwelling Tenure (own/rent)
- Length of Residence

In turn, the AGS demographic estimates are used as the foundation of Experian's U.S. MOSAIC segmentation system.

Now in its seventh year of use within the AGS estimates methodology, INSOURCE provides an excellent source of small area year-to-year change which greatly improves the quality of local estimates, especially in areas of growth.

The estimates and projections methodology combines the best current and projected information from the data sources noted above. It is supplemented by the extensive experience of Applied Geographic Solutions in creating accurate and reliable estimates and projections. A summary of the methodology for each of the major variable groups is included in the sections that follow.

## Population

The current population of the United States is obtained from the monthly Census population estimate. This is a very accurate and current estimate of the population and serves as the basis for projection and estimation at lower levels of geographic detail. The five and ten year projections have been derived from the middle-series projections of the Census Bureau.

The current year estimates rely heavily on the 2000 Census block level population counts, as these provide the most accurate recent data available. These 2000 Census counts replace the 1990 Census counts as the basis for undertaking estimates. In effect, the latest Census tabulation provides a baseline for the estimates and projections.

State and county level estimates are based on the compilation of data from a range of Federal and State authorities, including the latest county population estimates from the Census Bureau, the American Community Survey (ACS), reviews of building permit statistics, the current population survey (CPS), and additional local sources. Where required, the resulting estimates are then ratio-adjusted so that the sum of the county estimates is equal to the state total, and the state estimates equal to the national total. For the five- and ten-year projections, a similar method is employed. However, rather than using simple straight-line techniques, AGS uses straight-line methods only for growing areas. For declining areas, a log-normal extrapolation is used. This has the effect of slowing decline over time, which is characteristic of long-term population decline at the state level.

At the block group level, the population model consists of the application of a non-linear trend model which estimates

population given historical patterns, INSOURCE population counts, and the latest Census age distributions (using cohort-survival techniques). Special consideration is given to the population age 65+ by applying ZIP code level counts by age and sex of all Medicare eligible persons. This provides considerable improvement in the estimates of this important segment of the population. The final results are then carefully balanced to the county and city level population estimates to ensure consistency with current Census Bureau estimates.

The result is a comprehensive set of population estimates and projections which includes the knowledge of State, County, and private agencies about their detailed areas but also ensures that the total population is consistent with the Census Bureau estimates, which have proved extremely reliable over time.

## Population by Age, Sex, and Race

National and State level Census bureau projections of age by sex and race/Hispanic origin were used as overall controls to ensure consistency with the Census projections. Detailed forecasts by age, sex, and race, as well as Hispanic origin, were obtained from the Census Bureau 'middle series' projections.

At the state level, the projections of individual state agencies and ACS estimates were combined with the results of a cohort survival approach to obtain reliable state estimates by age and sex. The block group estimates were compiled using cohort survival methods, then balanced to both the estimated block group population totals and to the state level control totals. Consistency checks with the annual CPS (Current Population Survey) are used to ensure the validity of the resulting age/sex distributions. Further, INSOURCE population by age summaries were used to adjust local estimates for the adult population, with further adjustments applied using the ZIP code level Medicare eligibility statistics.

Trends in the racial distribution and Hispanic populations were used to derive preliminary estimates at the block group level, which were then adjusted to balance with appropriate control totals. This method allows the utilization of the historical changes in race and Hispanic origin distributions and projects those changes into the future while maintaining consistency with national level projections. Again, the CPS is used extensively to assist in the verification of the models.

## Households and Household Type

Total households were modeled by:

- projecting trends in the population per household over time at the national level to provide a control total;
- reviewing currently available household size statistics at the State level; and utilizing the current estimates of population by age and sex to determine household formation rates for small areas

The ACS data has been extensively used in order to bridge the gap between population estimates and dwelling/postal delivery counts.

All household based numbers are initially estimated / projected separately for family and non-family households. Non-family households have been growing in number at a higher rate than family households have over the past several decades. Average household sizes for family households have been decreasing for several decades. However, during the 1990's, the decline has stopped in most areas and has actually reversed in several states.

The group quarters population, that is population that is not in households (such as persons in institutions, military barracks, nursing homes, college dormitories, and homeless persons), is expected to increase slightly during the decade, but remain relatively constant as a percentage of the total population. This is a reflection of two trends: the decreasing armed forces employment since the 1980's and the longer term increasing elderly population which results in high populations in nursing homes and other institutions which cater to the elderly population. As a result, the total group quarters population has been relatively constant.

## Income

Income estimates include aggregate income by household type and income distributions as well as derived measures include per capita income, and various median income measures.

All income estimates produced by Applied Geographic Solutions are in current, rather than constant, dollars. In other words, a

projection of income for the year 2010 includes both an inflationary component and a 'real' component, the latter being the difference between the change in income and the change in inflation during the period. The 'real' component is normally attributed to productivity gains in the economy and to differences in the international competitiveness of the economy.

Aggregate income estimates for the current year are based on an analysis of income information from the SF3 database of Census 2000, supplemented heavily by the 2003 ACS estimates. The projections of aggregate income are based on a review of Bureau of Economic Analysis (BEA) projections, which assume an effective increase of 3.5% per annum in per capita incomes during the next ten years at the national level.

Income distributions are estimated and projected for both family households and non-family households separately. Total household income distributions are simply the aggregate of the two detailed distributions.

Income distributions were derived by using a complex distribution shifting technique which utilizes the changes in per family household and non-family household incomes as a means of adjusting the income distributions over time. The relative ratio between changes in per household average incomes and median incomes were used to adjust for above-average growth in high-income households within some geographic areas. The resulting distributions were then normalized to higher order totals and adjusted to national level expectations and were verified for internal consistency with respect to the mean and median measures.

As of the 2003 release, for the current year estimates, a new set of income breaks are provided for the \$150,000+ category, namely \$150000-\$199999, \$200000-\$249999, \$250000-\$499999, and \$500000+. Created by using logistic regression techniques that account for the local income distribution, these should be considered as maximum likelihood estimates. Although little data exists to substantiate incomes in these ranges, comparisons have been made to IRS taxation statistics to ensure that the results are consistent. Users are cautioned that these estimates are statistical in nature only.

## Other Variables

A number of other variables are also projected within the series. In large part, these are derived by using available current estimates and projections at the lowest possible level of geography as the base for the estimation procedures, relying heavily upon the annual release of the ACS. The CPS is used extensively to track changes using available cross-reference information related to age, race, sex, and income. Where possible, these CPS statistics are supplemented by INSOURCE estimates.

For example, current marital status estimates are available at the state level ACS from the Census Bureau as "control targets". The ACS is used in conjunction with the annual CPS surveys (both historical and current) are used to track the changes in marital status dependent upon other symptomatic variables such as age, sex, race, and income levels. These "micro-models" are then applied to the block group level changes between the census and the current period. This results in block group level data which is consistent with higher order levels but also reflects changes in marital status owing to shifting local demography.

On the other hand, vacant housing is tracked using state and regional indicators, then adjusted for seasonally vacant dwellings which are a significant component of the marketing landscape in many areas of the country.

## ***Consumer Expenditures***

### **Content**

The Consumer Expenditure database covers most major household expenditures in a multi-level hierarchical classification. Expenditures can be expressed either as aggregate expenditure or per household expenditure for any geographic level from the block group to national. The major categories represented are:

- Total Expenditure
  - Food and Beverages
  - Shelter
  - Utilities
  - Household Operations

- Household Furnishings/Equipment
- Apparel
- Transportation
- Health Care
- Entertainment
- Personal Care
- Reading
- Education
- Tobacco Products
- Miscellaneous Expenses
- Cash Contributions
- Personal Insurance
- Gifts

Most of these categories include two or three levels of sub-category detail. For example, a typical classification for an item in the food group is:

#### **Total Expenditure**

- FB Food and Beverage
  - FB1 Food At Home
    - FB106 Dairy Products
      - FB10604 Cheese

This structure permits ready analysis of expenditures at any level of detail and between levels of detail. It is possible to analyze any individual category within the context of its parent category (e.g. cheese expenditures as a share of total dairy product expenditures or total food at home expenditures).

#### **Methodology and Data Sources**

The consumer expenditure database consists of a multi-level hierarchical classification of household expenditures, which covers the majority of annual household expenditures. It is derived from an extensive modeling effort using the 2005 Consumer Expenditure Survey data from the Bureau of Labor Statistics. The BLS survey is a comprehensive survey that averages over 7,500 households four times a year using a rotating sampling frame. The use of several consecutive years of data provides a rich base of expenditure data from which to build expenditure models based on household demographics.

The database consists of a total of 396 base variables, which are aggregated in up to four levels of detail. A hierarchical structure is utilized throughout, so that it is possible to aggregate or disaggregate categories as required for analysis.

