# Appendix A. Household Survey Data Expansion

## Table A.1 ACS Household by Number of Vehicle, Household Size and District

| District | Household Size | No vehicle available | 1 vehicle available | 2 vehicles available | 3 or more vehicles available | Total |
|---|---|---|---|---|---|---|
| 1 | 1-person household: | | | | 41,893 | 50,410 |
| | 2-person household: | | 14,004 | 27,197 | 11,892 | 55,490 |
| | 3-person household: | | | 16,900 | 8,053 | 26,132 |
| | 4-or-more-person | 12,979 | 5,698 | 16,132 | 13,219 | 35,935 |
| 2 | 1-person household: | | | | 4,874 | 5,680 |
| | 2-person household: | | 1,585 | 4,098 | 1,718 | 7,647 |
| | 3-person household: | | | 1,570 | 1,228 | 2,812 |
| | 4-or-more-person | | | 2,165 | 1,751 | 3,977 |
| 3 | 1-person household: | | | | 2,700 | 2,889 |
| | 2-person household: | | 791 | 2,219 | 857 | 3,955 |
| | 3-person household: | | | 1,509 | 705 | 2,319 |
| | 4-or-more-person | | | 825 | 894 | 1,782 |
| 4 | 1-person household: | 3,134 | 3,625 | | 854 | 4,998 |
| | 2-person household: | | 1,598 | 4,330 | 1,289 | 7,361 |
| | 3-person household: | | | 1,283 | 1,056 | 3,084 |
| | 4-or-more-person | | 1,218 | 2,323 | 1,666 | 4,614 |
| 6 | 1-person household: | | 4,139 | | 3,078 | 7,573 |
| | 2-person household: | | 1,672 | 6,936 | 5,263 | 14,097 |
| | 3-person household: | | | 2,639 | 2,811 | 5,533 |
| | 4-or-more-person | | | 3,942 | 4,047 | 8,071 |
| 5 | 1-person household: | | | | 4,015 | 5,045 |
| | 2-person household: | 1,383 | 1,589 | 2,429 | 507 | 4,862 |
| | 3-person household: | | | 1,741 | 469 | 2,226 |
| | 4-or-more-person | | | 1,264 | 358 | 1,622 |
| Total | | 17,496 | 89,022 | 100,397 | 61,199 | 268,114 |

## Table A.2 Household Samples by Number of Vehicle, Household Size and District

| District | Household Size | No vehicle available | 1 vehicle available | 2 vehicles available | 3 or more vehicles available | Total |
|---|---|---|---|---|---|---|
| 1 | 1-person household: | | | | 99 | 110 |
| | 2-person household: | | 24 | 66 | 33 | 126 |
| | 3-person household: | | | 24 | 13 | 38 |
| | 4-or-more-person | 16 | 6 | 28 | 24 | 59 |
| 2 | 1-person household: | | | | 96 | 102 |
| | 2-person household: | | 27 | 106 | 43 | 179 |
| | 3-person household: | | | 31 | 24 | 55 |
| | 4-or-more-person | | | 29 | 29 | 58 |
| 3 | 1-person household: | | | | 74 | 81 |
| | 2-person household: | | 19 | 70 | 28 | 118 |
| | 3-person household: | | | 17 | 12 | 29 |
| | 4-or-more-person | 23 | | 18 | 11 | 30 |
| 4 | 1-person household: | | 105 | | 25 | 131 |
| | 2-person household: | | 36 | 98 | 46 | 181 |
| | 3-person household: | | | 26 | 23 | 63 |
| | 4-or-more-person | | 24 | 40 | 30 | 80 |
| 6 | 1-person household: | | 62 | | 32 | 97 |
| | 2-person household: | | 25 | 103 | 95 | 223 |
| | 3-person household: | | | 27 | 36 | 63 |
| | 4-or-more-person | | | 32 | 47 | 79 |
| 5 | 1-person household: | | | | 143 | 174 |
| | 2-person household: | 38 | 37 | 94 | 28 | 164 |
| | 3-person household: | | | 35 | 13 | 49 |
| | 4-or-more-person | | | 32 | 12 | 45 |
| Total | | 77 | 1,022 | 688 | 547 | 2,334 |

## Table A.3 ACS Household by Number of Workers, Household Size and District

| District | Household Size | No workers | 1 worker | 2 or more workers | Total |
|---|---|---|---|---|---|
| 1 | 1-person household: | 24,187 | 26,223 | - | 50,410 |
| | 2-person household: | 15,037 | 20,734 | 19,719 | 55,490 |
| | 3-person household: | | | 12,062 | 26,132 |
| | 4-or-more-person | 9,320 | 24,677 | 18,926 | 35,935 |
| 2 | 1-person household: | 2,801 | 2,879 | - | 5,680 |
| | 2-person household: | 1,879 | 2,724 | 3,044 | 7,647 |
| | 3-person household: | | | 1,657 | 2,812 |
| | 4-or-more-person | | 2,197 | 2,437 | 3,977 |
| 3 | 1-person household: | 3,538 | 4,349 | - | 2,889 |
| | 2-person household: | 1,033 | 1,401 | 1,521 | 3,955 |
| | 3-person household: | | | | 2,319 |
| | 4-or-more-person | | 1,637 | 2,152 | 1,782 |
| 4 | 1-person household: | | | - | 4,998 |
| | 2-person household: | 2,022 | 2,690 | 2,649 | 7,361 |
| | 3-person household: | | | 1,597 | 3,084 |
| | 4-or-more-person | | 3,023 | 2,591 | 4,614 |
| 5 | 1-person household: | 2,449 | 2,596 | - | 5,045 |
| | 2-person household: | 1,030 | 1,535 | 2,297 | 4,862 |
| | 3-person household: | | | 1,237 | 2,226 |
| | 4-or-more-person | | 1,380 | 987 | 1,622 |
| 6 | 1-person household: | 3,900 | 3,673 | - | 7,573 |
| | 2-person household: | 4,132 | 4,866 | 5,099 | 14,097 |
| | 3-person household: | | | 2,898 | 5,533 |
| | 4-or-more-person | | 4,714 | 4,615 | 8,071 |
| Total | | 71,328 | 111,298 | 85,488 | 268,114 |

## Table A.4 Household Sample by Number of Workers, Household Size and District

| District | Household Size | No workers | 1 worker | 2 or more workers | Total |
|---|---|---|---|---|---|
| 1 | 1-person household: | 39 | 71 | - | 110 |
| | 2-person household: | 38 | 42 | 46 | 126 |
| | 3-person household: | | | 23 | 38 |
| | 4-or-more-person | 19 | 35 | 35 | 59 |
| 2 | 1-person household: | 53 | 49 | - | 102 |
| | 2-person household: | 68 | 46 | 65 | 179 |
| | 3-person household: | | | 39 | 55 |
| | 4-or-more-person | | 39 | 32 | 58 |
| 3 | 1-person household: | 40 | 111 | - | 81 |
| | 2-person household: | 37 | 36 | 45 | 118 |
| | 3-person household: | | | | 29 |
| | 4-or-more-person | - | 23 | 36 | 30 |
| 4 | 1-person household: | 61 | | - | 131 |
| | 2-person household: | 73 | 49 | 59 | 181 |
| | 3-person household: | | | 39 | 63 |
| | 4-or-more-person | | 54 | 43 | 80 |
| 5 | 1-person household: | 75 | 99 | - | 174 |
| | 2-person household: | 35 | 67 | 62 | 164 |
| | 3-person household: | | | 28 | 49 |
| | 4-or-more-person | | 38 | 25 | 45 |
| 6 | 1-person household: | 46 | 51 | - | 97 |
| | 2-person household: | 79 | 72 | 72 | 223 |
| | 3-person household: | | | 41 | 63 |
| | 4-or-more-person | | 52 | 47 | 79 |
| Total | | 663 | 936 | 735 | 2,334 |

## Table A.5 ACS and Survey Person Distributions for Employment Status by Age and District

| *ACS Distribution* | | Employment Status | | |
|---|---|---|---|---|
| District | Age | Employed | Unemployed | Total |
| | Under 5 years: | | 32,385 | 32,385 |
| | 5-15 years: | | 61,065 | 61,065 |
| | 16 -17 years: | | | 11,683 |
| | 18-24 years | 34,007 | 30,684 | 53,008 |
| | 25-34 years: | 48,621 | 18,669 | 67,290 |
| | 35-64 years: | 114,715 | 50,272 | 164,987 |
| 1 | 65+ years: | 5,963 | 44,143 | 50,106 |
| | Under 5 years: | | 3,093 | 3,093 |
| | 5-15 years: | | 6,695 | 6,695 |
| | 16 -17 years: | | | 1,207 |
| | 18-24 years | 2,985 | 2,773 | 4,551 |
| | 25-34 years: | 5,988 | 1,640 | 7,628 |
| | 35-64 years: | 14,409 | 4,867 | 19,276 |
| 2 | 65+ years: | 698 | 5,909 | 6,607 |
| | Under 5 years: | | 1,511 | 1,511 |
| | 5-15 years: | | 3,845 | 3,845 |
| | 16 -17 years: | | | 636 |
| | 18-24 years | 1,433 | 1,376 | 2,173 |
| | 25-34 years: | 2,948 | 797 | 3,745 |
| | 35-64 years: | 8,534 | 2,651 | 11,185 |
| 3 | 65+ years: | 489 | 2,915 | 3,404 |
| | Under 5 years: | | 3,311 | 3,311 |
| | 5-15 years: | | 8,219 | 8,219 |
| | 16 -17 years: | | | 1,645 |
| | 18-24 years | 2,721 | 2,762 | 3,838 |
| | 25-34 years: | 5,372 | 1,833 | 7,205 |
| | 35-64 years: | 15,414 | 4,618 | 20,032 |
| 4 | 65+ years: | 665 | 6,327 | 6,992 |
| | Under 5 years: | | 1,590 | 1,590 |
| | 5-15 years: | | 3,395 | 3,395 |
| | 16 -17 years: | | | 460 |
| | 18-24 years | 2,623 | 2,196 | 4,359 |
| | 25-34 years: | 3,993 | 730 | 4,723 |
| | 35-64 years: | 8,726 | 2,863 | 11,589 |
| 5 | 65+ years: | 471 | 3,041 | 3,512 |
| | Under 5 years: | | 5,508 | 5,508 |
| | 5-15 years: | | 12,981 | 12,981 |
| | 16 -17 years: | | | 3,114 |
| | 18-24 years | 4,457 | 6,594 | 7,937 |
| | 25-34 years: | 6,926 | 2,635 | 9,561 |
| | 35-64 years: | 29,401 | 12,120 | 41,521 |
| 6 | 65+ years: | 1,601 | 9,678 | 11,279 |
| Total | | 323,160 | 365,691 | 688,851 |

| *Sample Distribution* | | Employment Status | | |
|---|---|---|---|---|
| District | Age | Employed | Unemployed | Total |
| | Under 5 years: | | 46 | 46 |
| | 5-15 years: | | 79 | 79 |
| | 16 -17 years: | | | 15 |
| | 18-24 years | 16 | 28 | 29 |
| | 25-34 years: | 72 | 28 | 100 |
| | 35-64 years: | 261 | 92 | 353 |
| 1 | 65+ years: | 25 | 98 | 123 |
| | Under 5 years: | | 47 | 47 |
| | 5-15 years: | | 90 | 90 |
| | 16 -17 years: | | | 23 |
| | 18-24 years | 30 | 33 | 40 |
| | 25-34 years: | 87 | 23 | 110 |
| | 35-64 years: | 279 | 101 | 380 |
| 2 | 65+ years: | 28 | 167 | 195 |
| | Under 5 years: | | 23 | 23 |
| | 5-15 years: | | 50 | 50 |
| | 16 -17 years: | | | 12 |
| | 18-24 years | 23 | 22 | 33 |
| | 25-34 years: | 39 | 14 | 53 |
| | 35-64 years: | 187 | 60 | 247 |
| 3 | 65+ years: | 22 | 95 | 117 |
| | Under 5 years: | | 69 | 69 |
| | 5-15 years: | | 129 | 129 |
| | 16 -17 years: | | | 15 |
| | 18-24 years | 15 | 24 | 24 |
| | 25-34 years: | 87 | 31 | 118 |
| | 35-64 years: | 333 | 96 | 429 |
| 4 | 65+ years: | 33 | 204 | 237 |
| | Under 5 years: | | 49 | 49 |
| | 5-15 years: | | 86 | 86 |
| | 16 -17 years: | | | 8 |
| | 18-24 years | 20 | 26 | 38 |
| | 25-34 years: | 109 | 28 | 137 |
| | 35-64 years: | 285 | 104 | 389 |
| 5 | 65+ years: | 25 | 112 | 137 |
| | Under 5 years: | | 47 | 47 |
| | 5-15 years: | | 133 | 133 |
| | 16 -17 years: | | | 20 |
| | 18-24 years | 34 | 38 | 52 |
| | 25-34 years: | 56 | 27 | 83 |
| | 35-64 years: | 396 | 153 | 549 |
| 6 | 65+ years: | 34 | 171 | 205 |
| Total | | 2,496 | 2,623 | 5,119 |

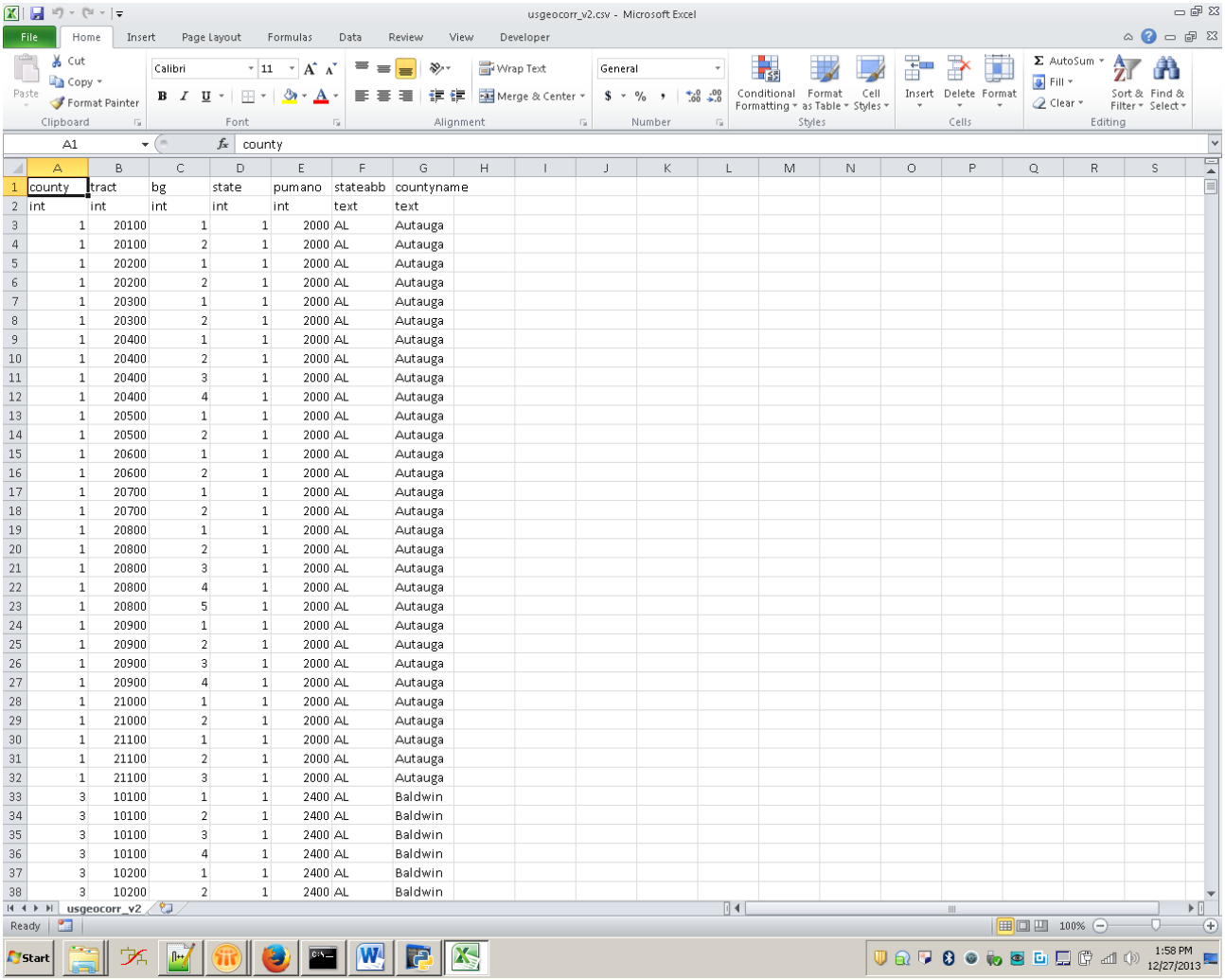
# Appendix B. Population Synthesis Results and Documentation

## B.1 An Example for a Sample PopGen Setup and Run

**Step 1: Prepare Input Files**

There are 5 input files.

1) Geographic correspondence file (shown below). Connects county, tract, blockgroup (TAZ), state, pumano, stateabb and countyname. This needs to be prepared first.

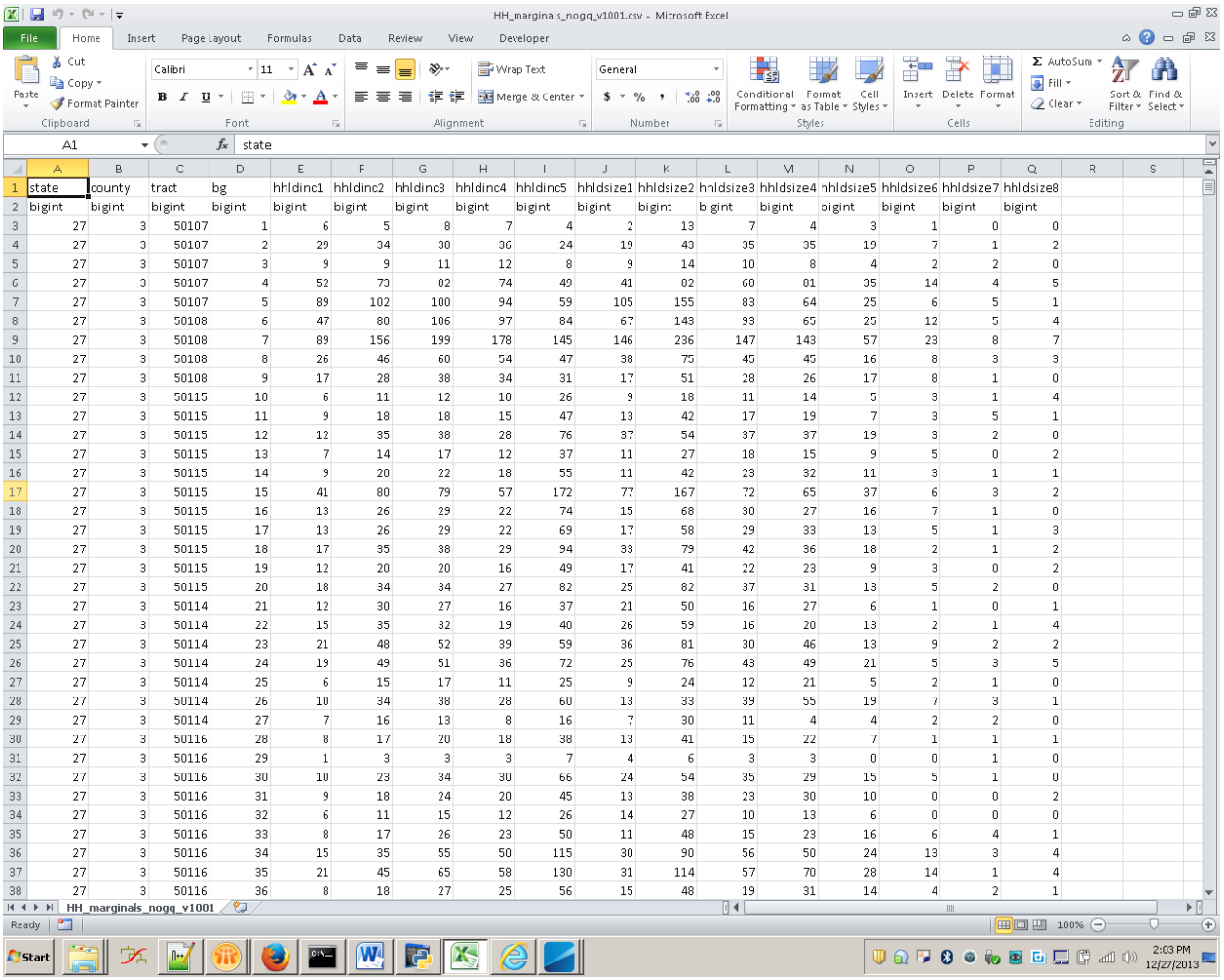


| county | tract | bg | state | pumano | stateabb | countyname |
|---|---|---|---|---|---|---|
| int | int | int | int | int | text | text |
| 1 | 20100 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20100 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20200 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20200 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20300 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20300 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20400 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20400 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20400 | 3 | 1 | 2000 | AL | Autauga |
| 1 | 20400 | 4 | 1 | 2000 | AL | Autauga |
| 1 | 20500 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20500 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20600 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20600 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20700 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20700 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20800 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20800 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20800 | 3 | 1 | 2000 | AL | Autauga |
| 1 | 20800 | 4 | 1 | 2000 | AL | Autauga |
| 1 | 20800 | 5 | 1 | 2000 | AL | Autauga |
| 1 | 20900 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 20900 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 20900 | 3 | 1 | 2000 | AL | Autauga |
| 1 | 20900 | 4 | 1 | 2000 | AL | Autauga |
| 1 | 21000 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 21000 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 21100 | 1 | 1 | 2000 | AL | Autauga |
| 1 | 21100 | 2 | 1 | 2000 | AL | Autauga |
| 1 | 21100 | 3 | 1 | 2000 | AL | Autauga |
| 3 | 10100 | 1 | 1 | 2400 | AL | Baldwin |
| 3 | 10100 | 2 | 1 | 2400 | AL | Baldwin |
| 3 | 10100 | 3 | 1 | 2400 | AL | Baldwin |
| 3 | 10100 | 4 | 1 | 2400 | AL | Baldwin |
| 3 | 10200 | 1 | 1 | 2400 | AL | Baldwin |
| 3 | 10200 | 2 | 1 | 2400 | AL | Baldwin |

2) Household marginal file. This contains total households by blockgroup (TAZ), household income categories, and household size.