The survey includes a wide range of demographic attributes related to "consumer units" (generally households), which have been modeled separately for each discrete expenditure category. The older surveys were first inflated to the current price levels using the detailed consumer price index series. For each individual expenditure category in the survey, summary statistics were calculated for each separate element in the list below. In several cases, it was possible to utilize cross tabulation data (e.g. income by age of head of household). These variables are listed below:

- geographic region (Northeast, South, Midwest, West)
- metropolitan status (metropolitan, non-metropolitan)
- housing tenure (owner or renter)
- age of head of household (<25 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, 65+ years)
- size of household (1 person, 2 persons, 3 persons, 4 persons, 5 persons, 6+ persons)
- household income (<5000, 5-10000, 10-15000, 15-20000, 20-30000, 30-40000, 40-50000, 50-70000, 70000+)
- race (White, Black, American Indian, Asian) and Hispanic/Non Hispanic
- number of vehicles (none, 1, 2+ vehicles per household)

The total sample was utilized to obtain an average expenditure for each item. For each expenditure item, a series of adjustment factors were derived for each unique demographic attribute. These adjustment factors were then applied to the block group level using the same demographic variables in order to create estimates at the local level, which are consistent with local characteristics. Consistency checks were undertaken in order to ensure that the results at the block group level were consistent in the aggregate with overall income levels and published expenditures. Finally, the estimates were inflated using detailed consumer price indexes to current levels.

In total, there are 396 detail categories that can be aggregated using the field name. The field name will in all cases begin with the three-character sequence XCY (for an average expenditure) or TCY (for total geographic area expenditure) in order to distinguish these variables from those of other databases and from other years. The next two characters are the major group (e.g. AP for apparel). The primary detail level is a one-digit number (e.g. AP1 is men's apparel). Two sequences of two digits then follow to indicate the remaining two levels of potential detail. The entire variable list is included in the file layout section.

In addition to providing average household expenditures, AGS also provides total market estimates for use in market share and demand analysis.

## ***Retail Potential***

### **Content**

The retail potential database consists of average household and total market potential estimates by each of sixty-eight retail store types. The store types are based on the NAICS classification and are listed below:

- 44111 New Car Dealers
- 44112 User Car Dealers
- 44121 Recreational Vehicle Dealers
- 44122 Motorcycle and Boat Dealers
- 44131 Auto Parts and Accessories
- 44132 Tire Dealers
- 44211 Furniture Stores
- 44221 Floor Covering Stores
- 44229 Other Home Furnishing Stores
- 44311 Appliances and Electronics Stores
- 44312 Computer Stores
- 44313 Camera and Photography Stores
- 44411 Home Centers
- 44412 Paint and Wallpaper Stores
- 44413 Hardware Stores
- 44419 Other Building Materials Stores
- 44421 Outdoor Power Equipment Stores
- 44422 Nursery and Garden Stores
- 44511 Grocery Stores
- 44512 Convenience Stores
- 44521 Meat Markets
- 44522 Fish and Seafood Markets
- 44523 Fruit and Vegetable Markets
- 44529 Other Specialty Food Markets
- 44531 Liquor Stores
- 44611 Pharmacy and Drug Stores
- 44612 Cosmetics and Beauty Stores
- 44613 Optical Goods Stores
- 44619 Other Health and Personal Care Stores
- 44711 Gasoline Stations with Convenience Stores
- 44719 Gasoline Stations without Convenience Stores

# Database Methodology Guide

2008 update

Released July 2008

44811 Men's Clothing Stores  
44812 Women's Clothing Stores  
44813 Childrens' and Infant's Clothing Stores  
44814 Family Clothing Stores  
44815 Clothing Accessory Stores  
44819 Other Apparel Stores  
44821 Shoe Stores  
44831 Jewelry Stores  
44832 Luggage Stores  
45111 Sporting Goods Stores  
45112 Hobby, Toy, and Game Stores  
45113 Sewing and Needlecraft Stores  
45114 Musical Instrument Stores  
45121 Book Stores  
45122 Record, Tape, and CD Stores  
45211 Department Stores  
45291 Warehouse Superstores  
45299 Other General Merchandise Stores  
45311 Florists  
45321 Office and Stationary Stores  
45322 Gift and Souvenir Stores  
45331 Used Merchandise Stores  
45391 Pet and Pet Supply Stores  
45392 Art Dealers  
45393 Mobile Home Dealers  
45399 Other Miscellaneous Retail Stores  
45411 Mail Order and Catalog Stores  
45421 Vending Machines  
45431 Fuel Dealers  
45439 Other Direct Selling Establishments  
7211 Hotels and Other Travel Accommodations  
7212 RV Parks  
7213 Rooming and Boarding Houses  
7221 Full Service Restaurants  
7222 Limited Service Restaurants  
7223 Special Food Services and Catering  
7224 Drinking Places

While similar to the SIC classification, the NAICS recognizes several retail types which did not exist at the time the SIC system was defined, including Computer Stores, Home Centers, and Gasoline Stations with Convenience Stores, to name a few.

## Methodology and Data Sources

The primary data sources used in the construction of the database include:

- Current year AGS Consumer Expenditure Estimates
- 2002 Census of Retail Trade, Merchandise Line Sales
- Census Bureau Monthly Retail Trade

The Census of Retail Trade presents a table known as the Merchandise Line summary, which relates approximately 120 merchandise lines (e.g. hardware) to each of the store types. For each merchandise line, the distribution of sales by store type can be computed, yielding a conversion table which apportions merchandise line sales by store type.

The AGS Consumer Expenditure database was re-computed to these merchandise lines by aggregating both whole and partial categories, yielding, at the block group level, a series of merchandise line estimates which are consistent with the AGS Consumer Expenditure database.

These two components were then combined in order to derive estimated potential by store type. The results were then

## Applied Geographic Solutions, Inc.

Newbury Park, CA 91320

(877) 944-4AGS

<http://www.appliedgeographic.com>

compared to current retail trade statistics to ensure consistency and completeness.

## ***BusinessCounts***

### **Introduction**

BusinessCounts is a geographic summary database of business establishments and employees for over twelve million companies and one hundred and ten million employees. The database is available at the block group level and higher, including all standard geographic aggregations.

BusinessCounts is a vital addition to residential demographic data, in that the success of many business establishments is dependent upon not only the residential population, but also the working population during the daytime. Based primarily on the InfoUSA business database and supplemented by various public data sources, BusinessCounts provides a clear look at the range and size of establishments and their employees within any geographic area.

As of the 2003 release, there were significant made to this data in order to eliminate variables which we do not feel are sufficiently well supported by source data, or have ceased to be sufficiently well supported, to merit inclusion in the database. These changes are outlined below:

### *Occupation*

We have expanded the occupation table from thirteen to seventeen categories as a result of an improvement to the underlying industry-occupation tables released by the Bureau of Labor Statistics. The sales occupation (BCCCYSALLES) has been split into two categories: Sales Professionals (BCCCYSALPR) and Sales Workers and Clerks (BCCCYSALCL). The clerical group (BCCCYCLER) has been split into Administrative Support Workers (BCCCYCLERO) and Technical, Sales, and Administrative Field Operations (BCCCYCLERF). The "Other" Services group has been broken into Site Based (BCCCYSERV1) and Field Based (BCCCYSERV2). Finally, the Production occupations group has been split into site and field based operations (BCCCYPRODS, BCCCYPRODF).

These changes will be very useful to retailers and food services companies who are most interested in the daytime population, as many of these field operations staff who may be assigned to a particular location are in all likelihood absent on most workdays from that location.

### *Business Functional Classification*

This set of 38 business types, based roughly on land use style, is adapted from one which was used several companies during the early 1990's and is, in our view, a very useful way of presenting business data to clients on the style of operation rather than the specific type of products produced. The variable ESCCYMFGSM, Manufacturing: Small Firms is a good indication of the presence of small industrial parks. A small manufacturer of plastics, for example, is likely to be different in its land use and employment profile than a large producer.

### *Administrative/Headquarters*

We have added counts of employees and establishments for administrative or headquarters firms, as identified in the InfoUSA file.

### *SIC Classifications*

The classification by two-digit SIC code has been slightly re-worked in order to merge certain small related categories while splitting other large categories. The result is a slightly shorter list of categories at this detail level. For example, the categories 01 (Crop production), 02 (Livestock production), 08 (Forestry), and 09 (Fishing) have been aggregated into a single category.

### *NAICS Classifications*

A few years ago, a new classification system known as NAICS (North American Industrial Classification System) was introduced to replace the aging and outdated SIC (Standard Industrial Classification) system, which has been used for several decades (with some modifications over time).

The NAICS classification is intended to more accurately reflect the growing reliance on information-based companies which

# Database Methodology Guide

2008 update

Released July 2008

were not well classified under the SIC system, reflect the changes in retailing towards large, multi-faceted retailing (e.g. home stores, grocery and drug stores, etc.), and to eliminate business types which simply no longer exist. In most cases, the orientation is more towards the product/service being offered and less towards the style or raw materials of its creation. For example, the textile manufacturing SIC group has been split by whether the product is an "end-user" product or is simply a raw material in further manufacturing. The manufacture of automobile seat fabrics is now classified under the automobile manufacturing group rather than the textile manufacturing group, reflecting the ultimate use of the product rather than its composition.

NAICS classification data are presented at the 3 and 4 digit detail levels. A 2 digit summary can be readily created using the formula features of Snap, should such be desirable. These are roughly comparable to the SIC "Major Industry" and 2 digit levels, although there are more categories under NAICS than under SIC.