Household income is provided with 4 income categories this is converted into 5 income categories by using the 2007-2011 ACS summaries at a county level.

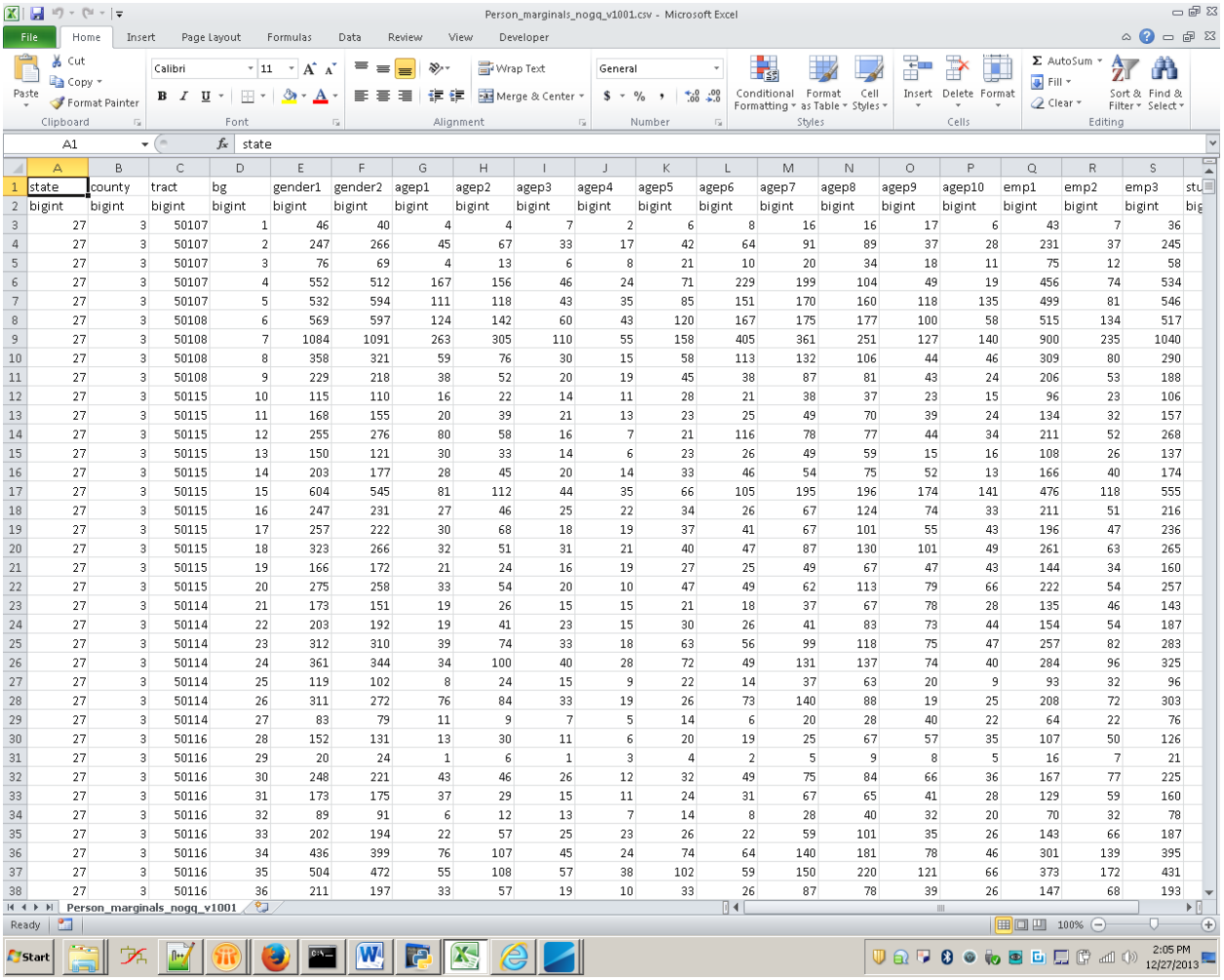
Check for consistency in totals i.e., sum of households by income categories should equal sum of households by household size. All number should be integers.



| state | county | tract | bg | hhldinc1 | hhldinc2 | hhldinc3 | hhldinc4 | hhldinc5 | hhldsize1 | hhldsize2 | hhldsize3 | hhldsize4 | hhldsize5 | hhldsize6 | hhldsize7 | hhldsize8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint |
| 27 | 3 | 50107 | 1 | 6 | 5 | 8 | 7 | 4 | 2 | 13 | 7 | 4 | 3 | 1 | 0 | 0 |
| 27 | 3 | 50107 | 2 | 29 | 34 | 38 | 36 | 24 | 19 | 43 | 35 | 35 | 19 | 7 | 1 | 2 |
| 27 | 3 | 50107 | 3 | 9 | 9 | 11 | 12 | 8 | 9 | 14 | 10 | 8 | 4 | 2 | 2 | 0 |
| 27 | 3 | 50107 | 4 | 52 | 73 | 82 | 74 | 49 | 41 | 82 | 68 | 81 | 35 | 14 | 4 | 5 |
| 27 | 3 | 50107 | 5 | 89 | 102 | 100 | 94 | 59 | 105 | 155 | 83 | 64 | 25 | 6 | 5 | 1 |
| 27 | 3 | 50108 | 6 | 47 | 80 | 106 | 97 | 84 | 67 | 143 | 93 | 65 | 25 | 12 | 5 | 4 |
| 27 | 3 | 50108 | 7 | 89 | 156 | 199 | 178 | 145 | 146 | 236 | 147 | 143 | 57 | 23 | 8 | 7 |
| 27 | 3 | 50108 | 8 | 26 | 46 | 60 | 54 | 47 | 38 | 75 | 45 | 45 | 16 | 8 | 3 | 3 |
| 27 | 3 | 50108 | 9 | 17 | 28 | 38 | 34 | 31 | 17 | 51 | 28 | 26 | 17 | 8 | 1 | 0 |
| 27 | 3 | 50115 | 10 | 6 | 11 | 12 | 10 | 26 | 9 | 18 | 11 | 14 | 5 | 3 | 1 | 4 |
| 27 | 3 | 50115 | 11 | 9 | 18 | 18 | 15 | 47 | 13 | 42 | 17 | 19 | 7 | 3 | 5 | 1 |
| 27 | 3 | 50115 | 12 | 12 | 35 | 38 | 28 | 76 | 37 | 54 | 37 | 37 | 19 | 3 | 2 | 0 |
| 27 | 3 | 50115 | 13 | 7 | 14 | 17 | 12 | 37 | 11 | 27 | 18 | 15 | 9 | 5 | 0 | 2 |
| 27 | 3 | 50115 | 14 | 9 | 20 | 22 | 18 | 55 | 11 | 42 | 23 | 32 | 11 | 3 | 1 | 1 |
| 27 | 3 | 50115 | 15 | 41 | 80 | 79 | 57 | 172 | 77 | 167 | 72 | 65 | 37 | 6 | 3 | 2 |
| 27 | 3 | 50115 | 16 | 13 | 26 | 29 | 22 | 74 | 15 | 68 | 30 | 27 | 16 | 7 | 1 | 0 |
| 27 | 3 | 50115 | 17 | 13 | 26 | 29 | 22 | 69 | 17 | 58 | 29 | 33 | 13 | 5 | 1 | 3 |
| 27 | 3 | 50115 | 18 | 17 | 35 | 38 | 29 | 94 | 33 | 79 | 42 | 36 | 18 | 2 | 1 | 2 |
| 27 | 3 | 50115 | 19 | 12 | 20 | 20 | 16 | 49 | 17 | 41 | 22 | 23 | 9 | 3 | 0 | 2 |
| 27 | 3 | 50115 | 20 | 18 | 34 | 34 | 27 | 82 | 25 | 82 | 37 | 31 | 13 | 5 | 2 | 0 |
| 27 | 3 | 50114 | 21 | 12 | 30 | 27 | 16 | 37 | 21 | 50 | 16 | 27 | 6 | 1 | 0 | 1 |
| 27 | 3 | 50114 | 22 | 15 | 35 | 32 | 19 | 40 | 26 | 59 | 16 | 20 | 13 | 2 | 1 | 4 |
| 27 | 3 | 50114 | 23 | 21 | 48 | 52 | 39 | 59 | 36 | 81 | 30 | 46 | 13 | 9 | 2 | 2 |
| 27 | 3 | 50114 | 24 | 19 | 49 | 51 | 36 | 72 | 25 | 76 | 43 | 49 | 21 | 5 | 3 | 5 |
| 27 | 3 | 50114 | 25 | 6 | 15 | 17 | 11 | 25 | 9 | 24 | 12 | 21 | 5 | 2 | 1 | 0 |
| 27 | 3 | 50114 | 26 | 10 | 34 | 38 | 28 | 60 | 13 | 33 | 39 | 55 | 19 | 7 | 3 | 1 |
| 27 | 3 | 50114 | 27 | 7 | 16 | 13 | 8 | 16 | 7 | 30 | 11 | 4 | 4 | 2 | 2 | 0 |
| 27 | 3 | 50116 | 28 | 8 | 17 | 20 | 18 | 38 | 13 | 41 | 15 | 22 | 7 | 1 | 1 | 1 |
| 27 | 3 | 50116 | 29 | 1 | 3 | 3 | 3 | 7 | 4 | 6 | 3 | 3 | 0 | 0 | 1 | 0 |
| 27 | 3 | 50116 | 30 | 10 | 23 | 34 | 30 | 66 | 24 | 54 | 35 | 29 | 15 | 5 | 1 | 0 |
| 27 | 3 | 50116 | 31 | 9 | 18 | 24 | 20 | 45 | 13 | 38 | 23 | 30 | 10 | 0 | 0 | 2 |
| 27 | 3 | 50116 | 32 | 6 | 11 | 15 | 12 | 26 | 14 | 27 | 10 | 13 | 6 | 0 | 0 | 0 |
| 27 | 3 | 50116 | 33 | 8 | 17 | 26 | 23 | 50 | 11 | 48 | 15 | 23 | 16 | 6 | 4 | 1 |
| 27 | 3 | 50116 | 34 | 15 | 35 | 55 | 50 | 115 | 30 | 90 | 56 | 50 | 24 | 13 | 3 | 4 |
| 27 | 3 | 50116 | 35 | 21 | 45 | 65 | 58 | 130 | 31 | 114 | 57 | 70 | 28 | 14 | 1 | 4 |
| 27 | 3 | 50116 | 36 | 8 | 18 | 27 | 25 | 56 | 15 | 48 | 19 | 31 | 14 | 4 | 2 | 1 |

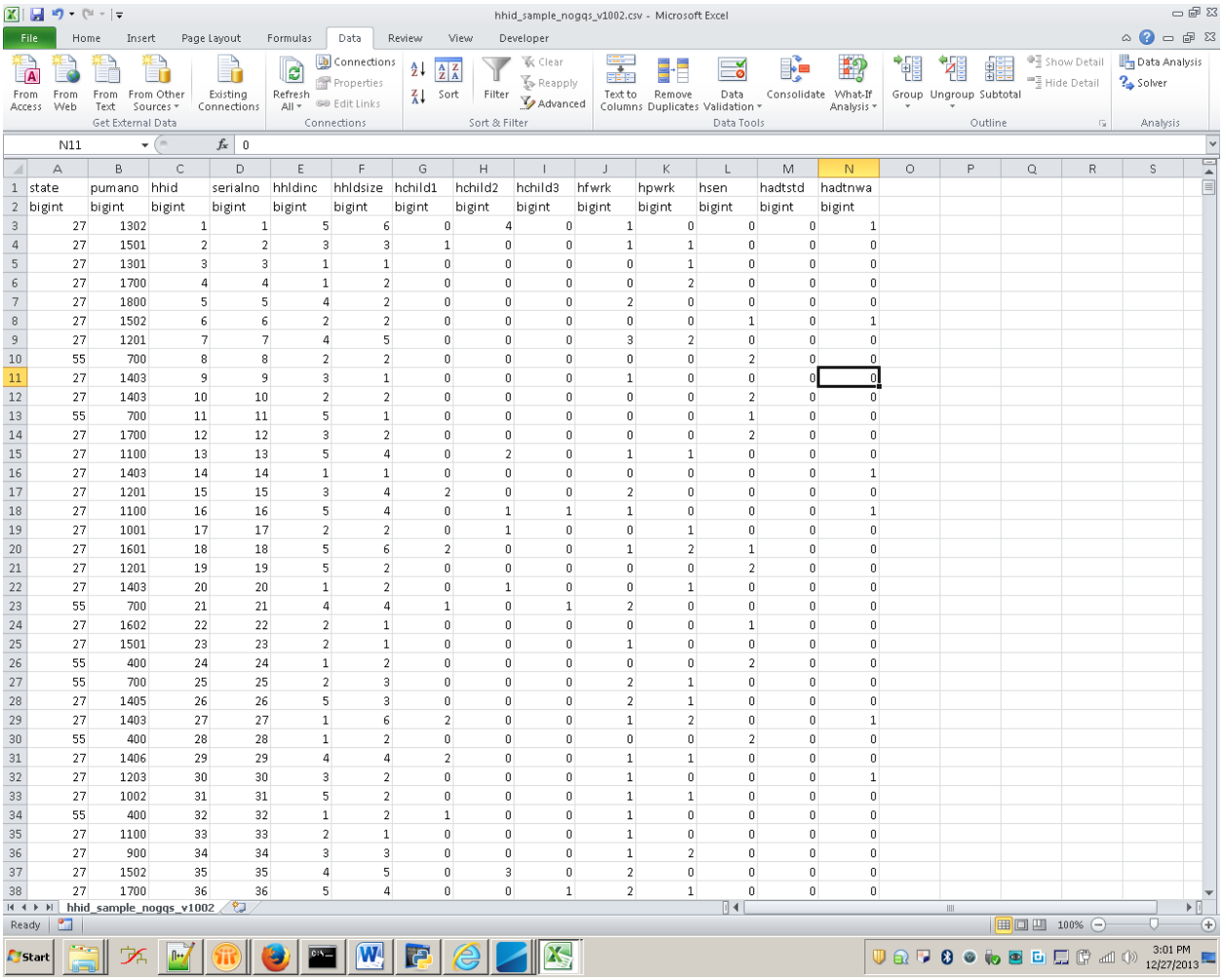
3) Person marginal file. This contains total households by blockgroup (TAZ), gender, age, employment status, and student status. The number of categories for each control variable is important.

Check for consistency in totals i.e., sum of persons by gender should equal sum of persons by employment status. All number should be integers.