## Content

BusinessCounts is a geographic summary database of business establishments and employees for nearly ten million businesses and one hundred and thirty million employees. The database is available for all standard levels of geography including block group.

BusinessCounts is a geographic compilation of the InfoUSA business list, supplemented by occupational data from the Bureau of Labor Statistics and the County Business Patterns program. The primary variables available include:

Total	Establishments, Employees,
Size	Establishments by size
Occupation	Employment by occupation
Major Industry	Establishments, Employees
NAICS	Establishments, Employees 3 and 4 digit

## Methodology and Data Sources

The core source for the InfoUSA Business Database that is built from a careful integration of commercial databases, compiled white and yellow page directory data, city directories, corporate annual reports, and securities filings. The BusinessCounts file is current to January 2008.

In years past, a different data source was used by AGS to compile this database, and users should review the notes at the end of this document that outline the type and scope of the impacts of the change in source data. The primary changes that will be noted by users include:

- The ability to release establishment level data for use in mapping applications, with selection based on company name, SIC, geographic area, and company size
- A greatly expanded number of establishments, many of which are small and unclassified, but nevertheless reflect changes in the corporate landscape
- Improved SIC coding at establishments which include more than one major industrial group
- Reduced duplication of records – and subsequent over counting of employment – at companies which contain multiple legal entities at the same address

The database has been thoroughly cleansed for address consistency and geocoded. Virtually all records within the database are geocoded, although in some cases with less positional accuracy than others.

## SIC Classification

A significant number of establishments are not SIC coded within business list files, most often including those small firms for which neither private nor public records exist. Many of these uncoded firms are simply individual holding companies, DBA ("doing business as") names, and new firms that have not yet been well documented in multiple sources.

## Employees

The file includes both a size classification (e.g. 1 to 4 employees) and in a significant number of records, verified employee counts. When an actual employee count was given, this was used directly.

In order to estimate employees for those establishments either only a size class range is available, the latest County Business Patterns (CBP) database, published annually by the Census Bureau was thoroughly analyzed. For each four digit SIC code,

## Applied Geographic Solutions, Inc.

Newbury Park, CA 91320

(877) 944-4AGS

<http://www.appliedgeographic.com>

the average number of employees per establishment of each size class was computed in order to provide a base estimate. These were further refined by using major industry average sizes by county, since much of the county level detail is suppressed within the CBP in order to avoid the possible disclosure of individual establishment employee, payroll, and sales volumes.

Once the initial estimates were applied, the results were evaluated on a county level basis in order to ensure consistency with county totals for each major SIC group, and nationally to ensure consistency with the detailed four-digit SIC level.

It should be noted that the employee size estimates for the Public Administration (SIC 91-98) major group are not particularly accurate. Employee estimates for individual government offices are simply not easily obtained and are generally afforded less attention by the major business list providers than private sector establishments. Further, neither the CBP nor the Economic Census databases cover this important sector of the economy. The total employee estimate is therefore rather low for this sector as a whole.

### **Occupation**

The occupation estimates were created using the 1996 Industry Staffing Patterns Estimate file, obtained from the Occupational Employment Statistics (OES) Survey, conducted by the Bureau of Labor Statistics. The occupational classification used by the OES was converted to the Census Occupational categorization using a translation file obtained from the NOICC Crosswalk and Data Center (NCDC), a technical resource center of the National Occupational Information Coordinating Committee (NOICC).

### **Comparability to Other Sources**

Several additional sources of national and state level estimates from the BLS (Bureau of Labor Statistics) and the Census Bureau were used to verify summary counts in the final database. In general, the database agrees substantially with these estimates. The major sources of difference occur in several areas. These areas of disagreement are noted below:

- Manufacturing employment in BusinessCounts is higher than corresponding statistics from the BLS. In large part, this reflects the use of the "primary" industry within BusinessCounts. Often, many manufacturing companies also have wholesale trade and finance divisions. The employment within these divisions is attributed to the main SIC category in this file, hence manufacturing estimates are higher than in BLS sources.
- Agricultural establishments, specifically farms, tend to be underreported in the database, so total agricultural establishment and employment counts are low relative to other sources. The so-called "primary" sector is not typically well represented in either the economic censuses or the annual County Business Patterns files, and is subsequently difficult to estimate with reliability.
- Service employment in the BusinessCounts file is higher than in equivalent BLS sources, primarily as a result of classification issues. In official BLS reports, educational institutions and employment is reported within the appropriate level of government (e.g. state versus local) whereas in BusinessCounts, these are reported in the educational services category.
- In addition, many public and quasi-public agencies are coded to the type of service they provide rather than as public sector establishments. Public sector estimates in BusinessCounts are therefore lower than published figures by an amount roughly equivalent to the over-count in services. In addition, the tendency within business list products such as InfoUSA is to put more emphasis on private sector establishments than on public sector establishments. Subsequently, in many cases not only is there no actual employee count, but often no size class information as well. Since the Census Bureau surveys of establishments typically exclude public sector establishments, and what statistics are available are typically only at a state level, the public sector employment estimates are substantially underestimated and should not be relied upon for many analytical applications.

## ***CrimeRisk***

### **Content**

CrimeRisk is a block group and higher level geographic database consisting of a series of standardized indexes for a range of serious crimes against both persons and property. It is derived from an extensive analysis of several years of crime reports

from the vast majority of law enforcement jurisdictions nationwide. The crimes included in the database are the "Part 1" crimes and include murder, rape, robbery, assault, burglary, theft, and motor vehicle theft. These categories are the primary reporting categories used by the FBI in its Uniform Crime Report (UCR), with the exception of Arson, for which data is very inconsistently reported at the jurisdictional level. Part II crimes are not reported in the detail databases and are generally available only for selected areas or at high levels of geography.

In accordance with the reporting procedures using in the UCR reports, aggregate indexes have been prepared for personal and property crimes separately, as well as a total index. While this provides a useful measure of the relative "overall" crime rate in an area, it must be recognized that these are unweighted indexes, in that a murder is weighted no more heavily than a purse snatching in the computation. For this reason, caution is advised when using any of the aggregate index values.

## Methodology

The primary source of CrimeRisk was a careful compilation and analysis of the FBI Uniform Crime Report databases. On an annual basis, the FBI collects data from each of about 16,000 separate law enforcement jurisdictions at the city, county, and state levels and compiles these into its annual Uniform Crime Report (UCR). The latest national crime report can be obtained either from the FBI web site in Adobe Portable Document (PDF) format or can be ordered directly from the FBI. While useful, the UCR provides detailed data only for the largest cities, counties, and metropolitan areas.

The original analysis was undertaken by obtaining detailed jurisdictional level data for the years 1990 through 1996, which were supplemented with 1999 preliminary UCR statistics at the State level and for cities and metropolitan areas where those have been released. We are now using UCR data from 1998-2006. The preliminary 2007 release data was used to balance the models to the latest available data.

A considerable effort was made to correct a number of problems that are prevalent within the FBI databases, including:

- The standardization of jurisdictional names: the FBI does not employ Census bureau codes in its databases and the jurisdictional names contain numerous typographical errors and format discrepancies which needed to be manually corrected
- Reporting by individual jurisdictions can be inconsistent from year to year, in that data for some jurisdictions is missing for one or more years and required handling
- Reporting for some crime types is inconsistent between jurisdictions. The FBI handles this by simply suppressing the statistics entirely for those areas. This primarily affects the rape category for Illinois, where statistics are suppressed for all but the largest jurisdictions. These missing values were handled via the modeling process, in which rape estimates were prepared for these jurisdictions by using a model which related rape incidence to other crime types
- The standardization of the database to account for jurisdictional overlaps. For example, the California Highway Patrol has jurisdiction over only state and Interstate highways in urban areas.
- Crime rates in general have been declining over the past several years, so it was necessary to adjust the historical data to reflect current crime rates.

Once this correction and standardization effort was completed, the database consisted of a time series of six years of data covering:

- All cities and towns which have their own police agency
- All cities and towns where policing for the local jurisdiction is contracted to a higher level agency but which tracks statistics separately (e.g. the city of Thousand Oaks, California contracts with the Ventura County Sheriff's Department for police services, but the incident reports are separately compiled)
- A record for each county which covers the population not covered by either of the two cases above. This is normally either a County Sheriff (or equivalent) or a State level jurisdiction which reports incidence of crime by county (e.g. in New York, the State Trooper).

For a very limited number of areas, such as New York City, the local jurisdiction spans several counties.

The initial models were undertaken using a subset of this database. In the smallest cities, a single murder will have a profound effect on the crime rate per 100,000 population that would severely distort the resulting models. Cities with less than 2,500 people were reassigned to their parent counties for the purpose of the analysis. A wide range of 1990 Census and current year demographic attributes was extracted from AGS' databases for the remaining areas (approximately 8,500

separate “jurisdictions”). This database was then used as the primary modeling database and was used later for scaling purposes.

Each of the seven crime types was modeled separately, using an initial range of about 65 socio-economic characteristics taken from the 2000 Census and AGS’ current year estimates. Separate models were constructed for each of the nine Census regions (e.g. New England, East North Central, Pacific) in order to account for regional differences in crime rates and the demographic characteristics which underlay them. The models constructed typically accounted for over 85% of the variance in crime rates at this “jurisdiction” level, although it should be noted that the results for property crimes were generally more reliable than for personal crimes.