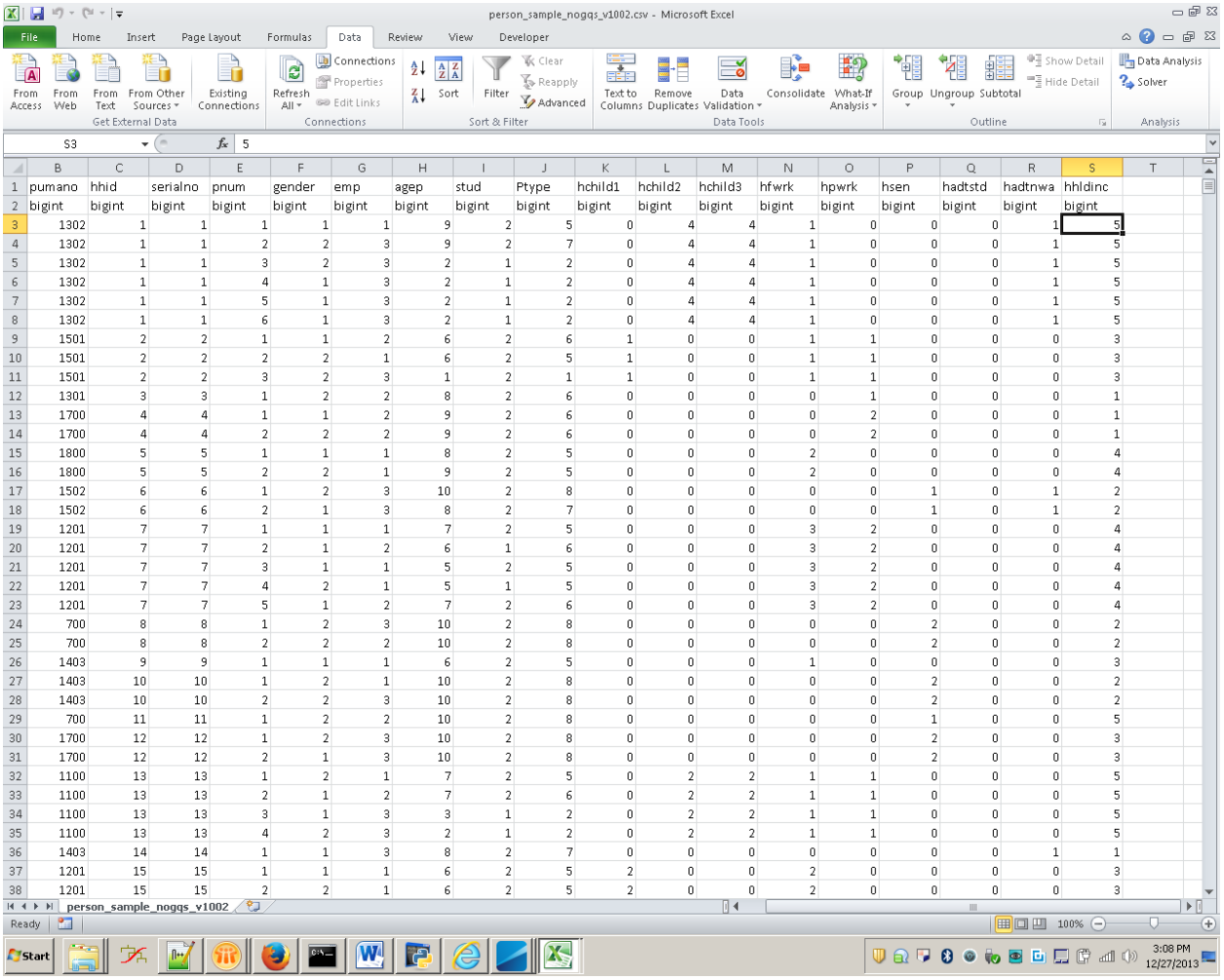
| state | county | tract | bg | gender1 | gender2 | agep1 | agep2 | agep3 | agep4 | agep5 | agep6 | agep7 | agep8 | agep9 | agep10 | emp1 | emp2 | emp3 | stu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | big |
| 27 | 3 | 50107 | 1 | 46 | 40 | 4 | 4 | 7 | 2 | 6 | 8 | 16 | 16 | 17 | 6 | 43 | 7 | 36 | |
| 27 | 3 | 50107 | 2 | 247 | 266 | 45 | 67 | 33 | 17 | 42 | 64 | 91 | 89 | 37 | 28 | 231 | 37 | 245 | |
| 27 | 3 | 50107 | 3 | 76 | 69 | 4 | 13 | 6 | 8 | 21 | 10 | 20 | 34 | 18 | 11 | 75 | 12 | 58 | |
| 27 | 3 | 50107 | 4 | 552 | 512 | 167 | 156 | 46 | 24 | 71 | 229 | 199 | 104 | 49 | 19 | 456 | 74 | 534 | |
| 27 | 3 | 50107 | 5 | 532 | 594 | 111 | 118 | 43 | 35 | 85 | 151 | 170 | 160 | 118 | 135 | 499 | 81 | 546 | |
| 27 | 3 | 50108 | 6 | 569 | 597 | 124 | 142 | 60 | 43 | 120 | 167 | 175 | 177 | 100 | 58 | 515 | 134 | 517 | |
| 27 | 3 | 50108 | 7 | 1084 | 1091 | 263 | 305 | 110 | 55 | 158 | 405 | 361 | 251 | 127 | 140 | 900 | 235 | 1040 | |
| 27 | 3 | 50108 | 8 | 358 | 321 | 59 | 76 | 30 | 15 | 58 | 113 | 132 | 106 | 44 | 46 | 309 | 80 | 290 | |
| 27 | 3 | 50108 | 9 | 229 | 218 | 38 | 52 | 20 | 19 | 45 | 38 | 87 | 81 | 43 | 24 | 206 | 53 | 188 | |
| 27 | 3 | 50115 | 10 | 115 | 110 | 16 | 22 | 14 | 11 | 28 | 21 | 38 | 37 | 23 | 15 | 96 | 23 | 106 | |
| 27 | 3 | 50115 | 11 | 168 | 155 | 20 | 39 | 21 | 13 | 23 | 25 | 49 | 70 | 39 | 24 | 134 | 32 | 157 | |
| 27 | 3 | 50115 | 12 | 255 | 276 | 80 | 58 | 16 | 7 | 21 | 116 | 78 | 77 | 44 | 34 | 211 | 52 | 268 | |
| 27 | 3 | 50115 | 13 | 150 | 121 | 30 | 33 | 14 | 6 | 23 | 26 | 49 | 59 | 15 | 16 | 108 | 26 | 137 | |
| 27 | 3 | 50115 | 14 | 203 | 177 | 28 | 45 | 20 | 14 | 33 | 46 | 54 | 75 | 52 | 13 | 166 | 40 | 174 | |
| 27 | 3 | 50115 | 15 | 604 | 545 | 81 | 112 | 44 | 35 | 66 | 105 | 195 | 196 | 174 | 141 | 476 | 118 | 555 | |
| 27 | 3 | 50115 | 16 | 247 | 231 | 27 | 46 | 25 | 22 | 34 | 26 | 67 | 124 | 74 | 33 | 211 | 51 | 216 | |
| 27 | 3 | 50115 | 17 | 257 | 222 | 30 | 68 | 18 | 19 | 37 | 41 | 67 | 101 | 55 | 43 | 196 | 47 | 236 | |
| 27 | 3 | 50115 | 18 | 323 | 266 | 32 | 51 | 31 | 21 | 40 | 47 | 87 | 130 | 101 | 49 | 261 | 63 | 265 | |
| 27 | 3 | 50115 | 19 | 166 | 172 | 21 | 24 | 16 | 19 | 27 | 25 | 49 | 67 | 47 | 43 | 144 | 34 | 160 | |
| 27 | 3 | 50115 | 20 | 275 | 258 | 33 | 54 | 20 | 10 | 47 | 49 | 62 | 113 | 79 | 66 | 222 | 54 | 257 | |
| 27 | 3 | 50114 | 21 | 173 | 151 | 19 | 26 | 15 | 15 | 21 | 18 | 37 | 67 | 78 | 28 | 135 | 46 | 143 | |
| 27 | 3 | 50114 | 22 | 203 | 192 | 19 | 41 | 23 | 15 | 30 | 26 | 41 | 83 | 73 | 44 | 154 | 54 | 187 | |
| 27 | 3 | 50114 | 23 | 312 | 310 | 39 | 74 | 33 | 18 | 63 | 56 | 99 | 118 | 75 | 47 | 257 | 82 | 283 | |
| 27 | 3 | 50114 | 24 | 361 | 344 | 34 | 100 | 40 | 28 | 72 | 49 | 131 | 137 | 74 | 40 | 284 | 96 | 325 | |
| 27 | 3 | 50114 | 25 | 119 | 102 | 8 | 24 | 15 | 9 | 22 | 14 | 37 | 63 | 20 | 9 | 93 | 32 | 96 | |
| 27 | 3 | 50114 | 26 | 311 | 272 | 76 | 84 | 33 | 19 | 26 | 73 | 140 | 88 | 19 | 25 | 208 | 72 | 303 | |
| 27 | 3 | 50114 | 27 | 83 | 79 | 11 | 9 | 7 | 5 | 14 | 6 | 20 | 28 | 40 | 22 | 64 | 22 | 76 | |
| 27 | 3 | 50116 | 28 | 152 | 131 | 13 | 30 | 11 | 6 | 20 | 19 | 25 | 67 | 57 | 35 | 107 | 50 | 126 | |
| 27 | 3 | 50116 | 29 | 20 | 24 | 1 | 6 | 1 | 3 | 4 | 2 | 5 | 9 | 8 | 5 | 16 | 7 | 21 | |
| 27 | 3 | 50116 | 30 | 248 | 221 | 43 | 46 | 26 | 12 | 32 | 49 | 75 | 84 | 66 | 36 | 167 | 77 | 225 | |
| 27 | 3 | 50116 | 31 | 173 | 175 | 37 | 29 | 15 | 11 | 24 | 31 | 67 | 65 | 41 | 28 | 129 | 59 | 160 | |
| 27 | 3 | 50116 | 32 | 89 | 91 | 6 | 12 | 13 | 7 | 14 | 8 | 28 | 40 | 32 | 20 | 70 | 32 | 78 | |
| 27 | 3 | 50116 | 33 | 202 | 194 | 22 | 57 | 25 | 23 | 26 | 22 | 59 | 101 | 35 | 26 | 143 | 66 | 187 | |
| 27 | 3 | 50116 | 34 | 436 | 399 | 76 | 107 | 45 | 24 | 74 | 64 | 140 | 181 | 78 | 46 | 301 | 139 | 395 | |
| 27 | 3 | 50116 | 35 | 504 | 472 | 55 | 108 | 57 | 38 | 102 | 59 | 150 | 220 | 121 | 66 | 373 | 172 | 431 | |
| 27 | 3 | 50116 | 36 | 211 | 197 | 33 | 57 | 19 | 10 | 33 | 26 | 87 | 78 | 39 | 26 | 147 | 68 | 193 | |

4) household sample file. This contains individual's households by state, pumano, hhid, household income, household size and other household level variables. This file is built using the 2005-2009 PUMS data.



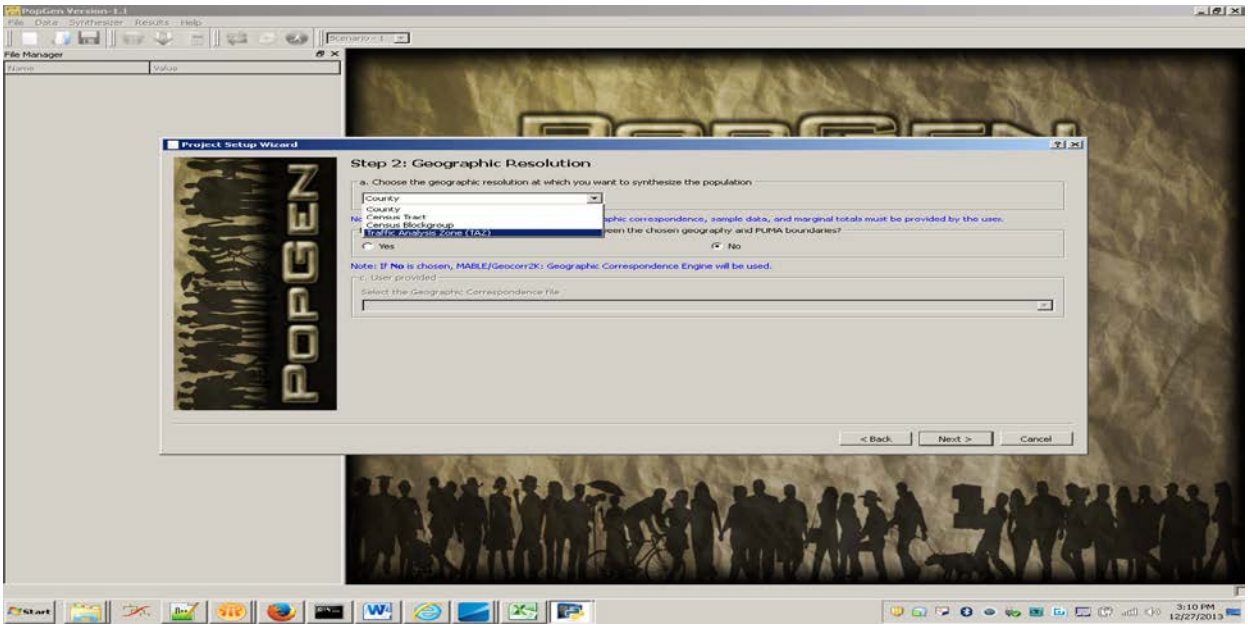

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | state | pumano | hhid | serialno | hhldinc | hhldsize | hchild1 | hchild2 | hchild3 | hfwrk | hpwrk | hsen | hadtstd | hadtnwa |
| 2 | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint |
| 3 | 27 | 1302 | 1 | 1 | 5 | 6 | 0 | 4 | 0 | 1 | 0 | 0 | 0 | 1 |
| 4 | 27 | 1501 | 2 | 2 | 3 | 3 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 5 | 27 | 1301 | 3 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 6 | 27 | 1700 | 4 | 4 | 1 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 7 | 27 | 1800 | 5 | 5 | 4 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 8 | 27 | 1502 | 6 | 6 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 9 | 27 | 1201 | 7 | 7 | 4 | 5 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 |
| 10 | 55 | 700 | 8 | 8 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 11 | 27 | 1403 | 9 | 9 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 12 | 27 | 1403 | 10 | 10 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 13 | 55 | 700 | 11 | 11 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 14 | 27 | 1700 | 12 | 12 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 15 | 27 | 1100 | 13 | 13 | 5 | 4 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 |
| 16 | 27 | 1403 | 14 | 14 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 17 | 27 | 1201 | 15 | 15 | 3 | 4 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 18 | 27 | 1100 | 16 | 16 | 5 | 4 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| 19 | 27 | 1001 | 17 | 17 | 2 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 20 | 27 | 1601 | 18 | 18 | 5 | 6 | 2 | 0 | 0 | 1 | 2 | 1 | 0 | 0 |
| 21 | 27 | 1201 | 19 | 19 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 22 | 27 | 1403 | 20 | 20 | 1 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 23 | 55 | 700 | 21 | 21 | 4 | 4 | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 0 |
| 24 | 27 | 1602 | 22 | 22 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 25 | 27 | 1501 | 23 | 23 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 26 | 55 | 400 | 24 | 24 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 27 | 55 | 700 | 25 | 25 | 2 | 3 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 |
| 28 | 27 | 1405 | 26 | 26 | 5 | 3 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 |
| 29 | 27 | 1403 | 27 | 27 | 1 | 6 | 2 | 0 | 0 | 1 | 2 | 0 | 0 | 1 |
| 30 | 55 | 400 | 28 | 28 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 31 | 27 | 1406 | 29 | 29 | 4 | 4 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 32 | 27 | 1203 | 30 | 30 | 3 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 33 | 27 | 1002 | 31 | 31 | 5 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 34 | 55 | 400 | 32 | 32 | 1 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 35 | 27 | 1100 | 33 | 33 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 36 | 27 | 900 | 34 | 34 | 3 | 3 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 |
| 37 | 27 | 1502 | 35 | 35 | 4 | 5 | 0 | 3 | 0 | 2 | 0 | 0 | 0 | 0 |
| 38 | 27 | 1700 | 36 | 36 | 5 | 4 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 |

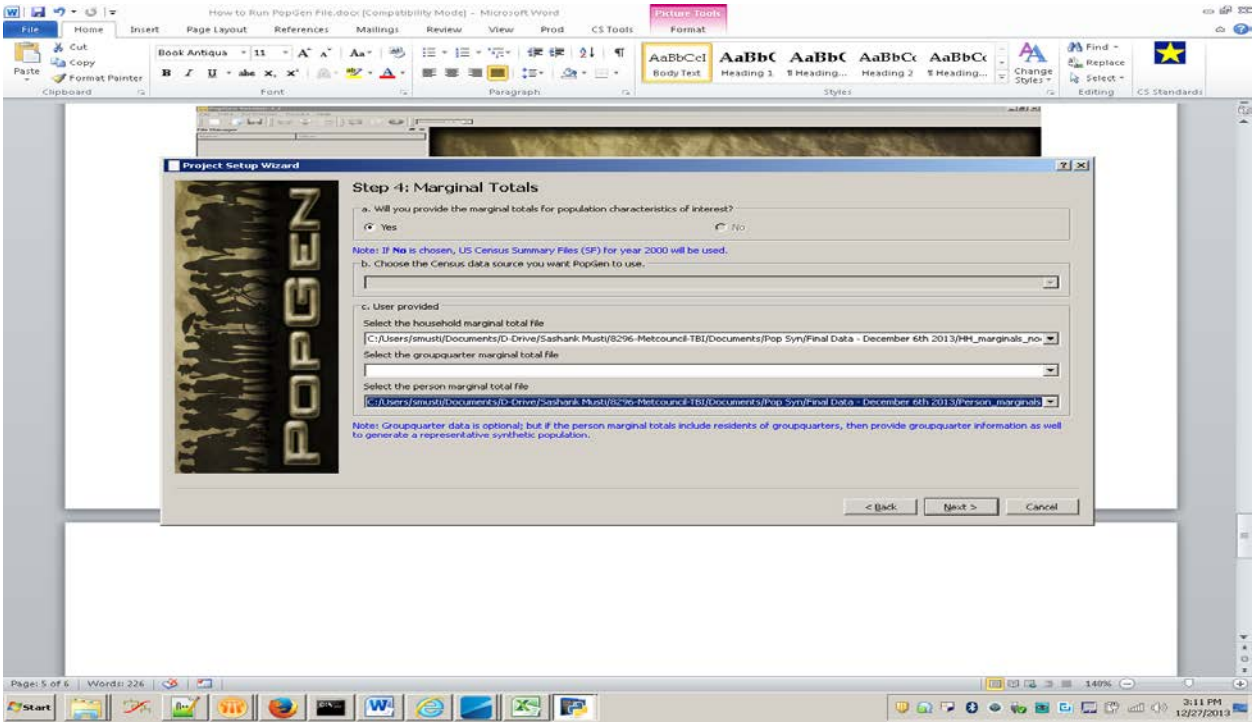
5) Person sample file. This contains individual persons in a household by state, pumano, hhid, gender, age, student status, employment status, household income and other household level variables. This file is built using the 2005-2009 PUMS data.



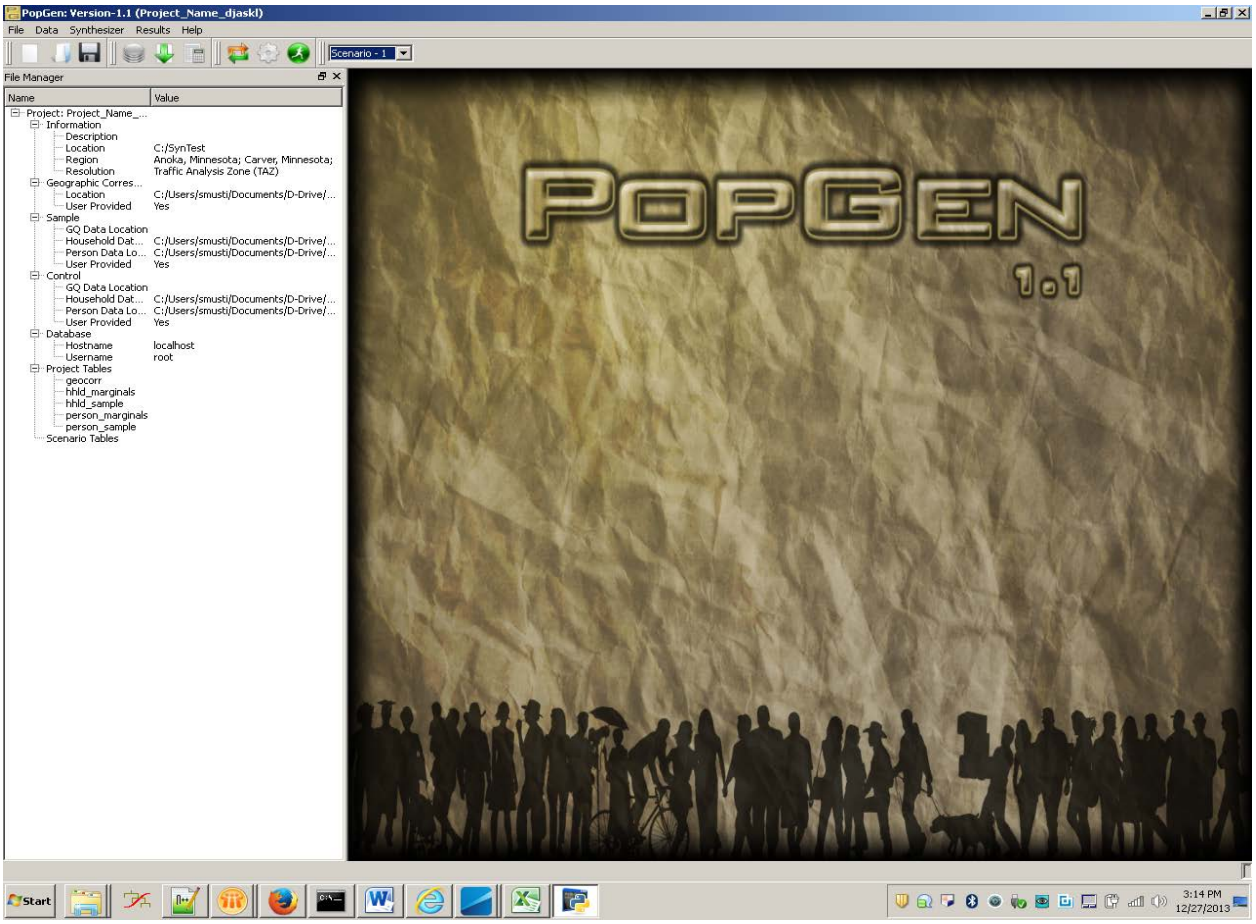
| pumano | hhid | serialno | pnum | gender | emp | agep | stud | Ptype | hchild1 | hchild2 | hchild3 | hfwrk | hpwrk | hsen | hadtstd | hadtnwa | hhldinc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint |
| 1302 | 1 | 1 | 1 | 1 | 1 | 9 | 2 | 5 | 0 | 4 | 4 | 1 | 0 | 0 | 0 | 1 | 5 |
| 1302 | 1 | 1 | 2 | 2 | 3 | 9 | 2 | 7 | 0 | 4 | 4 | 1 | 0 | 0 | 0 | 1 | 5 |
| 1302 | 1 | 1 | 3 | 2 | 3 | 2 | 1 | 2 | 0 | 4 | 4 | 1 | 0 | 0 | 0 | 1 | 5 |
| 1302 | 1 | 1 | 4 | 1 | 3 | 2 | 1 | 2 | 0 | 4 | 4 | 1 | 0 | 0 | 0 | 1 | 5 |
| 1302 | 1 | 1 | 5 | 1 | 3 | 2 | 1 | 2 | 0 | 4 | 4 | 1 | 0 | 0 | 0 | 1 | 5 |
| 1302 | 1 | 1 | 6 | 1 | 3 | 2 | 1 | 2 | 0 | 4 | 4 | 1 | 0 | 0 | 0 | 1 | 5 |
| 1501 | 2 | 2 | 1 | 1 | 2 | 6 | 2 | 6 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1501 | 2 | 2 | 2 | 2 | 1 | 6 | 2 | 5 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1501 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1301 | 3 | 3 | 1 | 2 | 2 | 8 | 2 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1700 | 4 | 4 | 1 | 1 | 2 | 9 | 2 | 6 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 |
| 1700 | 4 | 4 | 2 | 2 | 2 | 9 | 2 | 6 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 |
| 1800 | 5 | 5 | 1 | 1 | 1 | 8 | 2 | 5 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 4 |
| 1800 | 5 | 5 | 2 | 2 | 1 | 9 | 2 | 5 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 4 |
| 1502 | 6 | 6 | 1 | 2 | 3 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 |
| 1502 | 6 | 6 | 2 | 1 | 3 | 8 | 2 | 7 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 |
| 1201 | 7 | 7 | 1 | 1 | 1 | 7 | 2 | 5 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 4 |
| 1201 | 7 | 7 | 2 | 1 | 2 | 6 | 1 | 6 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 4 |
| 1201 | 7 | 7 | 3 | 1 | 1 | 5 | 2 | 5 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 4 |
| 1201 | 7 | 7 | 4 | 2 | 1 | 5 | 1 | 5 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 4 |
| 1201 | 7 | 7 | 5 | 1 | 2 | 7 | 2 | 6 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 4 |
| 700 | 8 | 8 | 1 | 2 | 3 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 700 | 8 | 8 | 2 | 2 | 2 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 1403 | 9 | 9 | 1 | 1 | 1 | 6 | 2 | 5 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 |
| 1403 | 10 | 10 | 1 | 2 | 1 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 1403 | 10 | 10 | 2 | 2 | 3 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 700 | 11 | 11 | 1 | 2 | 2 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 5 |
| 1700 | 12 | 12 | 1 | 2 | 3 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3 |
| 1700 | 12 | 12 | 2 | 1 | 3 | 10 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3 |
| 1100 | 13 | 13 | 1 | 2 | 1 | 7 | 2 | 5 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 5 |
| 1100 | 13 | 13 | 2 | 1 | 2 | 7 | 2 | 6 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 5 |
| 1100 | 13 | 13 | 3 | 1 | 3 | 3 | 1 | 2 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 5 |
| 1100 | 13 | 13 | 4 | 2 | 3 | 2 | 1 | 2 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 5 |
| 1403 | 14 | 14 | 1 | 1 | 3 | 8 | 2 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 1201 | 15 | 15 | 1 | 1 | 1 | 6 | 2 | 5 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 |
| 1201 | 15 | 15 | 2 | 2 | 1 | 6 | 2 | 5 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 |

## Step 2: Open PopGen and Load Input Files

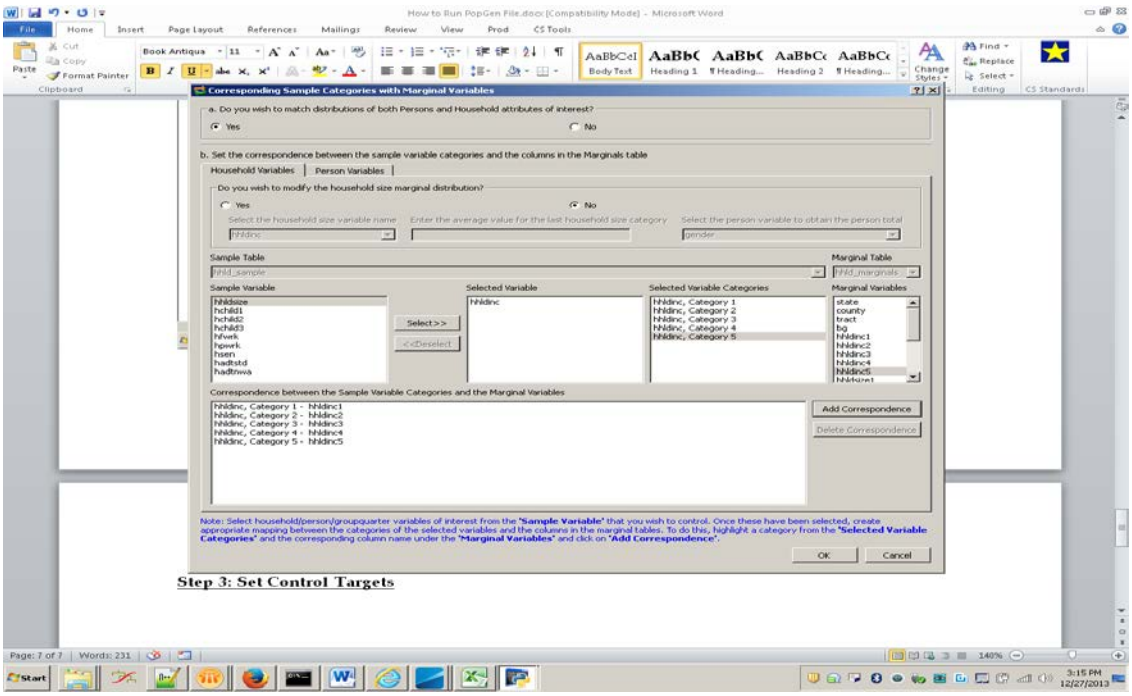


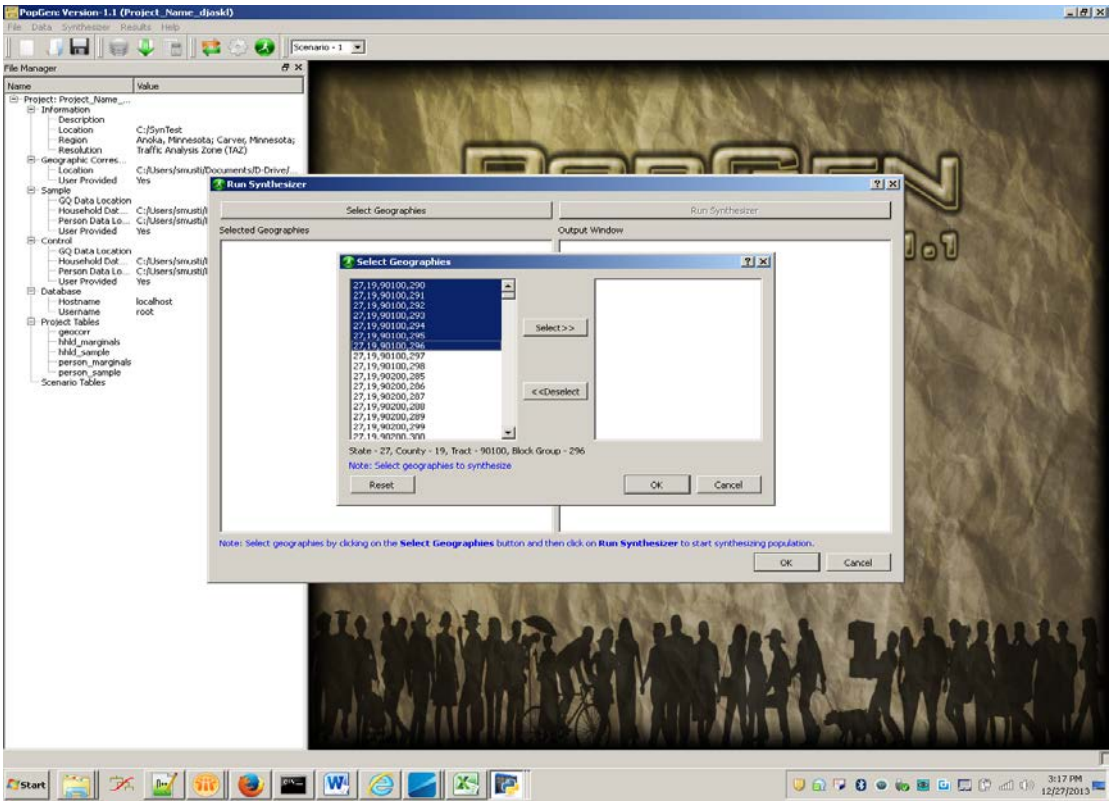
## Step 2 (Continued): Open PopGen and Load Input Files



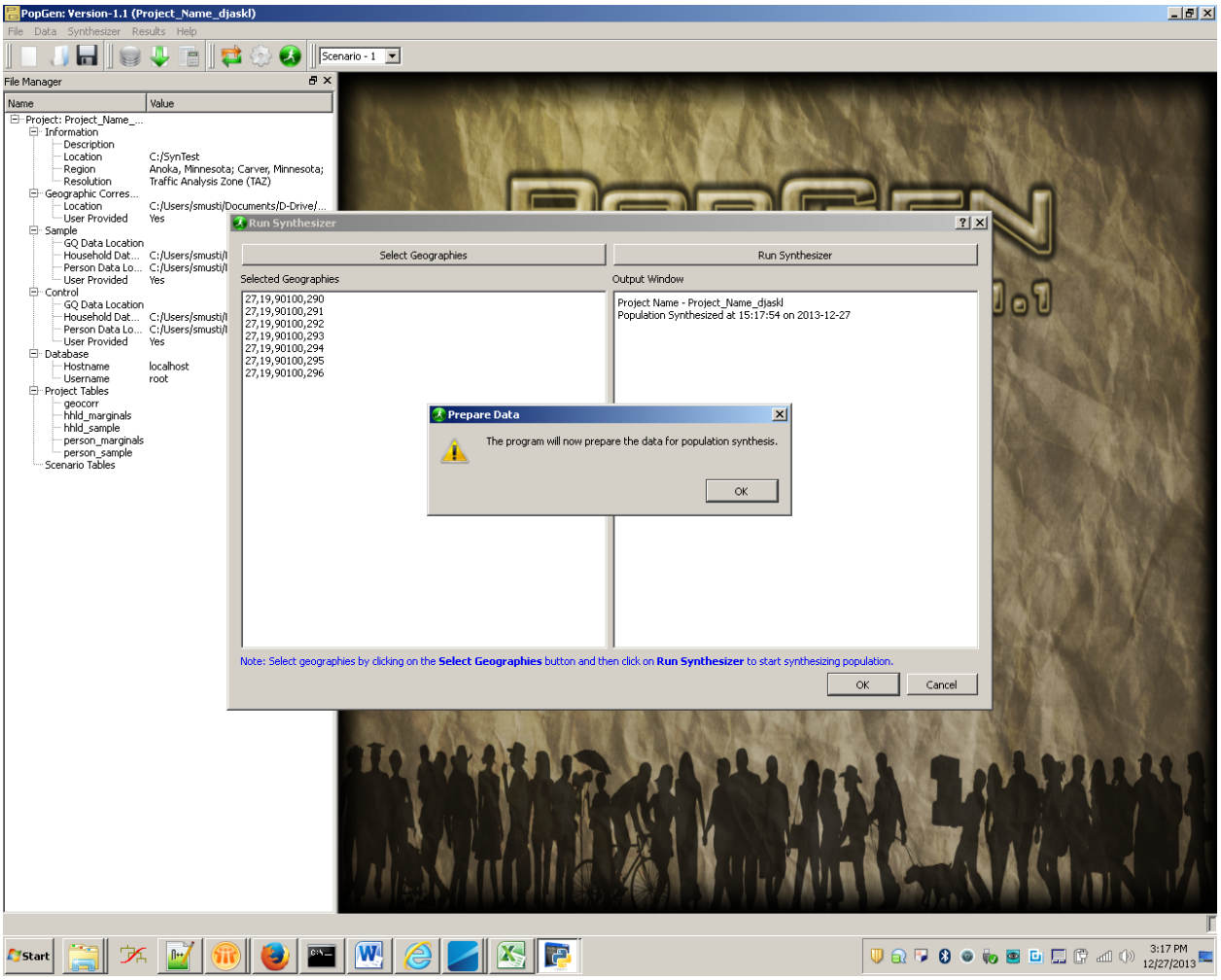
## Step 3: Set Control Targets




## Step 4: Select TAZ's for which population needs to be synthesized



## Step 5: Run PopGen



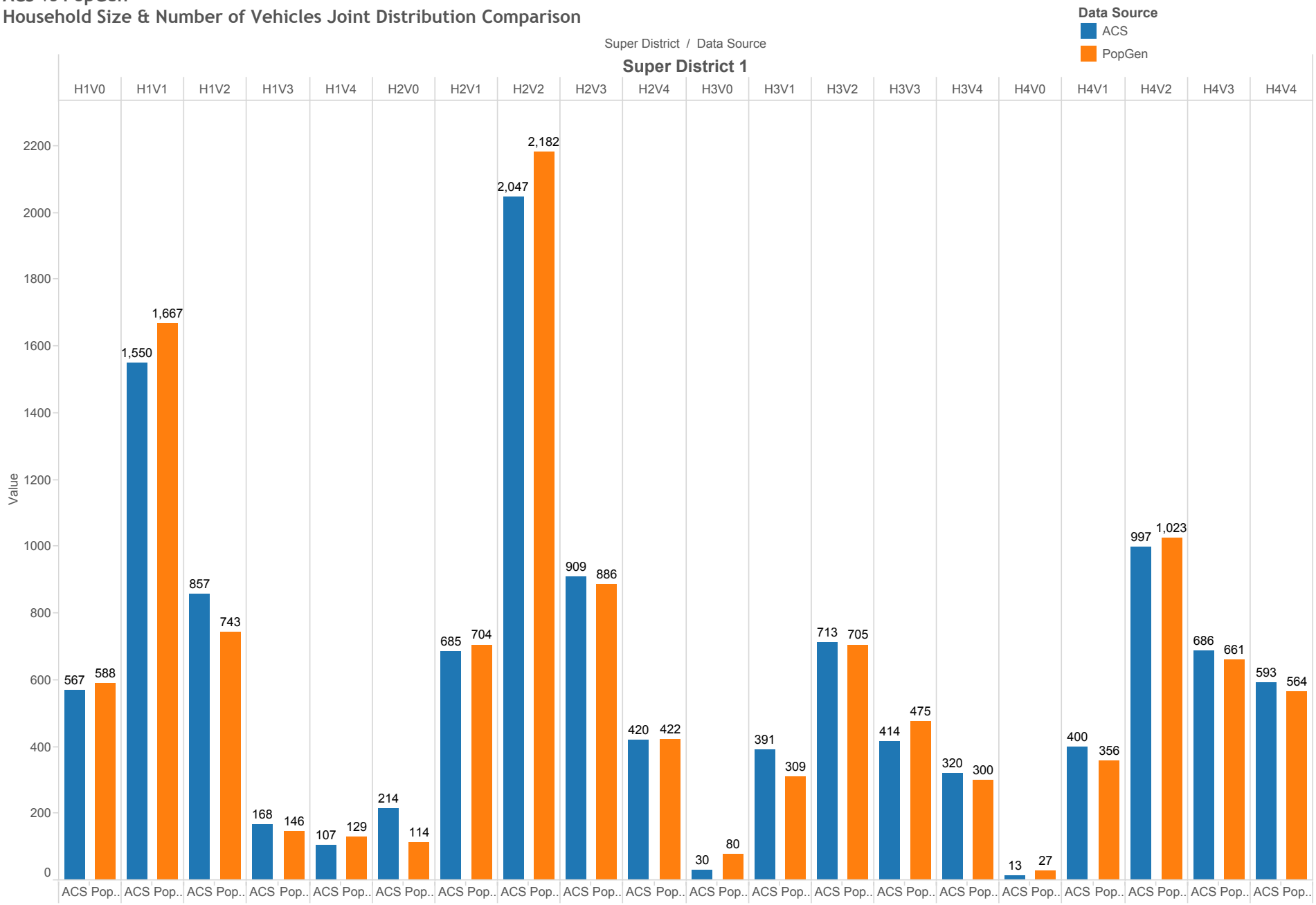
## B.2 Two-Dimensional Joint Distribution Comparisons between ACS and Synthetic Population for the TRPC Model Area

### a. Household Size and Number of Vehicles by District

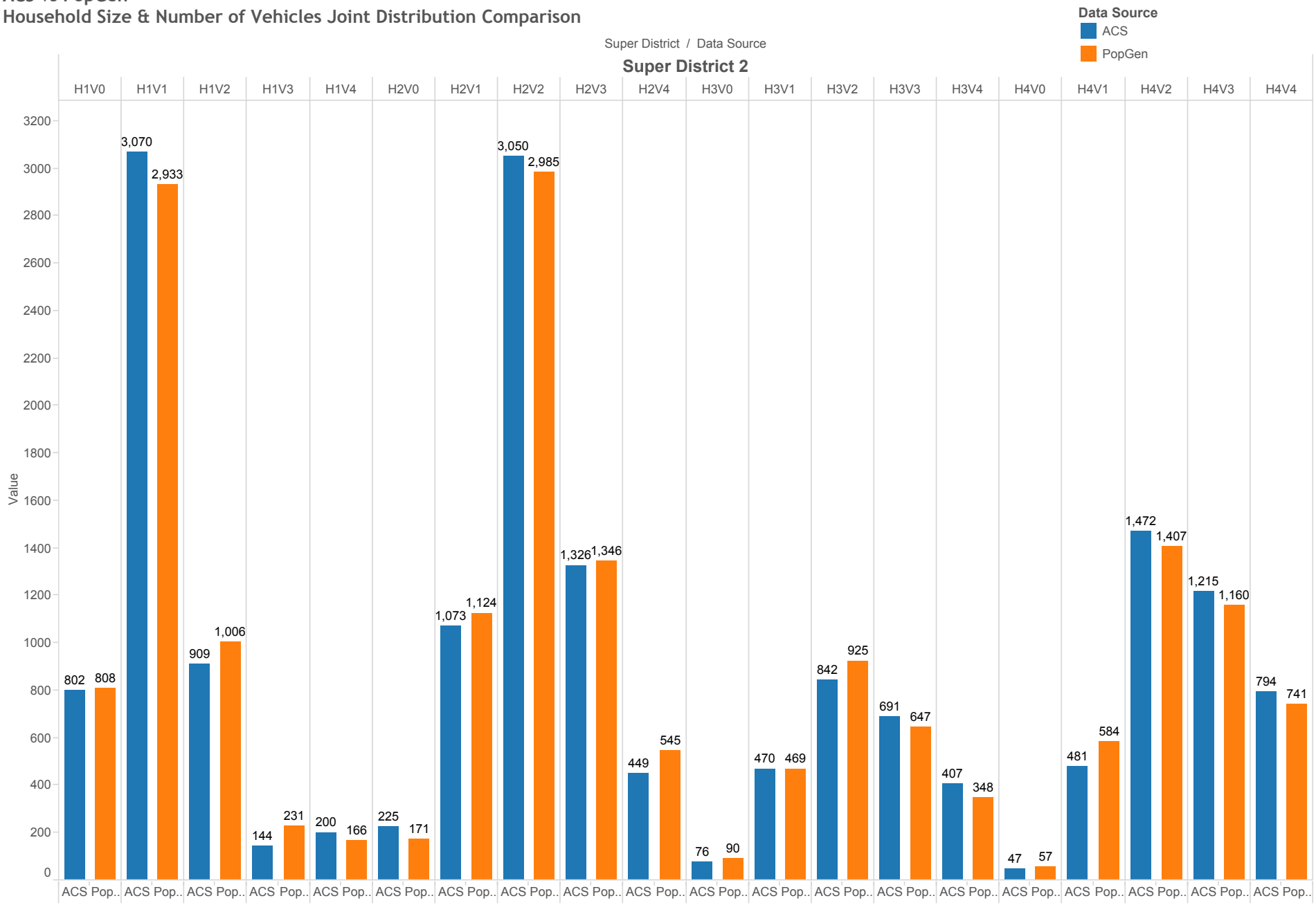
### b. Household Size and Number of Workers by District