The results of these models were then applied to the block group level using the same demographic attributes compiled at the block group level. The resulting estimates were then scaled to match the master database of 8,500 jurisdictions. For cities, the block groups within each city were scaled to match the city total. For areas outside of these cities (or for smaller centers), results were scaled to match the county total after adjusting for those cities scaled separately.

The final crime rate estimates were then weighted by population and aggregated to the national totals. The results were then scaled to match the 2007 preliminary estimates (at a state level) and converted to indexes relative to the national total.

## ***Environmental***

The AGS environmental databases consist of several separate database components, which include:

- WeatherRisk
- Hurricane Risk
- Tornado Risk
- Damaging Wind Risk
- Hail Risk
- Temperature
- Precipitation
- Degree Days
- Air Quality

## **WeatherRisk and QuakeRisk**

### **Content**

The WeatherRisk database consists of four separate types of weather-related hazards: hurricanes, tornadoes, hail, and damaging winds. The data are the results of a series of spatial analysis carried out on records compiled from publicly available USGS sources aimed at producing risk index estimates at the block group level and above.

The base data are made available for either analysis by the end user or map display for each of the four weather types. These files are currently available in MapInfo format as single nationwide files. Point files are available for both wind and hail incidents, while the hurricane data is represented as a series of lines. The tornado data is a mixture of point based (single known point of contact) and line based (path of contact). These files are included with the appropriate database files.

These cartographic databases, while certainly interesting, do not provide any “actionable” information to the user, as it is extremely difficult to interpret the likely risk for any given point using historical location data. The spatial analysis undertaken is based on several underlying facts:

- At a “macro” scale, there is a clear pattern of incidents of any type (e.g. “tornado alley”)
- At a “micro” scale, the particular path which a single tornado or hurricane takes, or the precise location of high wind

incidents or hail is essentially a random occurrence. It is only through the accumulation of a large number of historical records that the randomness at the local scale begins to show a pattern at a regional scale.

As such, a simple count of how many tornadoes have passed through any particular block group is of no value, as this certainly falls within the "micro" scale. Given a long enough historical record (e.g. several thousand years), this might be an appropriate technique for evaluating the potential risk. However, given the relative shortness of these data series, a simple arithmetic exercise is not sufficient. Instead, for any particular point occurrence (e.g. hail observation) a conical filter was applied using a simple distance decay measure. For path events (e.g. a tornado path), a distance-decayed linear filter was applied. For any particular point in space, the accumulated probabilities could then be calculated by summing the areas underneath these conical and linear filters.

All of the resulting indexes are "100" based, which means that a value of 100 for a particular level of geography is the average national value. A value of 200 indicates that the area has two times the average risk level, while a value of 50 indicates that the area is at half the average risk level. For example, a value of 200 for the "HailIndex" indicates that the particular area is two times as likely to suffer hail damage in any given time period than an area with a 100 score.

## Methodology and Data Sources

Hurricane track data was obtained from publicly available USGS records. Atlantic hurricane coverage is from 1896 to 1996, covering a total of 951 storms. Pacific hurricane coverage is from 1949 to 1996, covering a total of 661 storms. Storm locations are tracked every six hours while the storm maintains the minimum wind speed required to be classified as a tropical storm. Along with location, the database includes information on wind speed and barometric pressure.

The risk indexes were derived using a distance decay spatial filter along the line of the storm track with a width of 100 miles each side of the storm track. Statistics at the block group level were then compiled by computing summary statistics of hurricane impact at the block group centroid.

Tornado records published by the USGS from 1950 were analyzed for the purpose of identifying relative risk at the block group level. Unlike hurricanes, which are always presented as a hurricane path, tornadoes are presented either as a path or as a single touchdown point. A total of 38497 separate tornado events were analyzed. Similar spatial filters to those described under hurricanes were applied to both the point and path data.

Reports of damaging hail (over 0.75 inch in diameter) were compiled from USGS data sources, consisting of 86,675 records dating back to 1955. Point filters were applied to this database to derive relative frequency and intensity measures at the block group level.

The WindRisk data elements are based on reported events with wind speeds exceeding 50 knots, and consist of 115,814 separate events dating from 1955.

The composite risk index presents a unified risk index based on the relative damage expected from each of the four types of events. The relative weights of each of the source indexes were derived by weighting estimates of total annual damage caused by storms of each type.

The QuakeRisk database consists of two separate components. The first is a MapInfo point file showing the locations of significant earthquakes during this century. The quality of the additional information is significantly improved in recent years. Quakes in the 3.0 range are included only for the very recent past, while large quakes are tracked back to the turn of the century.

The second, and more important component, is a block group and higher level database which presents the risk of damaging earthquakes on a 100 based scale. This is currently available for only the continental United States and has been derived from USGS models using 0.1-degree grids (except in California and Nevada, where a 0.05-degree grid was used).

## Climate

### Content

The AGS climate database consists of the following variable groups:

Temperature:	Average, maximum, and minimum daily for January, July, and annual
Precipitation:	Annual rainfall and snowfall
Degree Days:	Annual average heating and cooling degree days
Air Quality:	Air quality indexes

## Methodology and Data Sources

The climate database was created from two separate sources. The temperature, precipitation, and degree days variables were derived from an analysis of weather observations from federal government sources. In order to derive values for individual block groups, the data for each observation point (over ten thousand in all) were analyzed using Vertical Mapper in order to estimate the likely values at block group centroids.

The air quality indexes were derived from data obtained from the EPA and modeled using similar methods.

## *MOSAIC™ Segmentation*

### Content

MOSAIC is a geodemographic segmentation system developed by Experian and marketed in over twenty countries worldwide. Each of the nearly one-quarter million block groups were classified into sixty segments on the basis of a wide range of demographic characteristics. The basic premise of geodemographic segmentation is that people tend to gravitate towards communities with other people of similar backgrounds, interests, and means. MOSAIC is linked to the systems in other nations through the Global MOSAIC classification, which consists of fourteen market segments found in every modernized country.

MOSAIC is one of over twenty neighborhood classification systems built by Experian staff, whose international segmentation experiences stretches back over twenty years. Along with the international experience applied in this product, some of the most experienced geodemographers in North America were involved with the development of MOSAIC. During the product refinement process, MOSAIC was compared to other clustering systems in a variety of tests. The MOSAIC assignments are updated annually by incorporating updated AGS demographics into the segmentation model, ensuring that the assignment is as accurate as possible given shifts in local area demographics.

AGS used a very wide range of variables from the 2000 Census at the block and block group levels in order to build the new MOSAIC system. In total, the number of variables used in the initial analysis was well in excess of 600. The group categories of variables included in the creation of the MOSAIC typology is listed below:

- Population by Age and Sex
- Population by Race and Hispanic origin
- Educational Attainment
- Educational Enrollment
- Marital Status
- Group quarters population by type
- Place of birth
- Foreign born by year of entry
  
- Households by type
- Size of household
- Household type by presence of children
- Age of head of household
- Language spoken at home and linguistic isolation
- Residence in 1995 (Stability)
- Tenure
- Vehicles available

# Database Methodology Guide

2008 update

Released July 2008

- Households by income
- Median income, average per capita income
- Median income by age
- Households by type of income
- Workers in family
- Income/Poverty ratio
  
- Labor force status by sex (incl. military)
- Labor force participation rate
- Employment by occupation
- Employment by industry
- Class of worker (e.g. private corporation, federal gov't, unpaid family, etc.)
- Veteran status
- Travel time to work
- Worked at home
- Dwellings by occupancy status (owned, rented, vacant)
- Housing value of owner occupied housing
- Median housing value
- Contract rent
- Median contract rent
- Units in structure
- Year structure built
- Median dwelling age
- Mortgage status (e.g. no mortgage, first only, first and second)
- Year moved in
  
- Population density
- MSA size
- Distance to MSA center

The resulting segmentation system consists of sixty segments which are presented as twelve separate groups:

- A Affluent Suburbia
- B Upscale America
- C Small-town Success
- D Blue-collar Backbone
- E American Diversity
- F Metro Fringe
- G Remote America
- H Aspiring Contemporaries
- I Rural Villages & Farms
- J Struggling Societies
- K Urban Essence
- L Varying Lifestyles

## Global Mosaic

The Global Mosaic system allows for the linkage of customer data and analyses between the U.S. and other major western markets. Global Mosaic has been recently rebuilt by Experian and released for –

- Australia
- China (Major metro areas only)
- Denmark
- Finland
- France

## Applied Geographic Solutions, Inc.

Newbury Park, CA 91320

(877) 944-4AGS

<http://www.appliedgeographic.com>

- Germany
- Greece
- Hong Kong
- Netherlands
- New Zealand
- Norway
- Republic of Ireland
- Spain
- Sweden
- United Kingdom
- United States

Within the coming months, the system will be extended to include –

- Austria
- Canada
- Czech Republic
- Italy
- Japan
- Switzerland

The Mosaic Global segments are:

- A Sophisticated Singles
- B Bourgeois Prosperity
- C Career and Family
- D Comfortable Retirement
- E Routine Service Workers
- F Hard Working Blue Collar
- G Metropolitan Strugglers
- H Low Income Elders
- I Post Industrial Survivors
- J Rural Inheritance

### **Methodology and Data Sources**

MOSAIC was originally constructed using the 1990 Census, and is now based on the 2000 Census and is updated on an annual basis using AGS demographic updates. In addition to the block group level segmentation, MOSAIC is available at the ZIP+4 level because of the analysis of Experian's household level records in conjunction with the block group assignments. The AGS estimates and projections are based in part on the same Experian household records, which provide a very accurate current demographic snapshot.

The MOSAIC system is documented more fully in separate handbook, methodology, and literature available from the AGS web site, <http://www.appliedgeographic.com>.

Note: Resellers must have a separate distribution agreement (as an attachment to their AGS Reseller agreement) with AGS in order to be licensed to resell these databases.

## ***Consumer Behavior***

### **INSOURCE (Experian) Profiles**

#### **Content**

The INSOURCE Profile database includes a range of direct marketing response and product ownership indices which include:

### **Applied Geographic Solutions, Inc.**

- Automobile ownership (average # vehicles, ownership type, retail sales price, current value, body style, and make)
- Direct marketing responsiveness (by type of program and channel)
- Contributions
- Product ownership (upscale products)
- Lifestyle

## Methodology

The INSOURCE Profile database is a MOSAIC profile database which has been built from several INSOURCE database components, including:

- INSOURCE MOR (Mail Order Response) variables
- Experian National Automobile Registration Database
- BehaviorBank

The MOR variables are compiled at the household level using information gathered on direct mail responsiveness to various types of campaigns and various modes of direct marketing.

The automobile registration database has been summarized to the block group level, and includes information on the number of vehicles, value, ownership type, and make/model.

BehaviorBank is a forty million household database that includes known responses to direct marketing campaigns, warranty registrations of products, and other data sources that clearly associate households with behavior.

Each of these databases were summarized by MOSAIC segment and indexed against the national average, yielding a series of index variables. In each case, a value of 100 indicates the national average.

## Simmons Profiles

### Content

Based on Simmons Market Research Bureau (SMRB) surveys, this consumer behavior database offers insight into the consumption patterns and preferences of consumers. A total of 2679 variables have been loaded from the Simmons survey. This is the latest 'doublebase' survey from 2005. Additional variables may be obtained from AGS, as Simmons has provided to us the Choices software which enables extraction of additional variables. The following general categories of information are provided:

Alcoholic Beverages

Apparel

Attitudes

- Fashion/Apparel
- Automobiles
- Food
- Finance
- General
- Health and dieting
- Internet
- Media
- Pharmaceuticals
- Product Placement
- Travel

- Technology

- Automotive
- Beverages
- Cable Television
- Collectibles
- Computers
- Contributions
- Demographics (Of Sample)
- Family Restaurants
- Fast Food Restaurants
- Financial
- Fitness and Sports
- Food
- Gambling
- Games and Toys
- Grocery Shopping
- Home Improvement
- Health and Medical
- Home Furnishings and Equipment
- Health Products
- Lawn and Garden
- Leisure
- Medical
- Media Quintiles
- Movies
- Music
- Parks
- Pets
- Radio Dayparts
- Reading
- Sports
- Telephone
- Theater
- Travel
- Television Dayparts

## Methodology and Data Sources

The Consumer Behavior database is derived from an analysis of the SMRB surveys using MOSAIC. The records in the SMRB survey are geocoded then assigned the MOSAIC code of the block group. The results are then summarized for each variable over the sixty segments, in effect providing the average value for each MOSAIC segment. For example, a variable such as "Visited Jack-In-The-Box" is computed by summarizing the records for each segment as a yes/no response, then finding the average percentage of households in each segment that went to Jack-In-The-Box. This is often referred to as a profile.

The profile is then applied to geographic areas by making the assumption that households in demographically similar neighborhoods will tend to have similar consumption patterns as a result of their similar economic means, life stage, and other characteristics. The result is a series of estimates for geographic areas which measure the relative propensity of consumers in each geographic area to eat at particular restaurants, own various household items, and engage in activities.

In most cases, these should be considered as relative indicators, since local differences may result in different behavior. In addition, in some cases, variables must be considered as potential only, since the activity or store may not be locally available.

The Consumer Behavior database is derived from an analysis of the SMRB surveys using MOSAIC. Each record in the SMRB survey is coded to a MOSAIC segment. The summarized profiles by MOSAIC segment are then used to derive indexes and penetrations that are applied to the block group level. The basic assumption is that people in demographically similar neighborhoods will tend to have similar consumption, ownership, and lifestyle preferences.

## *Assets, Debts, and Net Worth*

This database provides a unique glimpse into household finances – the type and amounts of household assets, the types and amounts of household debts, and the distribution of net worth by neighborhood. The data are derived from a statistical and geographic analysis of a Census Bureau survey known as the Survey of Consumer Finances.

As of 2007, this database has been completely remodeled, which results in some changes to the variable list – especially the net worth by value which has been eliminated as a result. The database no longer uses a median value as a result of difficulties in working with medians within geographic aggregations.

The Survey of Consumer Finances used is dated 2004. These data more accurately reflect the considerable rise in housing values and consumer debt over the past few years.

### **Content**

For assets (e.g. transaction accounts, life insurance with cash value, primary residence value) and debts (e.g. mortgage, credit card), separate tabulations are available for:

- Number of households with each asset or debt type
- Percentage of households with each asset or debt type
- Aggregate value of assets and debts
- Average value of assets and debts

A total of fifteen separate asset types are included, as well as six debt types.

With respect to net worth (effectively, the difference between a households assets and its debts) include:

- Average and aggregate neighborhood net worth

### **Data Sources and Methodology**

The primary source of this database is the 2004 Survey of Consumer Finances, issued by the Census Bureau.

The survey includes a wide range of demographic attributes related to “consuming units” (generally households), which have been modeled separately for each discrete expenditure category. The older surveys were first inflated to current price levels using the detailed consumer price index series. For each individual expenditure category in the survey, summary statistics were calculated for each separate element in the list below. In several cases, it was possible to utilize cross tabulation data (e.g. income by age of head of household). These variables are listed below:

- geographic region (Northeast, South, Midwest, West)
- metropolitan status (metropolitan, non-metropolitan) and size (e.g. > 4 million)
- housing tenure (owner or renter)
- age of head of household (<25 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, 65-74 years, and 75+ years)
- size of household (1 person, 2 persons, 3 persons, 4 persons, 5 persons, 6+ persons)
- household income (<5000, 5-10000, 10-15000, 15-20000, 20-30000, 30-40000, 40-50000, 50-70000, 70000+)
- race (White, Black, American Indian, Asian)
- number of vehicles (none, 1, 2+ vehicles per household)

For each item, a series of adjustment factors were derived for each unique demographic attribute. These adjustment factors were then applied to the block group level using the same demographic variables in order to create estimates at the local level,

which are consistent with local characteristics. Consistency checks were undertaken in order to ensure that the results at the block group level were consistent in the aggregate with overall published estimates.

## *Demographic Dimensions*

Demographic Dimensions is a modeling database at the block group and higher levels of geography that is useful in creating statistical models, site signature reports, and general executive summary information. Unlike discrete neighborhood classification systems, Demographic Dimensions provides continuous measurement scores across the dominant demographic components that differentiate neighborhoods.

Demographic Dimensions is based on the well-known data reduction tool of Principal Components Analysis, in which the common patterns found within a large number of variables are reduced to a core set of discriminating factors. By analyzing several hundred separate demographic variables at the block group level, sixteen dominant factors were identified. Together, these factors provide insights into the core dimensions of neighborhood differentiation.

The Dimensions database is normally provided as a set of continuous variables which are minimally auto correlated and have a mean of zero and unit variance. For graphic site signature charts, a consistent scale of 0 – 1000 is available.

Factors are useful in a broad spectrum of applications, including:

### Direct Marketing

Demographic Dimensions, when used in conjunction with MOSAIC and other targeting tools, can yield significant improvements in direct marketing results. By fine-tuning a MOSAIC profile, sub-groups of MOSAIC segments can be targeted effectively.

### Model Development

Dimensions are minimally correlated and are therefore very suitable for use in the construction of sales performance and site location models. Statistical models developed using Factors tend to be less prone to prediction error as a result of multicollinearity.

### Neighborhood Description

Factors can be used to effectively describe the dominant characteristics of neighborhoods for use in demographic reporting systems. Site “signatures” are easily defined and analyzed, since each of these factors is independent and reflect the dominant neighborhood differentiators.

### **Methodology**

Several hundred input variables were used in the analysis, which are summarized below by type of variable and source year. Note that in many cases, both average (or median) and distribution data were used (e.g. median age, % population age < 18, etc.). In most cases, with the exception of the housing characteristics tables, these were for estimates for 2003 rather than Census only.