### c. Number of Vehicles and Number of Workers by District

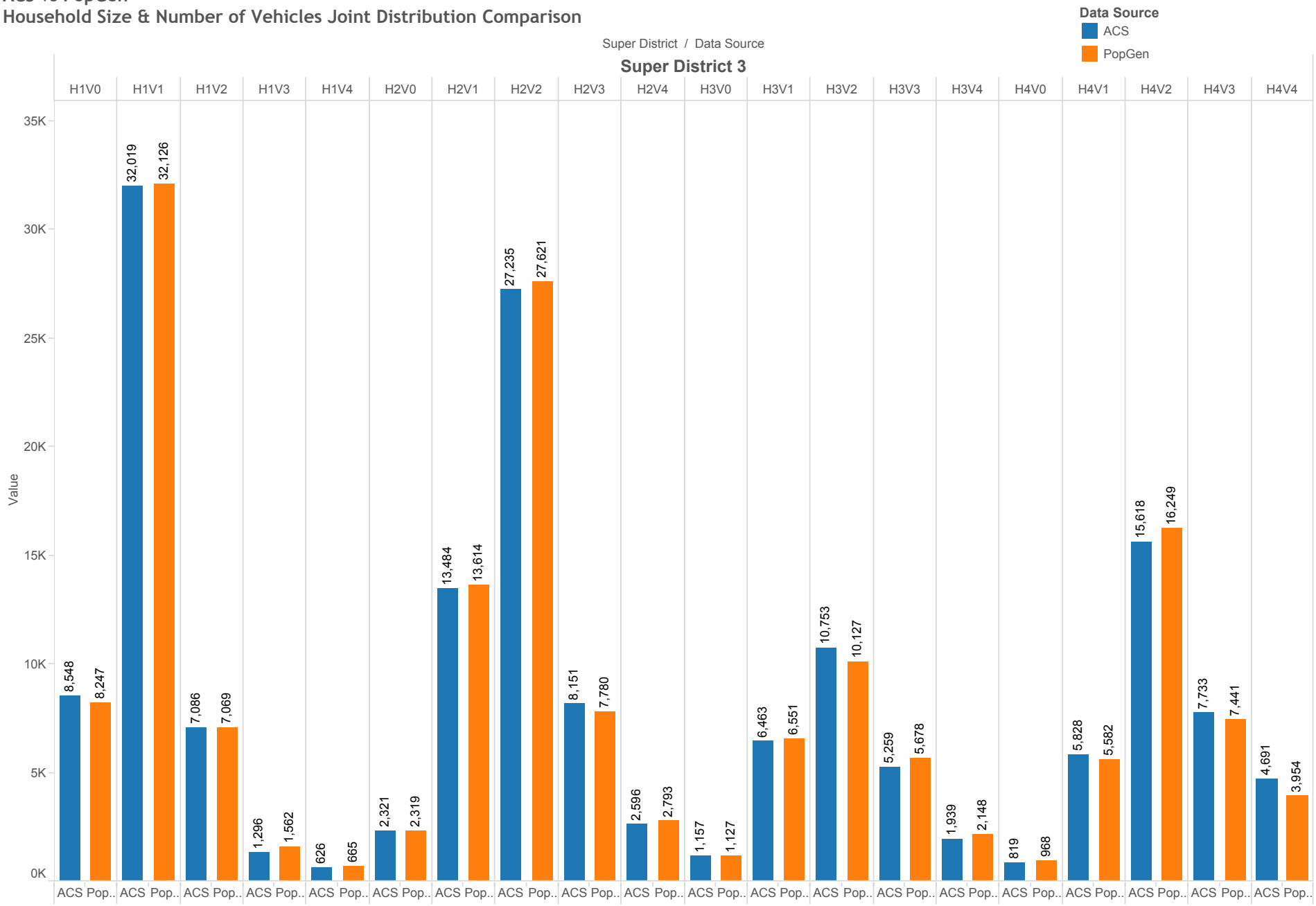
## ACS vs PopGen
## Household Size & Number of Vehicles Joint Distribution Comparison



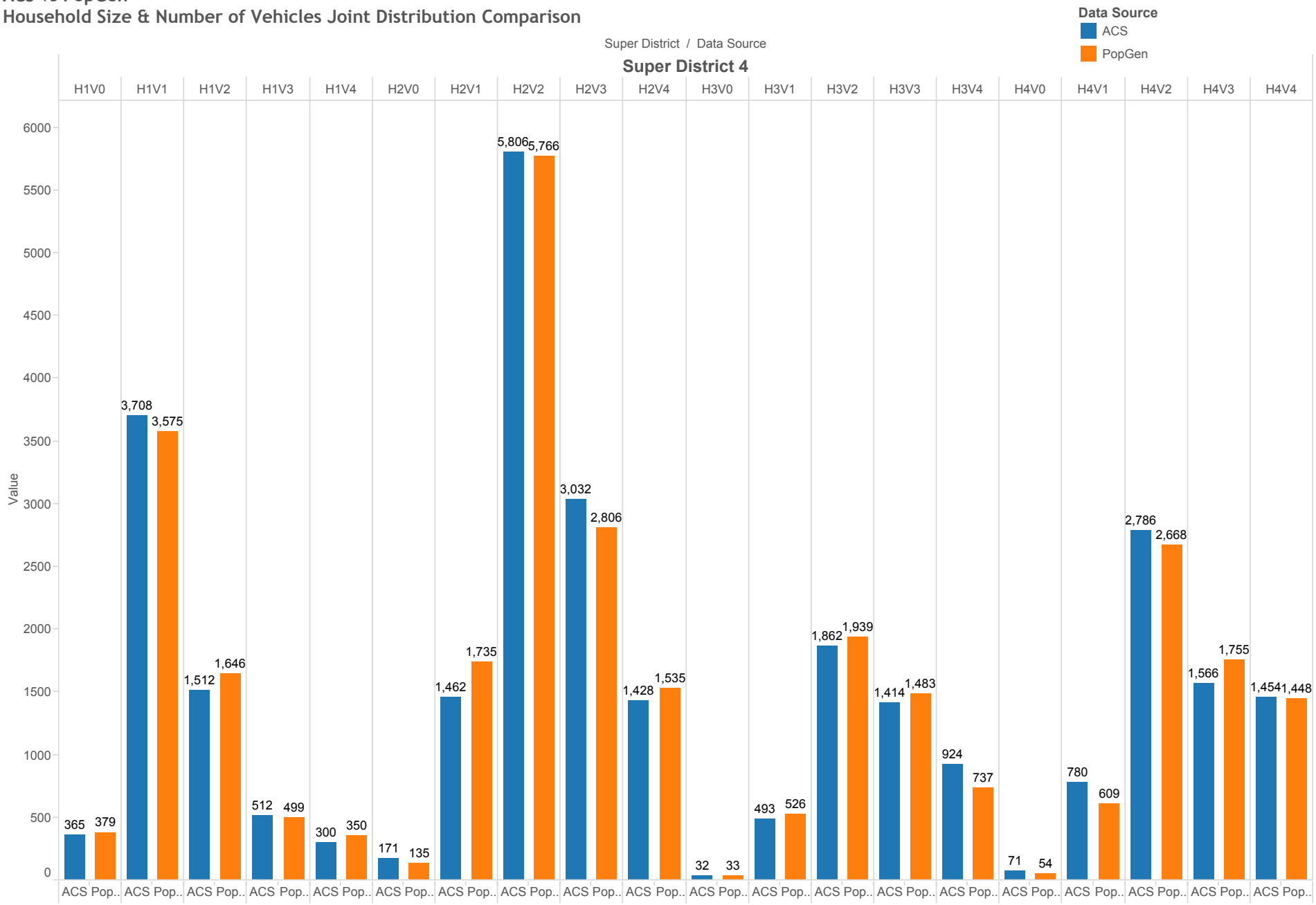
## ACS vs PopGen
## Household Size & Number of Vehicles Joint Distribution Comparison



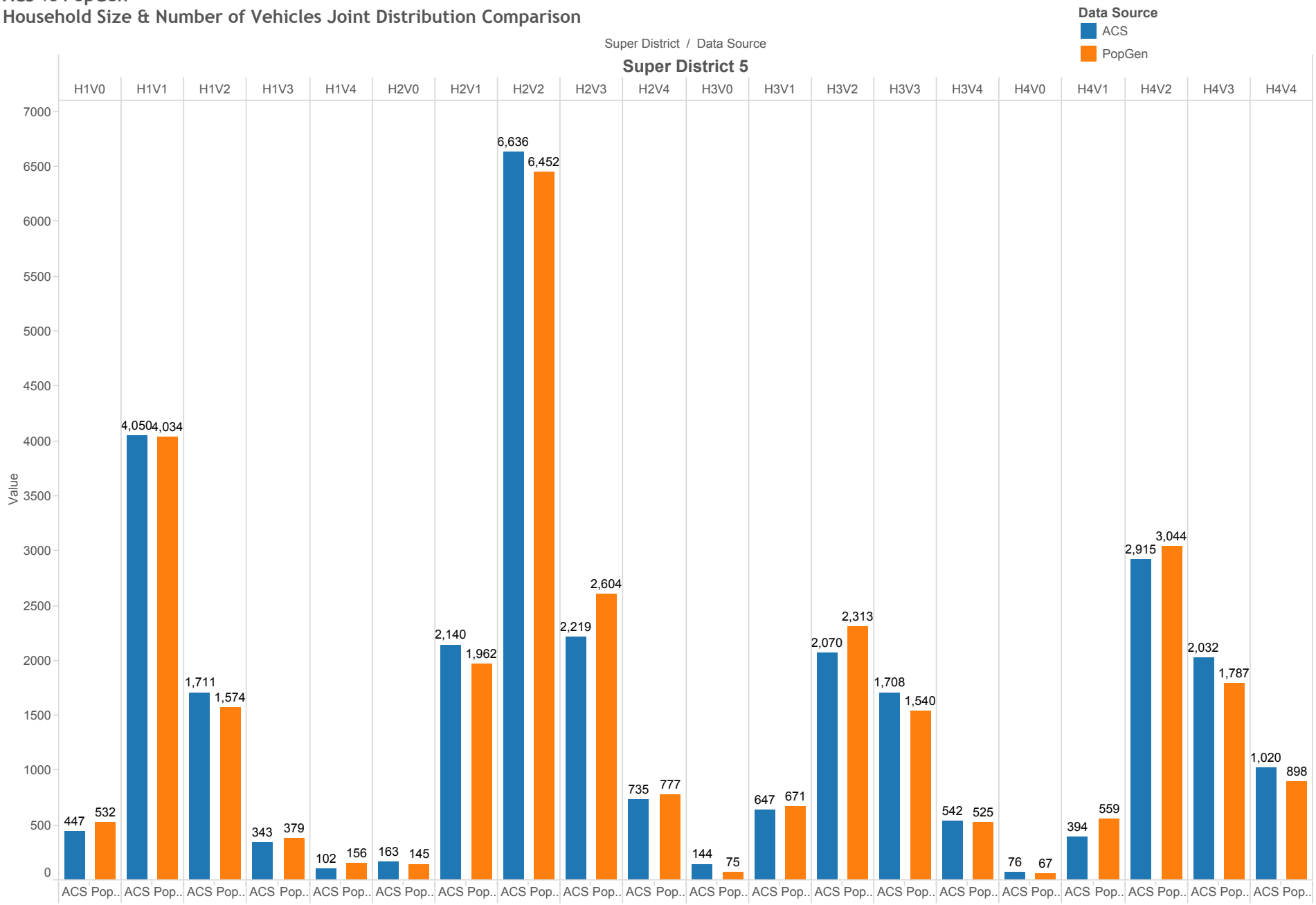
## ACS vs PopGen
## Household Size & Number of Vehicles Joint Distribution Comparison



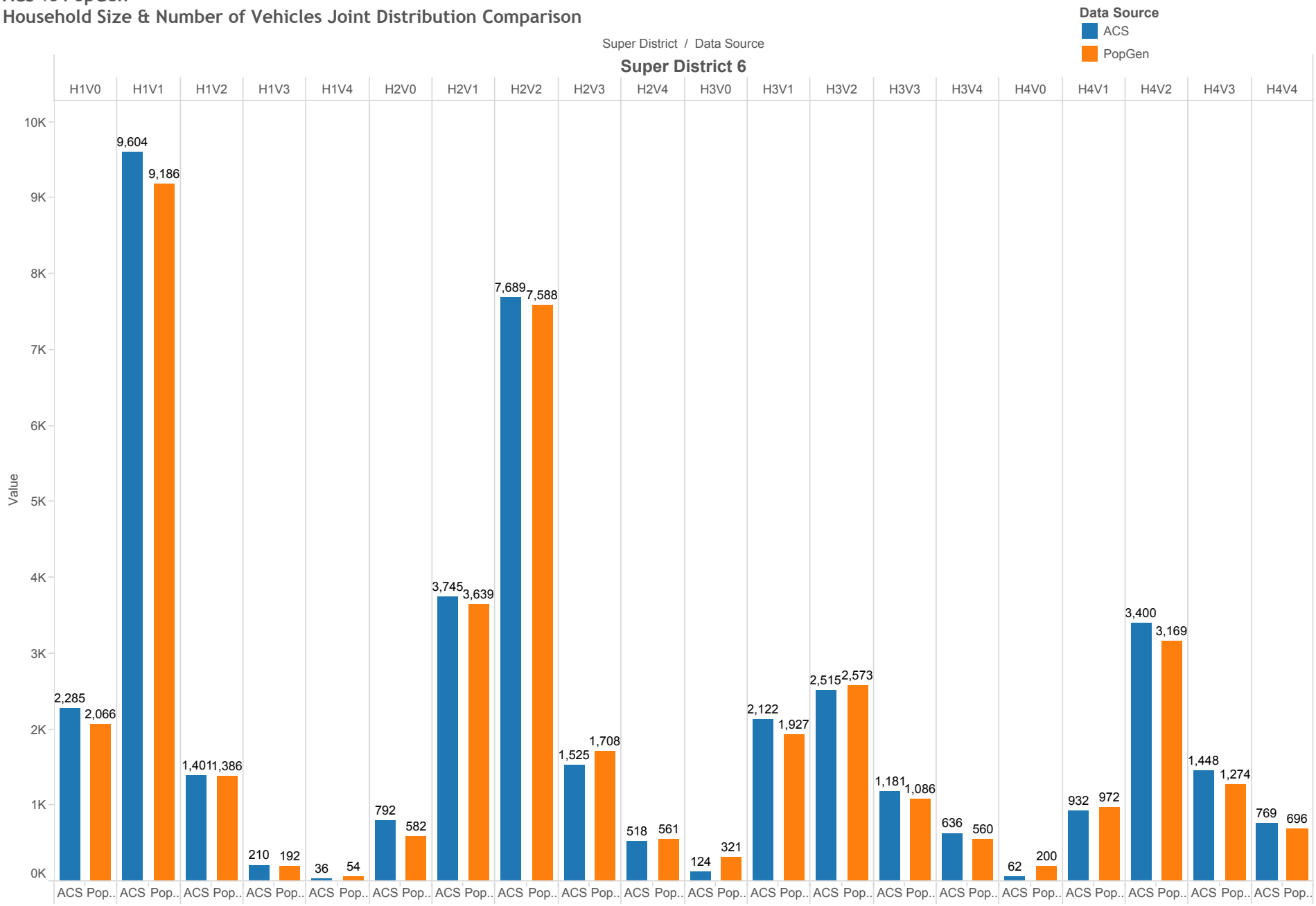
## ACS vs PopGen
## Household Size & Number of Vehicles Joint Distribution Comparison



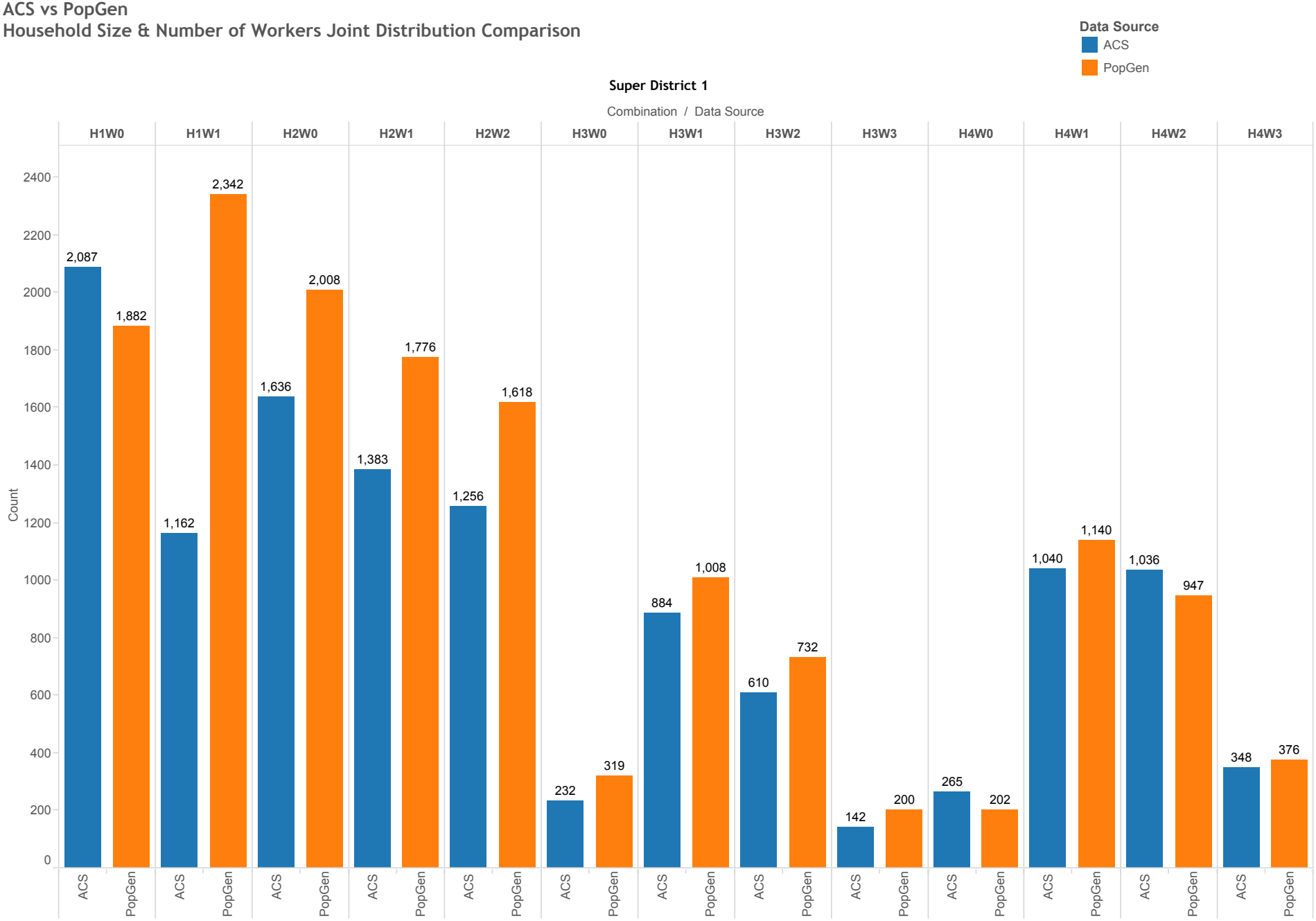
## ACS vs PopGen
## Household Size & Number of Vehicles Joint Distribution Comparison



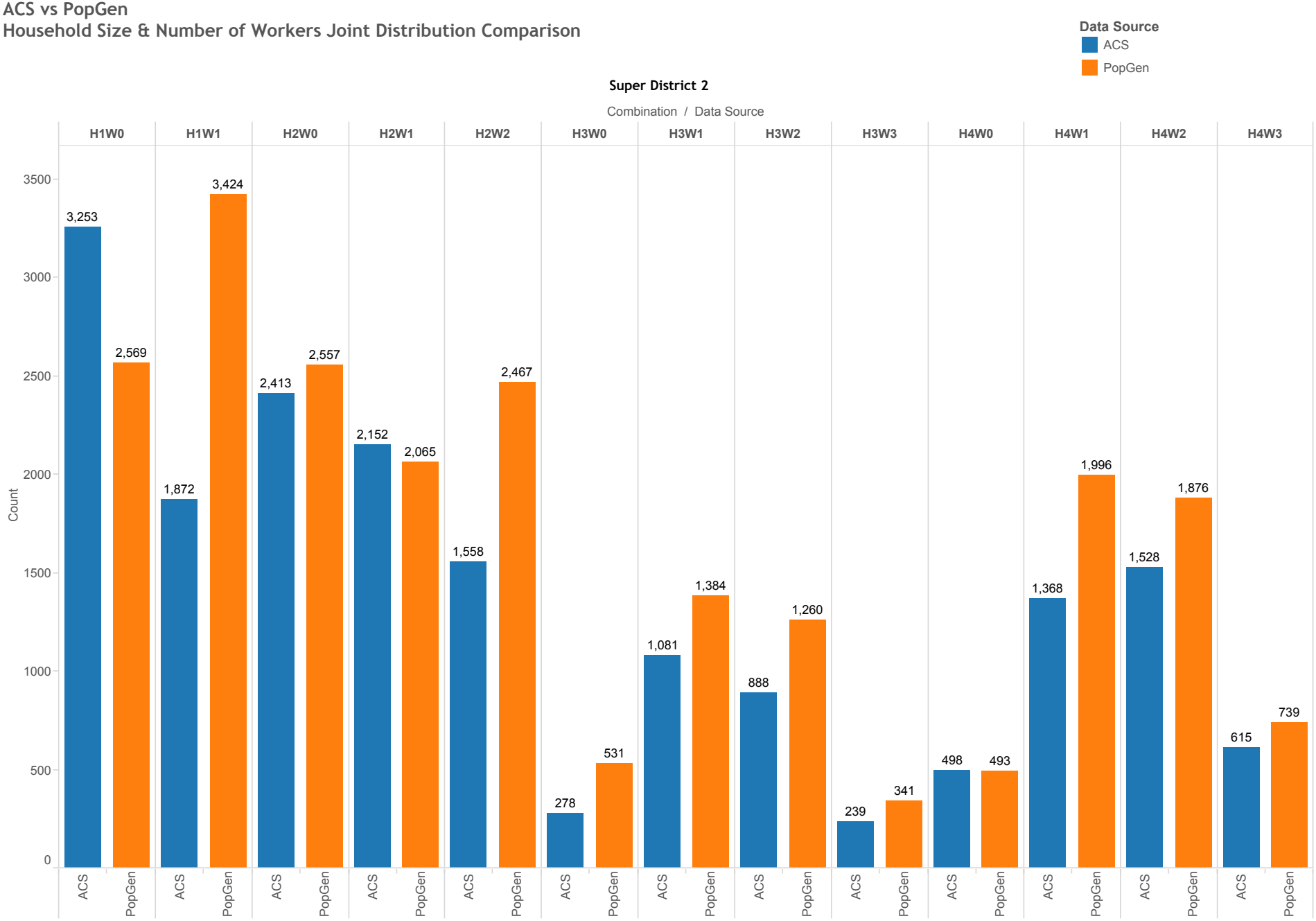
## ACS vs PopGen
## Household Size & Number of Vehicles Joint Distribution Comparison



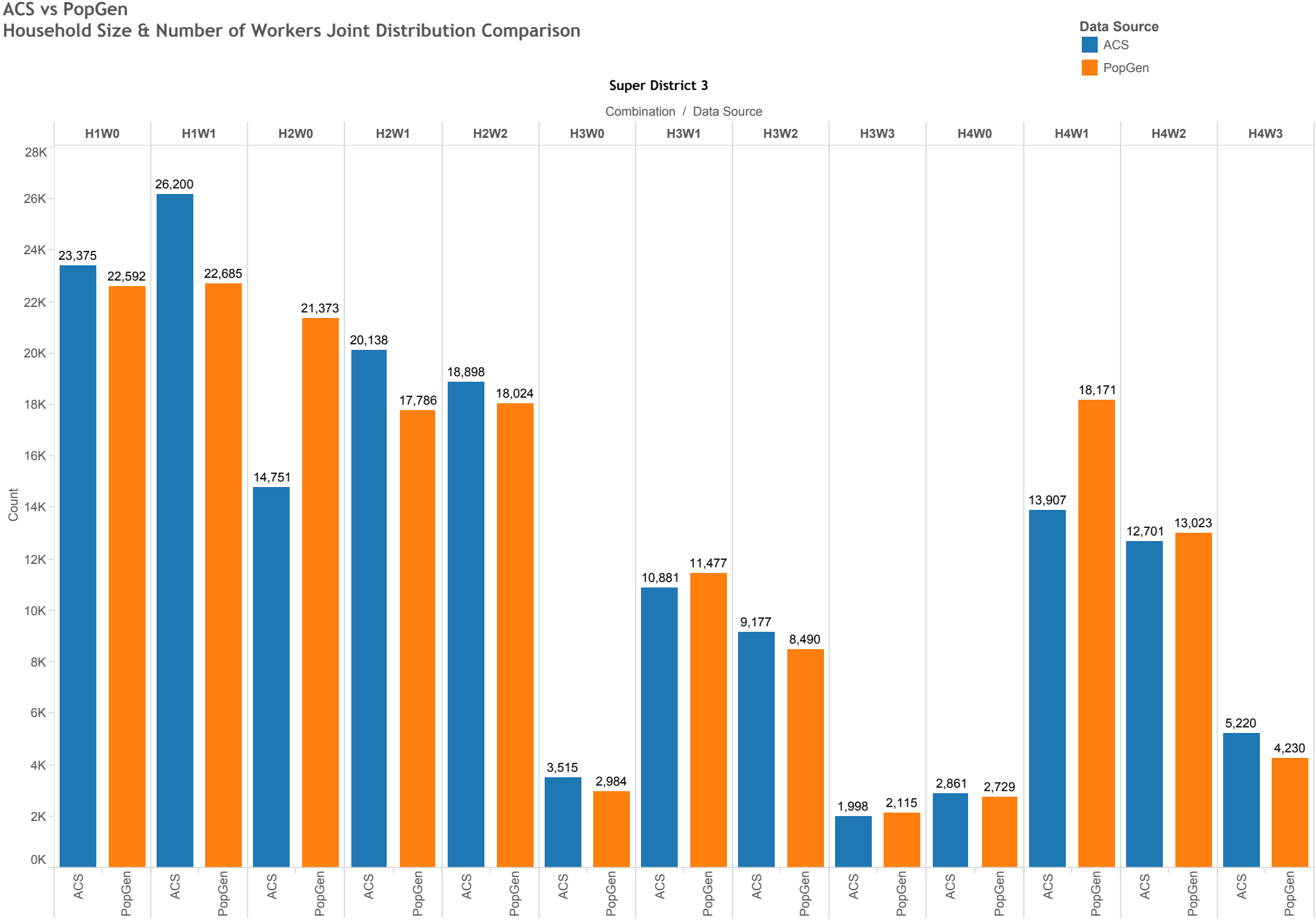
## ACS vs PopGen
## Household Size & Number of Workers Joint Distribution Comparison



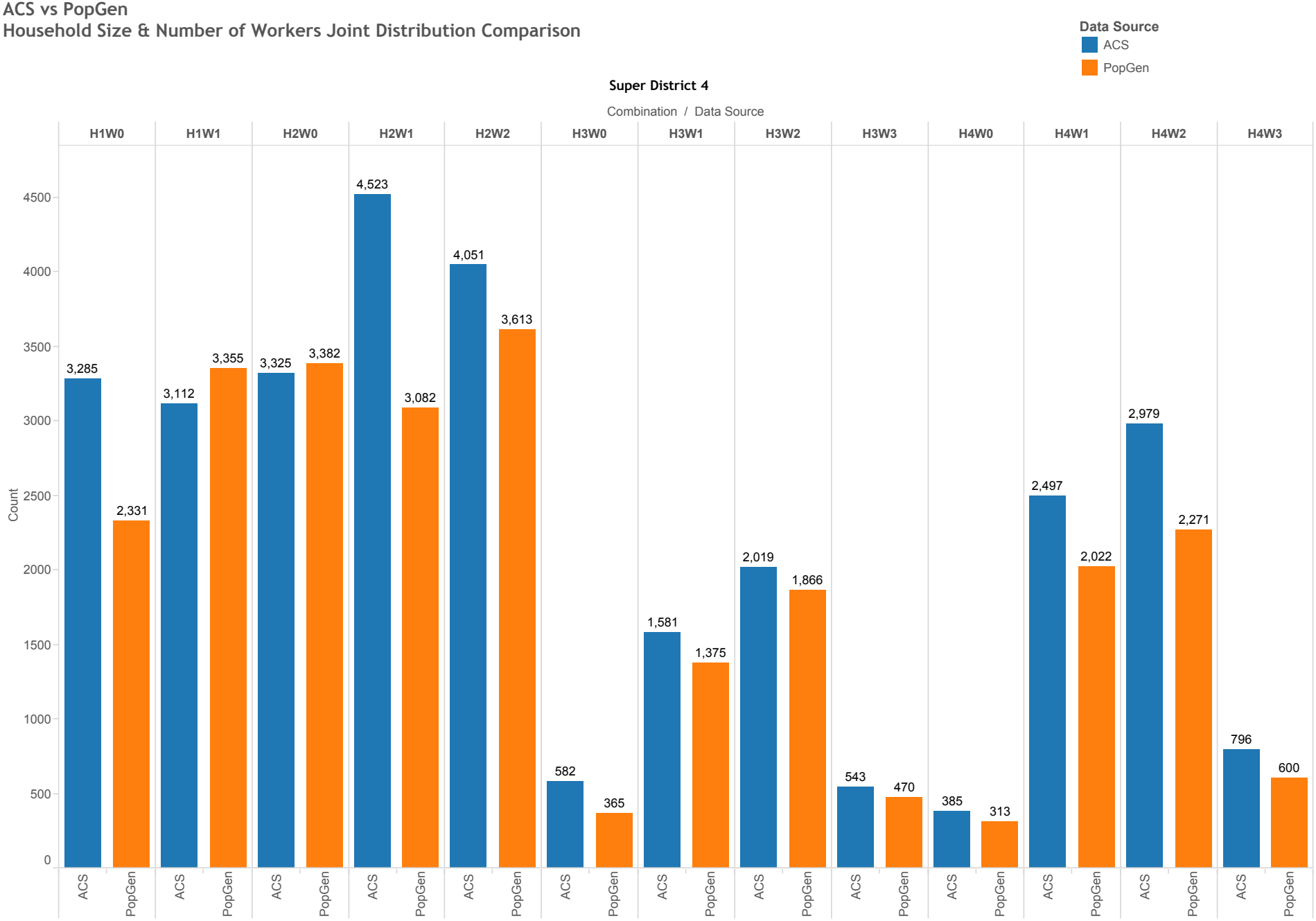
## ACS vs PopGen
## Household Size & Number of Workers Joint Distribution Comparison



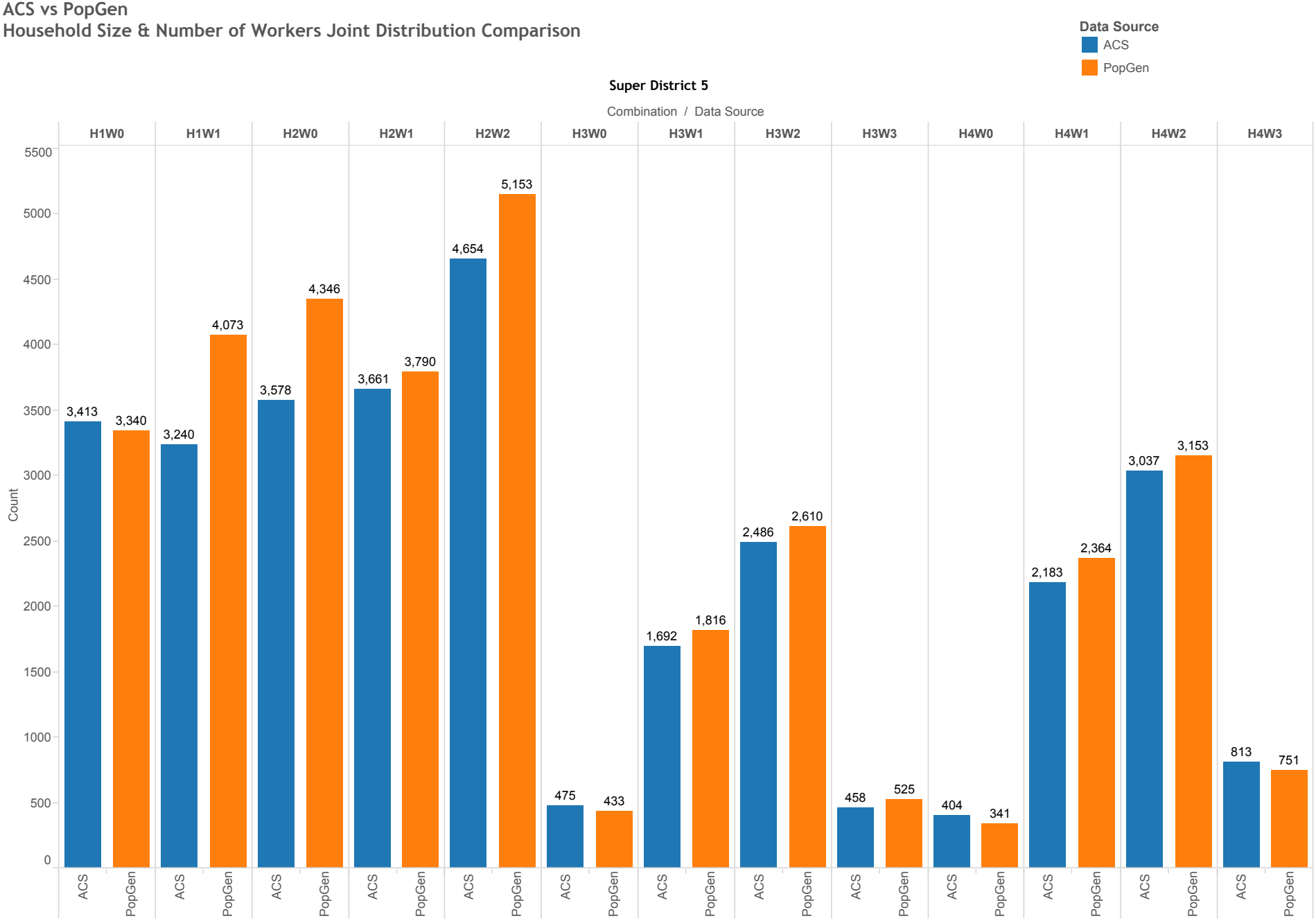
## ACS vs PopGen
## Household Size & Number of Workers Joint Distribution Comparison



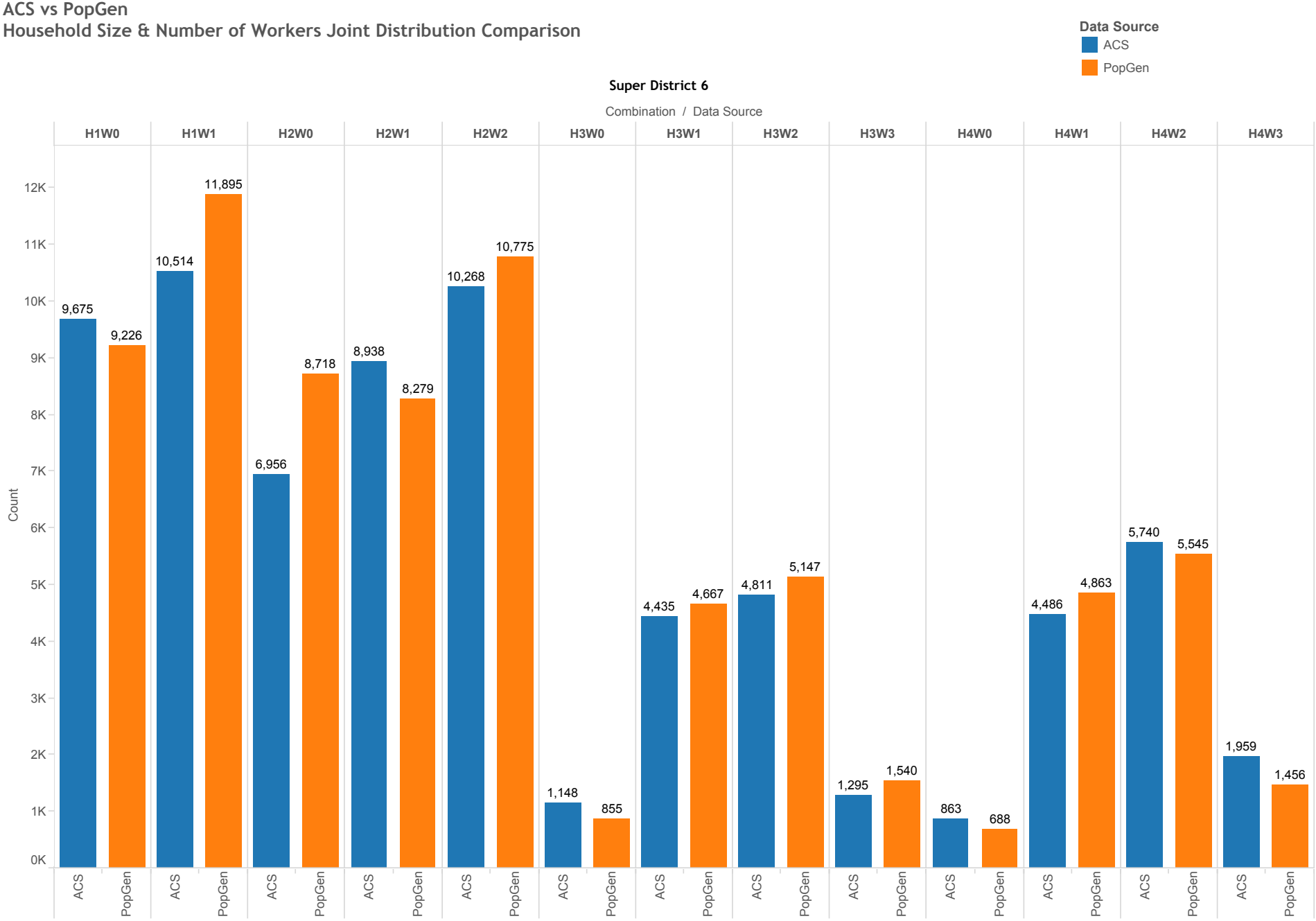
## ACS vs PopGen
## Household Size & Number of Workers Joint Distribution Comparison