### Geographic Characteristics

Urban core / urban fringe / rural Census classification  
Metropolitan status (e.g. metro, non-metropolitan area)

### Housing Characteristics

Units in structure (e.g. single family detached, apts 20+) Dwelling age  
Tenure  
Vacant dwellings by reason (e.g. seasonally vacant)  
Boarded up status (boarded up / not boarded up)  
Owner occupied dwellings by value  
Households by rent  
Dwellings by number of rooms  
Dwellings by heating type

Dwellings by water service and sewage service

### Household Characteristics

By type (family, non-family)

By size of household

By structure (e.g. married couple w children)

By age of householder

By length of residence (e.g. < 1 year, .... 10+ years)

### Population Characteristics

Recent and historical growth (1970-2000)

Projected growth (2000-2010)

Density

Age

Sex

Race

Hispanic origin

Detailed Hispanic Origin (e.g. Mexico, Puerto Rico)

Marital status

Highest level of education

Language spoken at home (% Spanish, % Asian)

School enrolment (public versus private)

Number of vehicles available

### Labor Force

Employment status (e.g. employed, unemployed)

Industry

Occupation

Employment of women with children (2000 only)

Unemployment rate

Travel time to work (2000 only)

Means of transportation to work (2000)

### Income

Sources of income (e.g. social security, wage and salary)

Households by income

Households by disposable income

Households by net worth

Households by income growth (1990-2000)

Households by income by age of householder

The SPSS principal components analysis module was used, with varimax rotation in order to maximize variable loading on each factor. Correlation between factors is minimal but non-zero in the resulting solution.

### Dimensions Variables

#### *01 Affluence*

Affluence is the single most important neighborhood discriminator and is most highly skewed. Affluence includes more than just income – it also reflects net worth, home ownership, and housing value and size.

#### *02 Family Status*

Family status, or household structure, is the second most important neighborhood differentiator. Ranging from areas populated with lone householders to married couple families with children, this factor varies most dramatically over the metropolitan scale.

#### *03 Occupational Status*

This factor measures the distinction between blue collar and white-collar occupations and lifestyles. Suburban, upscale neighborhoods of executives and professionals are contrasted with the blue-collar neighborhoods of smaller industrial towns and inner cities.

#### *04 Aging*

This important factor correlates highly with both the median age of residents and the percentage of residents over the age of 65. Residents in areas with high positive scores are most likely to be retired and receiving Social Security benefits, and often live alone. Residents in areas with high negative scores are likely to be young adults, often single, without children.

#### *05 African-American*

Areas with high scores consist of neighborhoods that are predominantly African-American. This factor tends to vary both at a metropolitan scale and regionally, with strong concentrations in the deep south and in the industrial cities of the northeast.

#### *06 Mexican-American*

The growth of the largely Mexican origin Hispanic population drives this increasingly important discriminating factor, which scores highest in the southwest states bordering Mexico.

#### *07 Housing Style*

This factor relates to the continuum of neighborhoods from single-family dwellings through dense high-rise apartment complexes.

#### *08 Agricultural Dominance*

Once the dominant discriminating factor of American life, the farm – non-farm dichotomy has been minimized with the wave of urbanization during the last century. High scores tend to occur in the generally rural states of the upper Great Plains and in the agricultural areas of Central California.

#### *09 College Campuses*

Areas with high scores on this factor are the distinctive neighborhoods on and around college campuses. These neighborhoods have a high percentage of young adults who have never been married, are enrolled in school, and may live in college dormitories.

#### *10 Growth and Stability*

Reflects the continuum between areas of rapid growth and change and stable, older neighborhoods. This factor highlights change areas both within metropolitan areas and at a national scale.

#### *11 Seasonal Areas*

Measuring the degree to which dwellings in the area are seasonally vacant, this factor is highest in the summer vacation areas of the Great Lakes and New England, the winter vacation areas of the Rocky Mountains, and on the non-urban coastlines of California and Florida.

#### *12 Native American*

Reflecting the distribution of Native Americans, this factor tends to be highest in the plains and southwest states, as well as Alaska.

#### *13 Asian-American*

Areas with high scores consist of neighborhoods that are predominantly Asian. Geographic variability is both at a metropolitan scale and regionally, with strong concentrations on the west coast and Hawaii.

#### *14 Institutional*

Areas scoring high on this factor are related to institutional land use – including both correctional facilities and long term care hospitals.

#### *15 Language Barriers*

Scores on this factor are high in areas where recent immigrants, often unable to speak English, have settled. Reflecting recent immigration trends, Spanish tends to be spoken in these neighborhoods.

#### *16 Military*

# Database Methodology Guide

2008 update

Released July 2008

Areas scoring high on this factor include both military bases and the nearby youthful and mobile neighborhoods that house military personnel.

## *2000 Census Database Releases*

Census 2000 is now fully released in terms of data that AGS intends to utilize. As of the 2003 release, SF1 and SF3 were fully incorporated into AGS data.

### **Census 2000 SF3 Balancing**

A substantial number of tables of the SF3 have been balanced to the SF1 totals, and several new tables have been extracted for this release. The following major tables have been balanced.

- Household Income
- Family Household Income
- Non-Family Household Income
- Housing Structure Type (units in structure)
- Housing Structure Type by Tenure
- Year Structure Built
- Year Structure Built by Tenure
- Year Moved In
- Year Moved in by Tenure
- Dwellings by Home Heating Fuel
- Dwellings by Number of Bedrooms
- Number of Bedrooms by Tenure
- Dwellings by Telephone Available
- Contract Rent
- Rent Asked for Vacant Units
- Gross Rent
- Housing Value, Selected Owner Occupied Units
- Housing Value, Vacant for Sale
- Housing Value, Owner Occupied Units
- Monthly Housing Costs, Rent as % of Income
- Monthly Mortgage Costs
- Monthly Housing Costs, Owner Occupied Housing with No Mortgage
- Vehicles Available
- Vehicles Available by Tenure
- Marital Status by Sex
- Educational Attainment by Sex
- Educational Enrollment by Sex
- Employment Status by Sex
- Number of Workers by Type of Family (Revised Format)
- Employment by Industry and Sex
- Employment by Occupation and Sex
- Transportation to Work: Travel Time
- Transportation to Work: Time Leaving for Work
- Transportation to Work: Mode of Transportation
- Disability Status (Additional Tables Added)

In addition, the databases have been reorganized into four rather than five physical databases. The SF1 database is now 00CEN1, and there are three SF3 databases labeled 00CEN2 through 00CEN4.

The process of balancing is used to overcome one of the basic shortcomings of the Census sampling frame, whereby it is possible to have discrepancies as high as 30% between SF1 and SF3 control totals (e.g. households, population), since

scaling factors for the sample are determined at the county level. In many cases, the result is confusion, since different tables which purport to be based on the same universe sum to different values.

The balancing process involves selecting the entire set of block groups for a county, and fitting this matrix to the county level control totals using a maximum entropy statistical technique. This has the desired result of forcing the relevant totals at the block group to match the expected values while maintaining the integrity of the summation at higher levels of geography.

## **The Census Sampling Issues (For Background Information only)**

As distributed, the Census data suffers from a fundamental problem that affects data at the block group level. The Census consists of two separate, but overlapping, questionnaires. The short form, which asks basic questions about the race and ethnicity of the population, yields a limited set of tables for a theoretical 100% count of population known as the SF1 tabulation. The long form, which includes many of the more interesting demographic questions related to income, is issued to approximately fifteen percent of the households. It includes the basic information on the short form, but in addition asks a wide range of additional questions of importance. The long form results in the SF3 tabulation set.

The Census Bureau normally makes full sampling adjustments to the SF3 tabulation at macro levels of geography in order to ensure that the SF1 and SF3 databases agree at these levels. At the block group level, however, adjustments are not undertaken. By way of example, if you were to sum the raw counts of households by income at the block group level from the SF3, you would not get the total household count shown in the SF1 tabulation. Likewise, if you sum the raw counts of vehicles available by household (e.g. none, 1, 2, 3, 4+), you would most often get a number which matches neither the SF1 count nor the SF3 households by income count.

The result is that each of the separate tabulations within the SF3 database normally requires the storage of a separate universe for each table, meaning that each tabulation (e.g. households by income, owner occupied dwellings by value) requires a separate universe, or total. The result is both a larger database and a significant potential for confusion. Most users of Census data expect that the "total households by income" will match the "total households by size of household" tabulation and so forth. Upon discovering that these are not the same "households", most users will immediately begin to question both the application that generated the data and the raw data itself.

The solution to this problem is to ensure that each tabulation within the SF3 sums, at the block group level, to the appropriate universe from the SF1 tabulations. The method by which AGS adjusts the SF3 counts to match the SF1 targets involves an entropy maximizing model which has the basic goal of adjusting the individual counts such that they sum to the desired totals in a way which minimizes the variance between the unadjusted and adjusted sets. The desired totals are obtained from the smallest parent unit for which the Census Bureau undertakes sampling adjustment and for which block groups are indivisible. In the 1990 Census, this was the county level.

By way of example, assume that we examine households by income at the block group level. If we sum the households by income class for any particular block group, we will not obtain the known households from the SF1 count in most cases. Likewise, if we take the block groups for a county and sum the household counts for any of the income classes, we will not get the total household count for that income group in the household (since this has been sample adjusted by the Census Bureau).

In effect, row and column sums will not add to the expected targets obtained from either the SF1 or from higher order geographic summaries. The goal, therefore, is to adjust the individual elements of this matrix such that the target totals are obtained and in such a way as to minimize the variation between the original matrix and the new matrix of values. This is a form of entropy maximization model, which fits the current matrix to the new totals with the minimum change possible to each individual cell of the matrix.