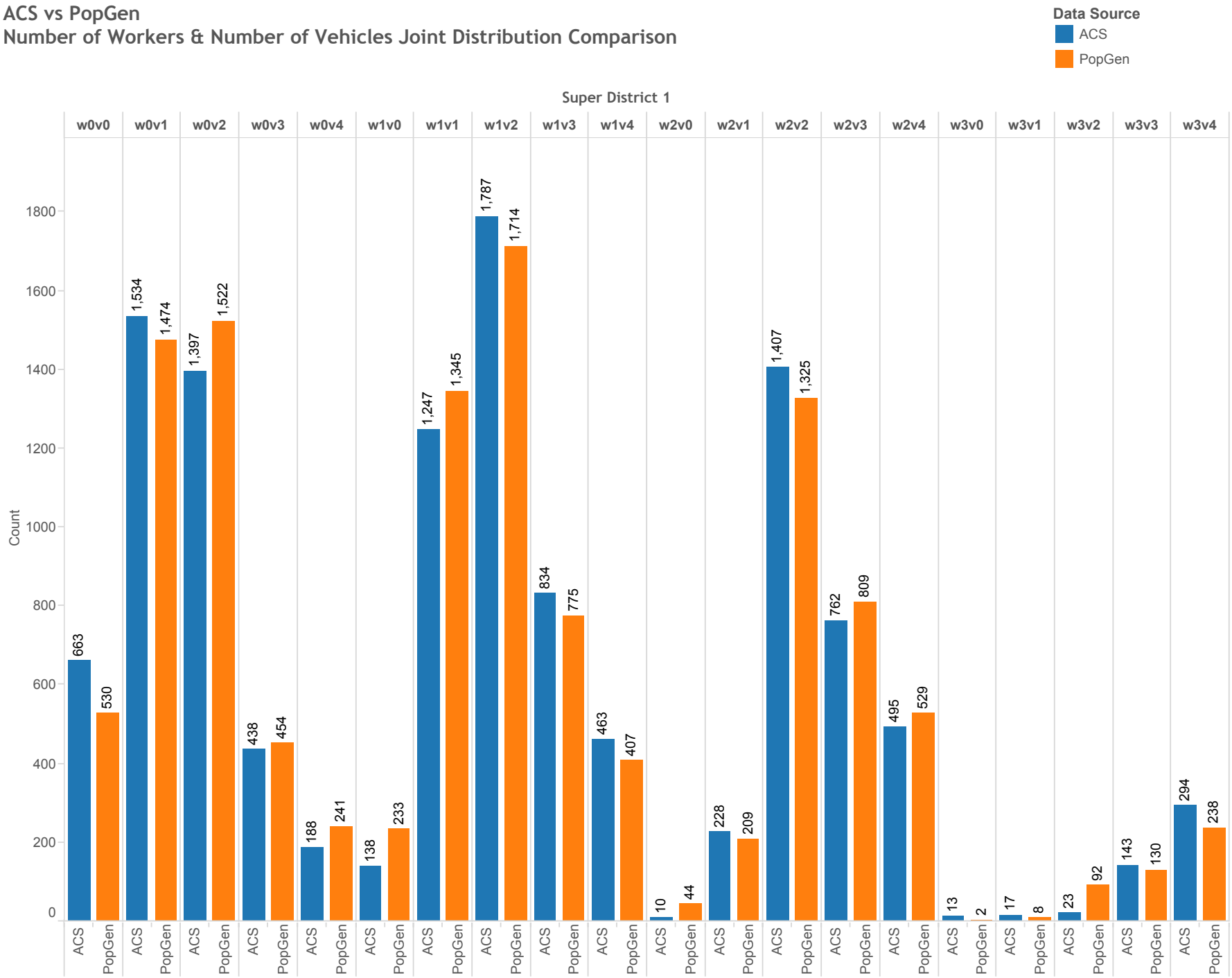
## ACS vs PopGen
## Household Size & Number of Workers Joint Distribution Comparison



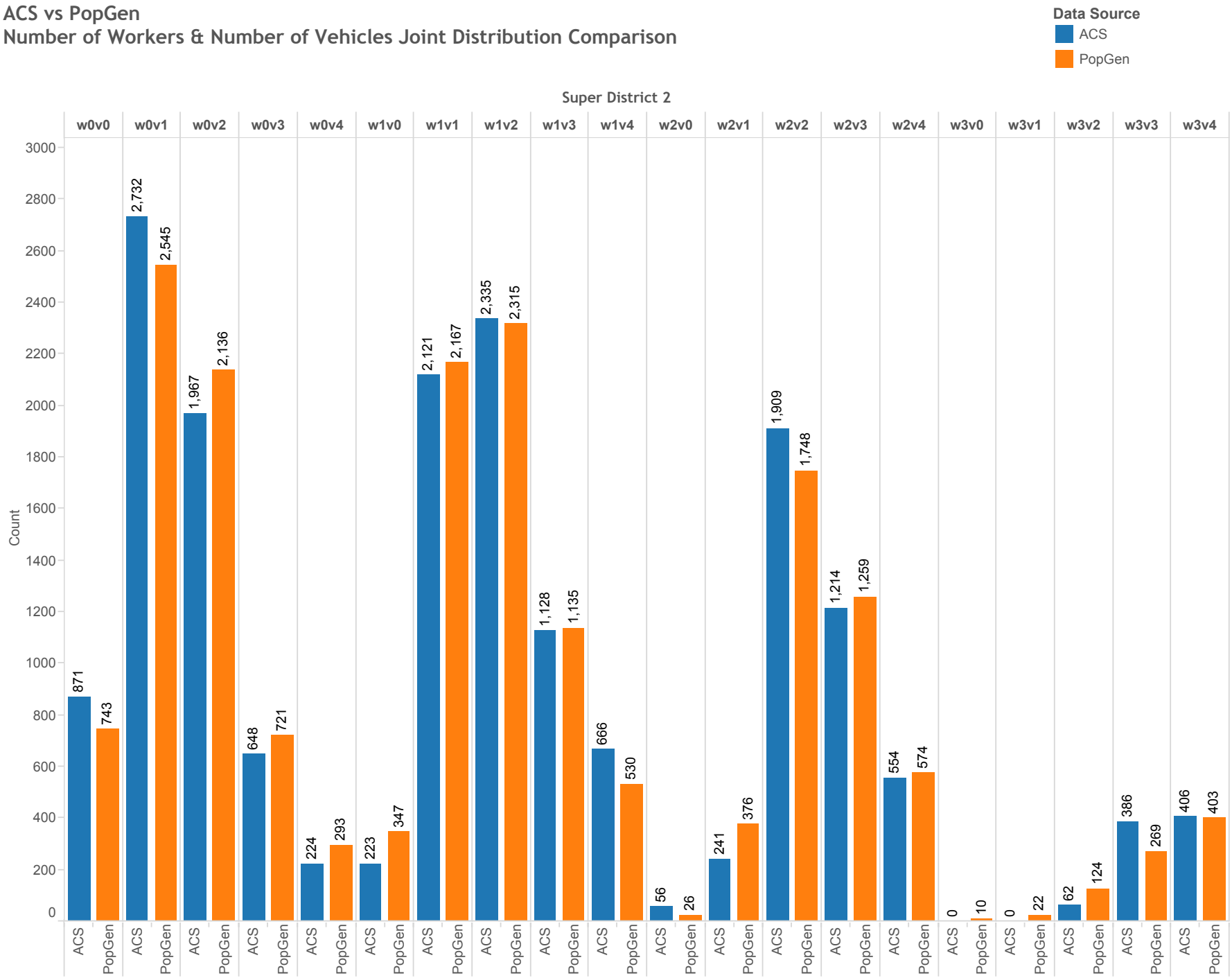
## ACS vs PopGen
## Household Size & Number of Workers Joint Distribution Comparison



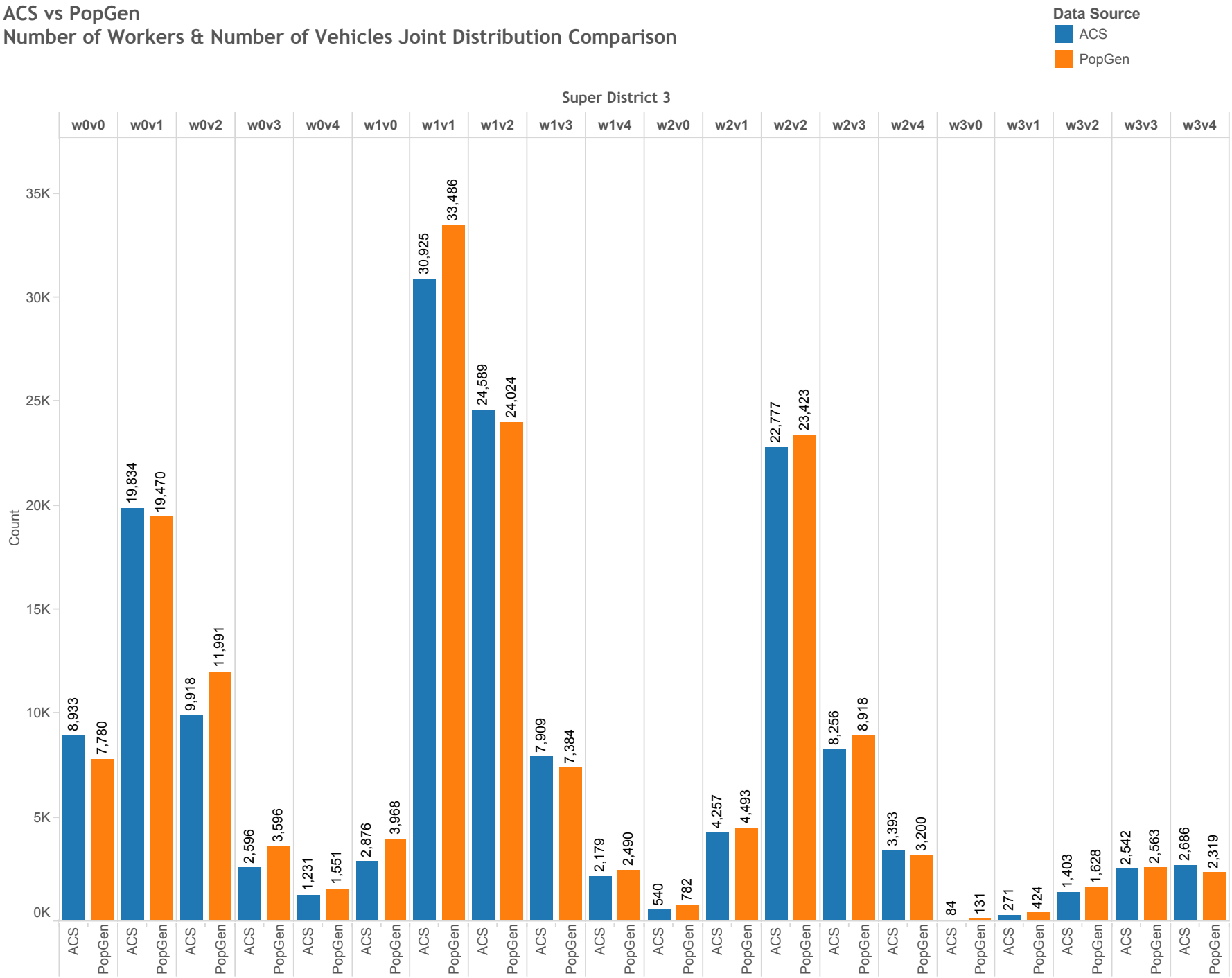
## ACS vs PopGen
## Number of Workers & Number of Vehicles Joint Distribution Comparison



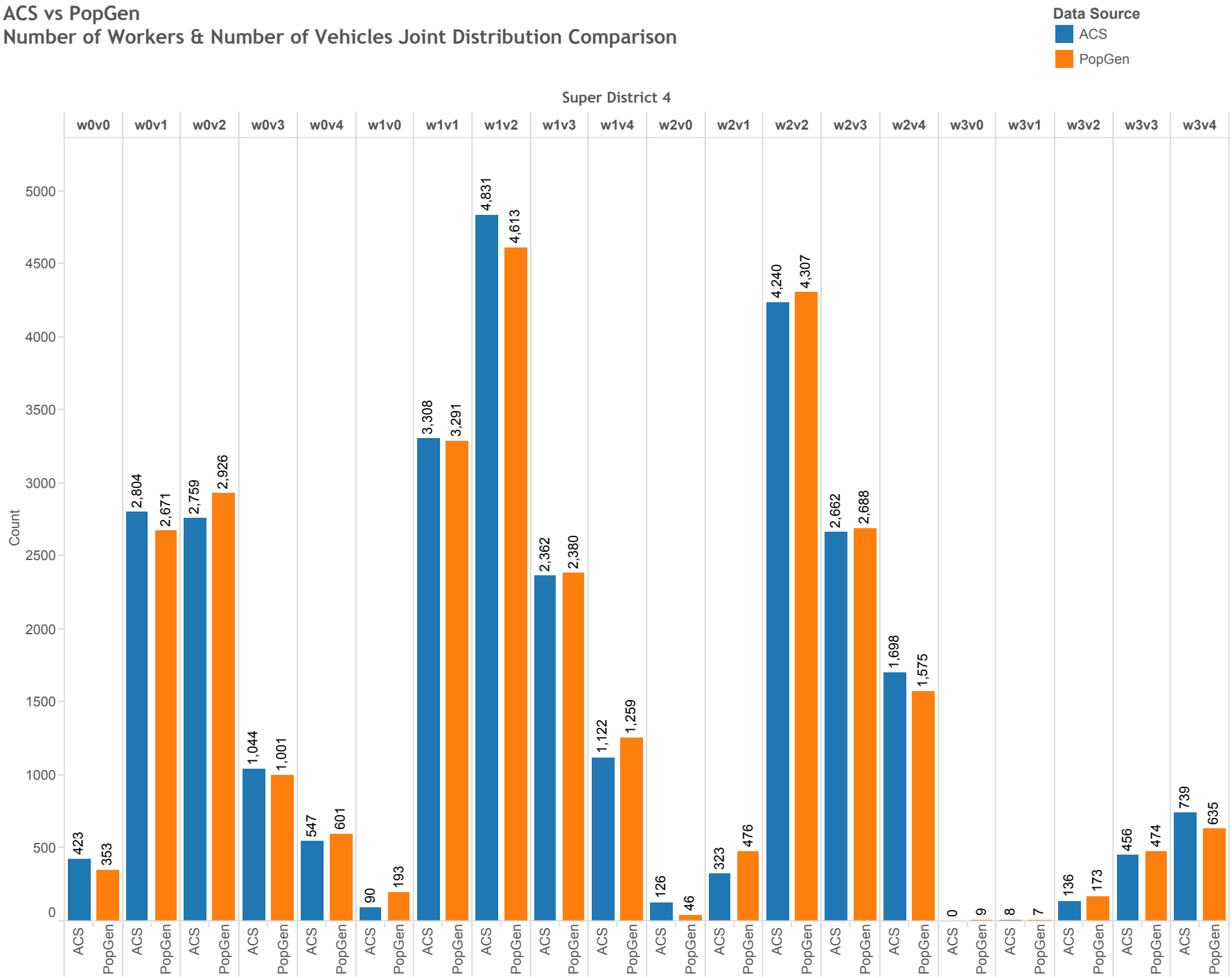
## ACS vs PopGen
## Number of Workers & Number of Vehicles Joint Distribution Comparison



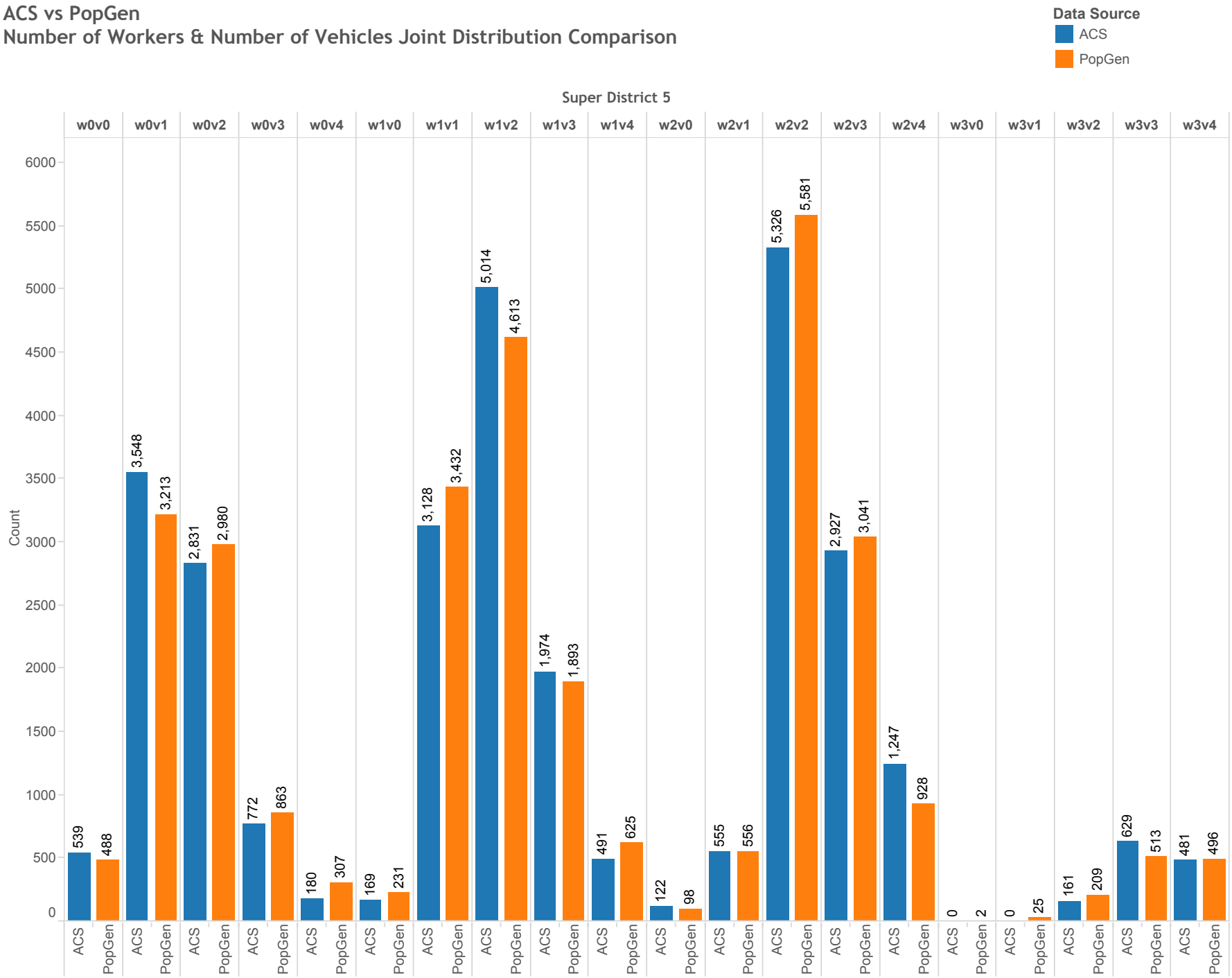
## ACS vs PopGen
## Number of Workers & Number of Vehicles Joint Distribution Comparison



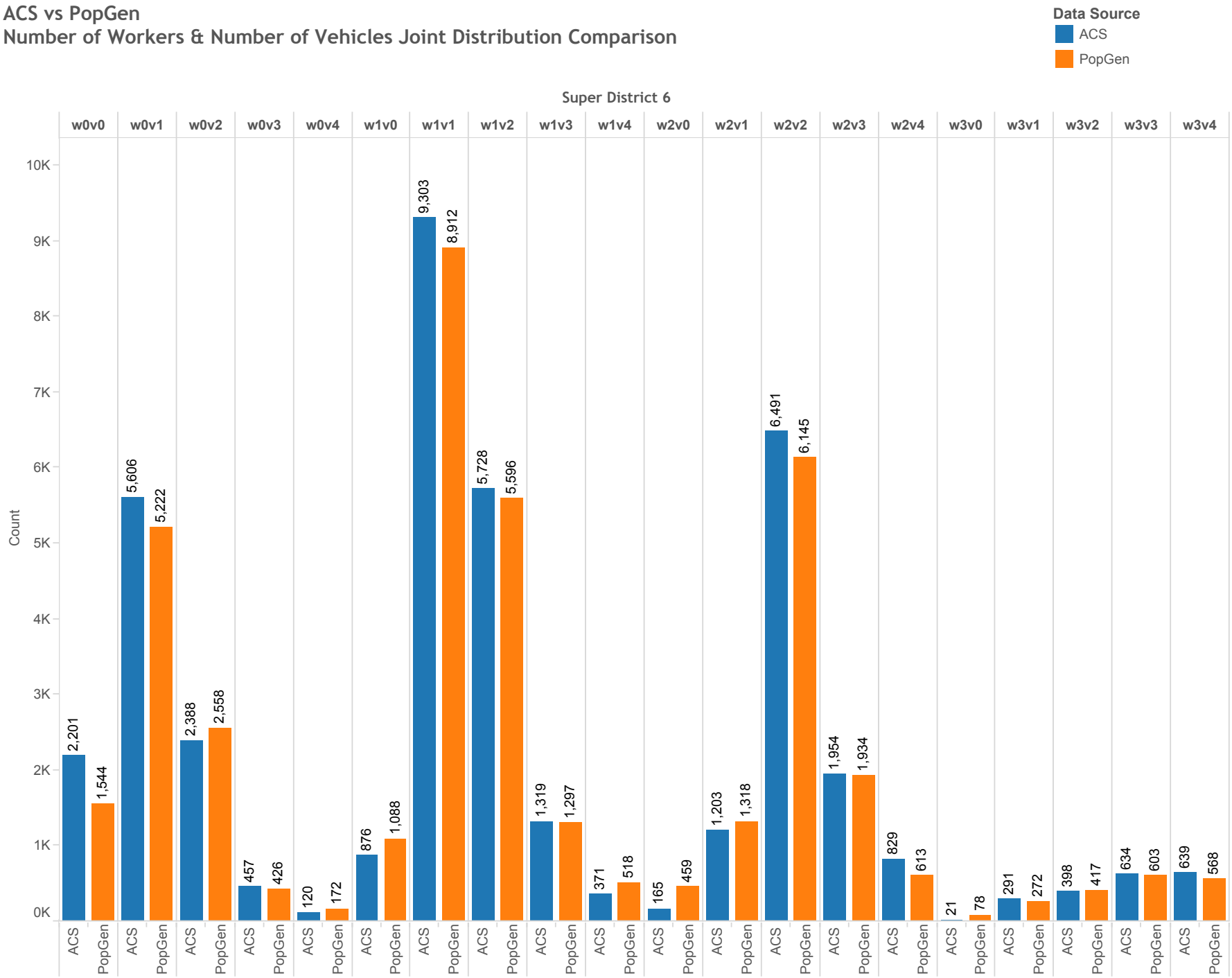
## ACS vs PopGen
## Number of Workers & Number of Vehicles Joint Distribution Comparison



## ACS vs PopGen
## Number of Workers & Number of Vehicles Joint Distribution Comparison



## ACS vs PopGen
## Number of Workers & Number of Vehicles Joint Distribution Comparison



## B.3 PopGen Training Materials

- *PopGen Training Workshop –Session I – Synthetic Population Generation for Travel Demand Forecasting; May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session II - PopGen Installation; May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session III – Data Structures May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session IV – Project Setup using Census Data; May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session V – Project Setup using User Provided Data; May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session VI – Modifying Data; May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session VII Configuring a Project; May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session VIII – Generating a Synthetic Population;; May 12-13, 2010, Arizona State University*
- *PopGen Training Workshop - Session IX – Visualization and Exploring Results; May 12-13, 2010, Arizona State University*

IRA A. FULTON SCHOOL OF ENGINEERING

Leading engineering discovery and innovative education for global impact on quality of life.