By so doing, we eliminate the need for both multiple bases (e.g. a separate base for total households, households by income, households by age of head of householder, and so forth) and in addition eliminate confusion on the part of the end users.

## **Historical Census Issues**

The data for small areas of geography are generally not comparable from one census to the next, as both administrative and statistical changes in the definitions of geographic areas are changed. Administrative changes typically include changes in

county or city definitions, which then are propagated down the base block group level. Statistical changes are more prevalent, as the Census Tract program is continually adjusted in order to maintain relatively equal population levels amongst the tracts.

In order to be able to use the historical Census data in any meaningful fashion, it is necessary to adjust the values for changes in boundaries over time. This is typically done by constructing a correspondence table between the base level of geography (block group) from one census to the next. The current 1970, 1980, and 1990 Census data provided by AGS has been standardized to the block group boundaries that were in effect at the time of the 1990 Census.

The means by which this is accomplished is to first create a correspondence for the three primary census base statistics – housing units, households, and population – at the block level. There are approximately seven million blocks for the 1990 Census and eight million for the 2000 Census. These are constructed cartographically using the TIGER cartography set from the Census Bureau. The two layers are then overlain in order to create a highly detailed “block-bit” set that contains well over ten million separate geographic areas. The 1990 Census base counts are first distributed from the block to block-bit level by removing any block bits that are known to be unpopulated (e.g. they are known to be non-residential in nature, such as airports, public buildings, and so on), after which the block bases are distributed to the block-bit level by an area apportionment. The resulting estimates are then re-aggregated to the 2000 Block and Block Group cartography, providing for each area a proportion of the 1990 housing units, households, and population to be allocated. These are then integer adjusted and balanced to higher order common geographies (e.g. the county) to ensure that the results are stable in preparation for further analysis.

At this point, the 1990 detailed tables are allocated into the 2000 geographic areas on the basis of the three bases (again, housing units, households, and population) on a table-by-table basis (e.g. households by income). The results are then integer adjusted and balanced to the corresponding county unit. In cases where the county unit has been changed, the results are balanced to the smallest aggregation of counties that has remained stable (e.g. the group of counties which were modified). The latter problem affects only a limited number of areas for the 2000 Census in Alaska (where the “county” is actually a statistical construct) and Virginia.

The resulting database is therefore an estimation of the distribution of the historical census data on the current geographic boundaries, which results in an ability to compare areas over time.

## **Census 2000 Geographic Areas**

The primary innovation of the 2000 Census is the use of “Collection” versus “Tabulation” blocks and block groups. One of the issues of the 1990 Census delineation was that many of the block groups were very small and not particularly useful, therefore, as geographic summary areas. This problem has been circumvented by creating for the 2000 Census two sets of geographies, one used for collection, the other for tabulation and presentation. The result is the somewhat surprising fact that the number of block groups has declined from 1990 to 2000, while the number of tracts has increased. However, there are far fewer small block groups and overall, the relative scale of the range of block group population is lower.

For the 2004 release, there were two major adjustments to the 2000 Census block and block group rosters. First, a new county 08014 Broomfield was carved out of four adjacent counties west of Denver. Second, the town of Clifton Forge VA was folded back into its parent county. In addition to these, a number of minor changes were made, largely corrective of the cartographic base, which have an effect on the rosters. In most cases, these changes have no impact on block group demographics.

The result, however, is that the number of block groups has been adjusted to 208,703.

The Metropolitan Areas have been completely revised for 2004 in accordance with the newest release by the Office of Management and Budget. MSA related codes are no longer 4 characters but are now 5 characters in length.

## Standard AGS Geographic Areas

BG	Block Group
CO	County
CS	County Subdivision
CB	County Based Metropolitan Areas (includes "Micropolitan" and Metropolitan areas) (2008)
MA	Metropolitan Statistical Areas (the "metro" not "micro" areas, formerly MS) (2008)
NC	New England City/Place Areas (formerly NE) (2008)
CA	Consolidated Metropolitan Areas (formerly CM) (2008)
PL	Place / Census Designated Place
ST	State
TR	Census Tract
US	United States
ZI	ZIP Codes (2nd Quarter 2008, TeleAtlas)
DM	Designated Marketing Areas

## *Geographic Definition Changes*

The metropolitan definitions are current to the latest release, OMB Bulletin 08-01, which is available at <http://www.whitehouse.gov/omb/bulletins/fy2008/b08-01.pdf>.

These changes are made on an annual basis as a result of a review of the annual Census Bureau place and county estimates. In most cases, the changes do not affect the geographic definitions, but rather only the order and list of city names included in the metropolitan area name.

### Metropolitan Statistical Areas

There have been some changes to the metropolitan areas, most minor, but including name changes which resulted in changes to the code assigned. For the 2008 release, we have updated the definitions for metropolitan and micropolitan areas in accordance with the latest OMB definitions.

These can be obtained from <http://www.whitehouse.gov/omb/bulletins/fy2008/b08-01.pdf>. Several changes have been made which will affect either record counts, names, or codes:

#### New Areas

43320 Show Low, AZ (Micropolitan Area) consisting of Navajo County, AZ

#### New Titles

- Atlantic City NJ now Atlantic City-Hammonton, NJ
- Summerville, SC added to the name of Charleston-North Charleston-Summerville, SC
- Winterhaven, FL added to the name of Lakeland-Winter Haven, FL
- Myrtle Beach-North Myrtle Beach-Conway, SC (order of secondary cities changed)
- Kennewick-Pasco-Richland, WA (order of secondary cities changed)

#### New Code

- Sarasota-Bradenton-Venice, FL is now Bradenton-Sarasota-Venice, FL and its code has been changed to 14600, as Sarasota is no longer considered the largest city. Please note that the CA (494 Sarasota-Bradenton-Punta Gorda) was not changed.
- Helena-West Helena, AR is now 25760. It previously was West Helena-Helena, AR
- Washington Court House, OH is 47920. It previously was simply Washington, OH

### Combined Statistical Areas

There were no changes for 2008.

The combined areas are defined by the OMB as:

"If specified criteria are met, adjacent Metropolitan and Micropolitan Statistical Areas, in various combinations, may become the components of a new set of complementary areas called Combined Statistical Areas. For instance, a Combined Statistical Area may comprise two or more Metropolitan Statistical Areas, a Metropolitan Statistical Area and a Micropolitan Statistical Area, two or more Micropolitan Statistical Areas, or multiple Metropolitan and Micropolitan Statistical Areas that have social and economic ties as measured by commuting, but at lower levels than are found among counties within Metropolitan and Micropolitan Statistical Areas."

The scale and population of these areas varies considerably, and a useful role for these areas in analysis is doubtful.

## *Geographic Area Reference*

Most of the information presented here is taken from the 2000 Census information published by the Census Bureau [www.census.gov](http://www.census.gov), and updated to include references to more recent documents which affect metropolitan area definitions.

### **Block Group**

A statistical subdivision of a census tract (or, prior to Census 2000, a block numbering area), consisting of a cluster of census blocks having the same first digit of their identifying numbers within that tract. For example, for Census 2000, BG 3 within a census tract includes all blocks numbered from 3000 to 3999. (A few BGs consist of a single block.) BGs generally contain between 300 and 3,000 people, with an optimum size of 1,500 people. The BG is the lowest-level geographic entity for which the U.S. Census Bureau tabulates sample data from a decennial census. See tribal block group.

In 2002, the Census Bureau created and released a slightly modified block group roster which included changes made to the geographical definitions post-census. AGS products utilize this revised 2002 roster.

### **Block**

An area bounded on all sides by visible and/or invisible features shown on a map prepared by the U.S. Census Bureau. A block is the smallest geographic entity for which the Census Bureau tabulates decennial census data. See block boundary, block number. For collecting information for Census 2000, each census block was identified uniquely within a county (or statistically equivalent entity) by a 4- or 5-digit number, which could be followed by an alphabetic suffix. All the collection blocks within a county had either four or five digits. For tabulating data for Census 2000, each census block is identified uniquely within a census tract by a 4-digit number. A 1990 census block number had three digits, with a potential alphabetic suffix.

### **County Subdivision: Census County Division (CCD)**

A statistical subdivision of a county, delineated by the U.S. Census Bureau in cooperation with state and local government officials for data presentation purposes. The Census Bureau has established CCDs in 21 states that do not have minor civil divisions suitable for data presentation; that is, minor civil divisions have not been legally established, do not have governmental or administrative purposes, have boundaries that are ambiguous or change frequently, and/or generally are not well known to the public.

### **Place: Census Designated Place (CDP)**

A geographic entity that serves as the statistical counterpart of an incorporated place for the purpose of presenting census data for an area with a concentration of population, housing, and commercial structures that is identifiable by name, but is not within an incorporated place. CDPs usually are delineated in cooperation with state, Puerto Rico, Island Area, local, and tribal officials based on U.S. Census Bureau guidelines. For Census 2000, for the first time, CDPs did not need to meet a minimum population threshold to qualify for tabulation of census data. See place.

### **Place: City**

A type of incorporated place in all states and the District of Columbia. In agreement with the state of Hawaii, the U.S. Census Bureau does not recognize the city of Honolulu for presentation of census data. In Virginia, all cities are not part of any county, and so the Census Bureau treats them as equivalent to a county for data presentation purposes, as well as treating them as places; there also is one such independent city in each of three states: Maryland, Missouri, and Nevada. In 23 states and the District of Columbia, some or all cities are not part of any minor civil division, in which case the Census Bureau also treats these entities as county subdivisions for data presentation purposes.

### **Place: Incorporated Place**

A type of governmental unit, incorporated under state law as a city, town (except in New England, New York, and Wisconsin), borough (except in Alaska and New York), or village, generally to provide a wide array of specific governmental services for a concentration of people within legally prescribed boundaries. New for Census 2000 are city and borough and municipality, which serve as both place- and county-level entities in Alaska. A few incorporated places do not have a legal description. See consolidated city, independent city, place.

### **Place: Independent City**

An incorporated place that is independent of i.e., not part of any county. All incorporated places classified as cities in Virginia are independent cities, as are Baltimore, Maryland; St. Louis, Missouri; and Carson City, Nevada. The U.S. Census Bureau

treats an independent city as equivalent to a county and county subdivision and as an incorporated place for data presentation purposes.

## **Census Tract**

A small, relatively permanent statistical subdivision of a county or statistically equivalent entity, delineated for data presentation purposes by a local group of census data users or the geographic staff of a regional census center in accordance with U.S. Census Bureau guidelines. Designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions at the time they are established, census tracts generally contain between 1,000 and 8,000 people, with an optimum size of 4,000 people. Census tract boundaries are delineated with the intention of being stable over many decades, so they generally follow relatively permanent visible features. However, they may follow governmental unit boundaries and other invisible features in some instances; the boundary of a state or county is always a census tract boundary. When data are provided for American Indian entities, the boundary of a federally recognized American Indian reservation and off-reservation trust land is always the boundary of a tribal census tract. See block numbering area, tribal census tract.

## **Consolidated City**

The U.S. Census Bureau refers to a governmental unit for which the functions of an incorporated place and its county or minor civil division have merged as a consolidated government. If one or more other incorporated places continue to function as separate governmental units within a consolidated government, the Census Bureau refers to the primary incorporated place as a consolidated city.

## **Consolidated Metro Area (CA)**

Refer to the section below "About Metropolitan and Micropolitan Statistical Areas"

## **County**

The primary legal division of every state except Alaska and Louisiana. A number of geographic entities are not legally designated as a county, but are recognized by the U.S. Census Bureau as equivalent to a county for data presentation purposes. These include the boroughs, city and boroughs, municipality, and census areas in Alaska; parishes in Louisiana; and cities that are independent of any county in Maryland, Missouri, Nevada, and Virginia. They also include the municipios in Puerto Rico, districts and islands in American Samoa, municipalities in the Northern Mariana Islands, and islands in the Virgin Islands of the United States. Because they contain no primary legal divisions, the Census Bureau treats the District of Columbia and Guam each as equivalent to a county (as well as equivalent to a state) for data presentation purposes. (A county is a minor civil division in American Samoa.)

## **County Subdivision**

The primary legal or statistical division of a county or statistically equivalent entity.

## **Metro Statistical Area (MA)**

Refer to the section below "About Metropolitan and Micropolitan Statistical Areas"

## **County Subdivision: (CS)**

The primary governmental or administrative division of a county or statistically equivalent entity in many states and statistically equivalent entities. The U.S. Census Bureau recognizes MCDs in 28 states, the District of Columbia, Puerto Rico, and the Island Areas. In 20 states and American Samoa, all or many MCDs are active general-purpose governmental units. See county subdivision, governmental unit, legal entity.

## **New England City/Town area (NC)**

Refer to the section below "About Metropolitan and Micropolitan Statistical Areas"

## **Place**

A concentration of population either legally bounded as an incorporated place, or delineated for statistical purposes as a census designated place (in Puerto Rico, a comunidad or zona urbana). See census designated place, consolidated city, incorporated place, independent city.

## **State**

A primary governmental division of the United States. The U.S. Census Bureau treats the District of Columbia as the equivalent of a state for data presentation purposes. It also treats a number of entities that are not legal divisions of the United

States as the equivalent of states for data presentation purposes (see Island Areas).

## United States

The 50 states and the District of Columbia.

## *About Metropolitan and Micropolitan Statistical Areas*

The United States Office of Management and Budget (OMB) defines metropolitan and micropolitan statistical areas according to published standards that are applied to Census Bureau data. The general concept of a metropolitan or micropolitan statistical area is that of a core area containing a substantial population nucleus, together with adjacent communities having a high degree of economic and social integration with that core. Currently defined metropolitan and micropolitan statistical areas are based on application of 2000 standards [ [PDF](#) | [Plain text](#) ] (which appeared in the *Federal Register* on December 27, 2000) to 2000 decennial census data. Current metropolitan and micropolitan statistical area definitions were announced by OMB effective December 2007.

Standard definitions of metropolitan areas were first issued in 1949 by the then Bureau of the Budget (predecessor of OMB), under the designation "standard metropolitan area" (SMA). The term was changed to "standard metropolitan statistical area" (SMSA) in 1959, and to "metropolitan statistical area" (MSA) in 1983. The term "metropolitan area" (MA) was adopted in 1990 and referred collectively to metropolitan statistical areas (MSAs), consolidated metropolitan statistical areas (CMSAs), and primary metropolitan statistical areas (PMSAs). The term "core based statistical area" (CBSA) became effective in 2000 and refers collectively to metropolitan and micropolitan statistical areas.

OMB has been responsible for the official metropolitan areas since they were first defined, except for the period 1977 to 1981, when they were the responsibility of the Office of Federal Statistical Policy and Standards, Department of Commerce. The standards for defining metropolitan areas were modified in 1958, 1971, 1975, 1980, 1990, and 2000.

## Defining Metropolitan and Micropolitan Statistical Areas

The 2000 standards provide that each CBSA must contain at least one urban area of 10,000 or more population. Each metropolitan statistical area must have at least one urbanized area of 50,000 or more inhabitants. Each micropolitan statistical area must have at least one urban cluster of at least 10,000 but less than 50,000 population.

Under the standards, the county (or counties) in which at least 50 percent of the population resides within urban areas of 10,000 or more population, or that contain at least 5,000 people residing within a single urban area of 10,000 or more population, is identified as a "central county" (counties). Additional "outlying counties" are included in the CBSA if they meet specified requirements of commuting to or from the central counties. Counties or equivalent entities form the geographic "building blocks" for metropolitan and micropolitan statistical areas throughout the United States and Puerto Rico.

If specified criteria are met, a metropolitan statistical area containing a single core with a population of 2.5 million or more may be subdivided to form smaller groupings of counties referred to as "metropolitan divisions."

As of June 6, 2000, there are 362 metropolitan statistical areas and 560 micropolitan statistical areas in the United States. In addition, there are 8 metropolitan statistical areas and 5 micropolitan statistical areas in Puerto Rico.

## Principal Cities and Metropolitan and Micropolitan Statistical Area Titles

The largest city in each metropolitan or micropolitan statistical area is designated a "principal city." Additional cities qualify if specified requirements are met concerning population size and employment. The title of each metropolitan or micropolitan statistical area consists of the names of up to three of its principal cities and the name of each state into which the metropolitan or micropolitan statistical area extends. Titles of metropolitan divisions also typically are based on principal city names but in certain cases consist of county names.

## Defining New England City and Town Areas

In view of the importance of cities and town in New England, the 2000 standards also provide for a set of geographic areas that are defined using cities and towns in the six New England states. The New England city and town areas (NECTAs) are defined using the same criteria as metropolitan and micropolitan statistical areas and are identified as either metropolitan or micropolitan, based, respectively, on the presence of either an urbanized area of 50,000 or more population or an urban cluster of at least 10,000 but less than 50,000 population. If the specified criteria are met, a NECTA containing a single core with a population of at least 2.5 million may be subdivided to form smaller groupings of cities and towns referred to as New England city and town area divisions.

## Changes in Definitions over Time

Changes in the definitions of these statistical areas since the 1950 census have consisted chiefly of:

- the recognition of new areas as they reached the minimum required city or urbanized area population, and
- the addition of counties (or cities and towns in New England) to existing areas as new decennial census data showed them to qualify.

In some instances, formerly separate areas have been merged, components of an area have been transferred from one area to another, or components have been dropped from an area. The large majority of changes have taken place on the basis of decennial census data. However, Census Bureau data serve as the basis for intercensal updates in specified circumstances.

Because of these historical changes in geographic definitions, users must be cautious in comparing data for these statistical areas from different dates. For some purposes, comparisons of data for areas as defined at given dates may be appropriate; for other purposes, it may be preferable to maintain consistent area definitions. Historical metropolitan area definitions are available for 1999, 1993, 1990, 1983, 1981, 1973, 1970, 1963, 1960, and 1950.

For more information, contact the Population Distribution Branch at (301) 763-2419.

*Source: U.S. Census Bureau, Population Division,  
Population Distribution Branch*