# Synthetic Population Generation for Travel Demand Forecasting

## PopGen Training Workshop

May 12 and 13, 2010

http://urbanmodel.asu.edu/popgen.html

ARIZONA STATE UNIVERSITY

# Acknowledgements


- *Software Development:* Karthik Konduri and Bhargava Sana
- *Graphic Support and Documentation:* Keith Christian
- *Methodology:* Xin Ye, University of Maryland; Hillel Bar-Gera, Ben-Gurion University, Israel
- *Sponsors:*
  - Arizona State University, School of Sustainable Engineering and the Built Environment, Ira A. Fulton School of Engineering
  - Exploratory Advanced Research Program (EARP), Federal Highway Administration, US Department of Transportation

# Outline

- Motivation for population synthesis
- What is population synthesis?
  - Standard IPF procedure
- Motivation for enhanced population synthesis
- Design of a new population synthesizer
- New Iterative Proportional Updating (IPU) Algorithm
  - Explanation of procedure
  - Geometric Interpretation
- Test Application
  - Computing household weights
  - Generating a synthetic population
  - Algorithm performance
- Demonstration of PopGen Open Source Software Package

# Microsimulation Models of Travel

- Increasing interest in microsimulation models for travel demand forecasting
- Microsimulation models simulate travel at the level of the individual decision-maker while recognizing inter-dependencies among activities, trips, persons, time, and space
- Microsimulation models of travel increasingly based on activity-based paradigm of travel behavior
  - Explicit recognition of derived nature of travel demand
  - Enhanced representation of time-space interactions and constraints

# Microsimulation Models of Travel (continued)

- Activity-based microsimulation modeling approaches offer ability to address emerging policy questions of interest
- By simulating activities and travel at the level of the individual traveler, these models are able to address impacts of:
  - Greenhouse gas emissions reduction targets
  - Flexible working arrangements
  - Impact of information and communication technology (ICT)
  - Interactions between micro-scale land use changes and travel
  - Pricing-based policies
  - Non-motorized transportation mode enhancements

# Why Population Synthesis?

- We need disaggregate household and person socio-demographic data for entire population of model region
- Such data for the entire population is generally not available
- This leads to the need to synthesize a regional population from known statistical distributions on the population
- We have:
  - Disaggregate data for a sample of the population (PUMS, travel surveys)
  - Marginal distributions for the entire region (census summary files, agency forecasts)

# What is Population Synthesis?

***Population synthesis involves generating a synthetic population by expanding the disaggregate sample data to mirror known aggregate distributions of household and person variables of interest.***

# Standard IPF-Based Procedure

- Standard IPF (iterative proportional fitting)-based procedure based on Beckman et al (1996)
- Procedure
  - Choose household-level control variables
  - Obtain the marginal distributions on these variables from census summary files (SF)
  - Generate a seed matrix of the joint distribution from a microdata sample data set (PUMS, travel survey)
  - Expand the seed matrix using an IPF-procedure to match the given marginal control totals while maintaining the joint distribution implied by the seed matrix

# Standard IPF-Based Procedure (continued)

- Selection probabilities are estimated for households in the microdata sample
- Households are drawn using the selection probabilities to match the expanded cell frequencies
- The resulting synthetic population is checked for goodness-of-fit and households are redrawn if necessary
- The synthetic population is comprised of all individuals within the synthesized (drawn) households

# Illustration of IPF Procedure

## Sample Seed Data and Summary Marginal Distributions

| | | Income | | Total | Household Size Marginals |
|---|---|---|---|---|---|
| | | Low | High | | |
| Household Size | Adjustment | -- | -- | | |
| 1 | -- | 3.0 | 1.0 | 4.0 | 30.0 |
| 2 | -- | 2.0 | 4.0 | 6.0 | 40.0 |
| 3 or more | -- | 2.0 | 1.0 | 3.0 | 30.0 |
| Total | | 7.0 | 6.0 | | |
| Income Marginals | | 60.0 | 40.0 | | |

Marginal Distributions

Seed Data

# Illustration of IPF Procedure (continued)

## Iteration 1: Adjustment for Income

| | | Income | | Total | Household Size Marginals |
|---|---|---|---|---|---|
| | | Low | High | | |
| Household Size | Adjustment | 60/7 = 8.57 | 6.67 | | |
| 1 | -- | 3 x 8.57 = 25.7 | 6.7 | 32.4 | 30.0 |
| 2 | -- | 17.1 | 26.7 | 43.8 | 40.0 |
| 3 or more | -- | 17.1 | 6.7 | 23.8 | 30.0 |
| Total | | 60.0 | 40.0 | | |
| Income Marginals | | 60.0 | 40.0 | | |

# Illustration of IPF Procedure (continued)

## Iteration 1: Adjustment for Household Size

| | | Income | | Total | Household Size Marginals |
|---|---|---|---|---|---|
| | | Low | High | | |
| Household Size | Adjustment | -- | -- | | |
| 1 | 30.0/32.4 = 0.93 | 25.7 x 0.93 = 23.8 | 6.2 | 30.0 | 30.0 |
| 2 | 0.91 | 15.7 | 24.3 | 40.0 | 40.0 |
| 3 or more | 1.26 | 21.6 | 8.4 | 30.0 | 30.0 |
| Total | | 61.1 | 38.9 | | |
| Income Marginals | | 60.0 | 40.0 | | |

# Illustration of IPF Procedure (continued)

## After 3 Iterations, convergence is achieved

| | | Income | | Total | Household Size Marginals |
|---|---|---|---|---|---|
| | | Low | High | | |
| Household Size | Adjustment | -- | -- | | |
| 1 | 1.00 | 23.6 | 6.4 | 30.0 | 30.0 |
| 2 | 1.00 | 15.2 | 24.8 | 40.0 | 40.0 |
| 3 or more | 1.00 | 21.3 | 8.7 | 30.0 | 30.0 |
| Total | | 60.0 | 40.0 | | |
| Income Marginals | | 60.0 | 40.0 | | |

Multiway frequency table matching known marginal distributions

# Summary of IPF Procedure

- The standard IPF-based procedure explained in detail in Beckman et al (1996)
- The IPF-based procedure has been implemented widely in various population synthesizers
- Following the estimation of the cell frequencies in the joint distribution, households are drawn probabilistically

# Motivation for Enhancement

- Key limitation of the standard IPF-based procedure
  - Controls only for household attributes and not person attributes
  - Synthetic populations fail to match distributions of person characteristics of interest
  - The method ignores differences in household composition among households within a cell
- ***Hence the need to re-assign weights to sample households based on household composition***

# Recent Literature Addresses Issue

- Guo and Bhat (2007)
  - "… deviation (in person attributes) could severely affect the accuracy of the subsequent microsimulation outcome …"
  - Household- and person- joint distributions are estimated using IPF procedure
  - Household selection probabilities computed based on target distributions of household types
  - A sample household is drawn so long as the household and person level frequency counts are within a certain threshold of the given distributions

# Recent Literature (continued)

- Arentze and Timmermans (2007)
  - Person level marginal constraints are converted into household level constraints using relational matrices
  - Household constraints and the converted person level constraints are used to estimate household joint distributions using the standard IPF procedure

# Recent Literature (continued)

- Pritchard and Miller (2009)
  - IPF implemented with a sparse list-based data structure that can accommodate a large number of control variables
  - A conditional Monte Carlo drawing procedure is adopted to simultaneously fit household and person marginal distributions
  - Persons within households are drawn from a pool while maintaining person to household relationships
  - Enhances the fit to person distributions while maintaining the match to household marginals

# Recent Literature (continued)

- Srinivasan et al (2009)
  - A “fitness value” is calculated for each sample household
  - “Fitness value” captures the contribution of the sample household in matching both household and person distributions
  - Synthetic population is generated by selecting sample households with the highest fitness values
  - Drawing process continues until the expected number of households are drawn or all fitness values become negative

# PopGen: A New Population Synthesizer

- Incorporates a new Iterative Proportional Updating (IPU) algorithm for estimating household weights
- The algorithm estimates sample household weights such that ***BOTH*** household and person distributions are matched
- Simple, practical, and computationally tractable algorithm with an intuitive interpretation
- Basic idea behind IPU algorithm in PopGen
  - ***Reallocate weights among sample households of a type to account for differences in household composition***

# PopGen Methodology

**Step 1: Estimate Household and Person Type Constraints**

- household and person sample data
- household and person level marginal distributions

- Adjust priors to account for zero-cell problem
- Adjust marginals to account for the zero-marginal problem
- Run Iterative Proportional Fitting (IPF) procedure to estimate ***household and person type constraints***

# PopGen Methodology (continued)

**Step 2: Estimate Household Weights**

- household and person sample data
- household and person type constraints from Step 1

- Run the Iterative Proportional Updating (IPU) algorithm to estimate sample ***household weights*** that satisfy both household and person type constraints

# PopGen Methodology (continued)

**Step 3: Generate the Synthetic Population**

- household and person sample data
- household weights from Step 2

- Apply rounding procedures to get the frequency of different household types in the synthetic population
- Estimate household selection probabilities using the computed weights
- Draw sample households based on selection probabilities for each household to match cell frequencies
- Repeat the process until a ***synthetic population*** with the best fit is obtained

# PopGen Terminology

- Household Type
  - Not to be confused with a household attribute 'household type'
  - Refers to a combination of household-level variables of interest
  - Represents a cell in the joint distribution of a set of household-level variables
- Person Type
  - Similar to above – formed by a combination of multiple person-level variables of interest

# PopGen Terminology (continued)

- A measure of fit ($\delta$ value)
  - Measures the absolute relative deviation between the IPU-adjusted cell frequency and the IPF-estimated household/person type constraints
  - Average $\delta$ value across all constraints is used as a goodness-of-fit measure
  - Average $\delta$ value is also used to monitor and set convergence criterion for the IPU algorithm

# PopGen Terminology (continued)

- A measure of fit ($\delta$ value)

$$\delta_j = \frac{\left| d_{i,j} w_i - c_j \right|}{c_j}$$

$d_{i,j}\, w_i$ = adjusted cell frequency

$c_j$ = the $j^{th}$ IPF-estimated constraint

# Illustration of IPU Algorithm

## Frequency Matrix

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 |
| **Weighted Sum** | | 3.00 | 5.00 | 9.00 | 7.00 | 7.00 |
| **Constraints** | | 35.00 | 65.00 | 91.00 | 65.00 | 104.00 |
| $\delta_0$ | | 0.9143 | 0.9231 | 0.9011 | 0.8923 | 0.9327 |

# Illustration of IPU Algorithm (continued)

## Adjustment with respect to household type constraints

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Weights 1 | Weights 2 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 11.67 | 11.67 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 11.67 | 11.67 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | 11.67 | 11.67 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | 1.00 | 13.00 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | 1.00 | 13.00 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | 1.00 | 13.00 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 |
| **Weighted Sum** | | 3.00 | 5.00 | 9.00 | 7.00 | 7.00 | 35/3 = 11.67 | 65/5 = 13.00 |
| **Constraints** | | 35.00 | 65.00 | 91.00 | 65.00 | 104.00 | | |
| $\delta_0$ | | 0.9143 | 0.9231 | 0.9011 | 0.8923 | 0.9327 | | |
| **Weighted Sum 1** | | ***35.00*** | 5.00 | 51.67 | 28.33 | 28.33 | | |
| **Weighted Sum 2** | | 35.00 | ***65.00*** | 111.67 | 88.33 | 88.33 | | |

# Illustration of IPU Algorithm (continued)

## Adjustment with respect to person type constraints

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Weights 1 | Weights 2 | Weights 3 | Weights 4 | Weights 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 11.67 | 11.67 | 9.51 | 8.05 | 12.37 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 11.67 | 11.67 | 9.51 | 9.51 | 14.61 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | 11.67 | 11.67 | 9.51 | 8.05 | 8.05 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | 1.00 | 13.00 | 10.59 | 10.59 | 16.28 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | 1.00 | 13.00 | 13.00 | 11.00 | 16.91 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 10.59 | 8.97 | 8.97 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | 1.00 | 13.00 | 10.59 | 8.97 | 13.78 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 10.59 | 8.97 | 8.97 |
| Weighted Sum | | 3 | 5 | 9 | 7 | 7 | | | | | |
| Constraints | | 35 | 65 | 91 | 65 | 104 | 35/3 = 11.67 | 65/5 = 13.00 | 91/111.67 = 0.81 | 65/76.80 = 0.85 | 104/67.68 = 1.54 |
| $\delta_0$ | | 0.9143 | 0.9231 | 0.9011 | 0.8923 | 0.9327 | | | | | |
| Weighted Sum 1 | | *35.00* | 5.00 | 51.67 | 28.33 | 28.33 | | | | | |
| Weighted Sum 2 | | 35.00 | *65.00* | 111.67 | 88.33 | 88.33 | | | | | |
| Weighted Sum 3 | | 28.52 | 55.38 | *91.00* | 76.80 | 74.39 | | | | | |
| Weighted Sum 4 | | 25.60 | 48.50 | 80.11 | *65.00* | 67.68 | | | | | |
| Weighted Sum 5 | | 35.02 | 64.90 | 104.84 | 85.94 | *104.00* | | | | | |
| $\delta_1$ | | 0.0006 | 0.0015 | 0.1521 | 0.3222 | 0 | | | | | |

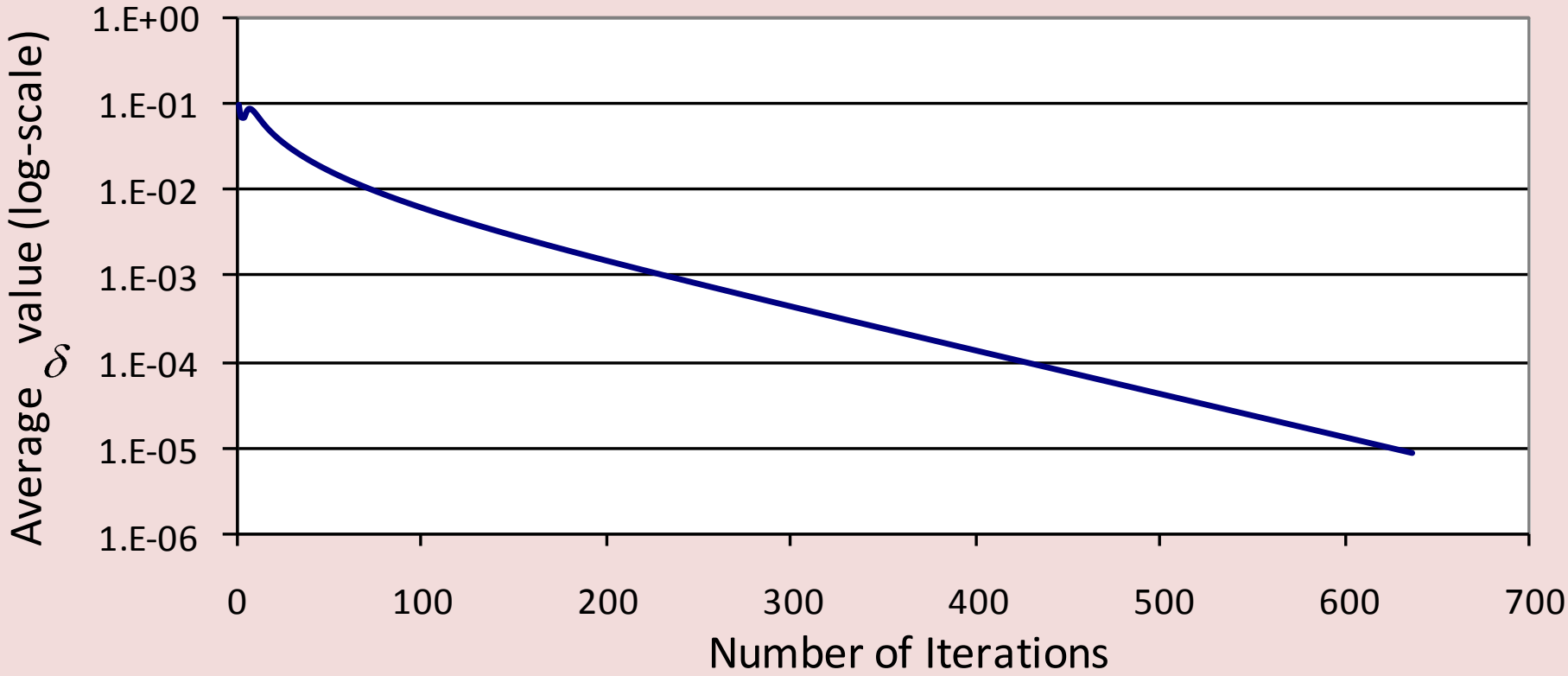
# Illustration of IPU Algorithm (continued)

## Final Results

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | IPU (New) Weights |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1.36 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 25.66 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | 7.98 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | 27.79 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | 18.45 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 8.64 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | 1.47 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | 8.64 |
| **Constraints** | | 35.00 | 65.00 | 91.00 | 65.00 | 104.00 | |
| $\delta_0$ | | 0.9143 | 0.9231 | 0.9011 | 0.8923 | 0.9327 | |
| $\delta_{IPU}$ | | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

# Illustration of IPU Algorithm (continued)

## Improvement in Average $\delta$ Value



# IPU: Geometric Interpretation

- Consider the following household structure and population constraints

| Household ID | Household Type 1 | Person Type 1 | Weights |
|---|---|---|---|
| 1 | 1 | 0 | $w_1$ |
| 2 | 1 | 1 | $w_2$ |
| ***Constraints*** | ***4*** | ***3*** | |

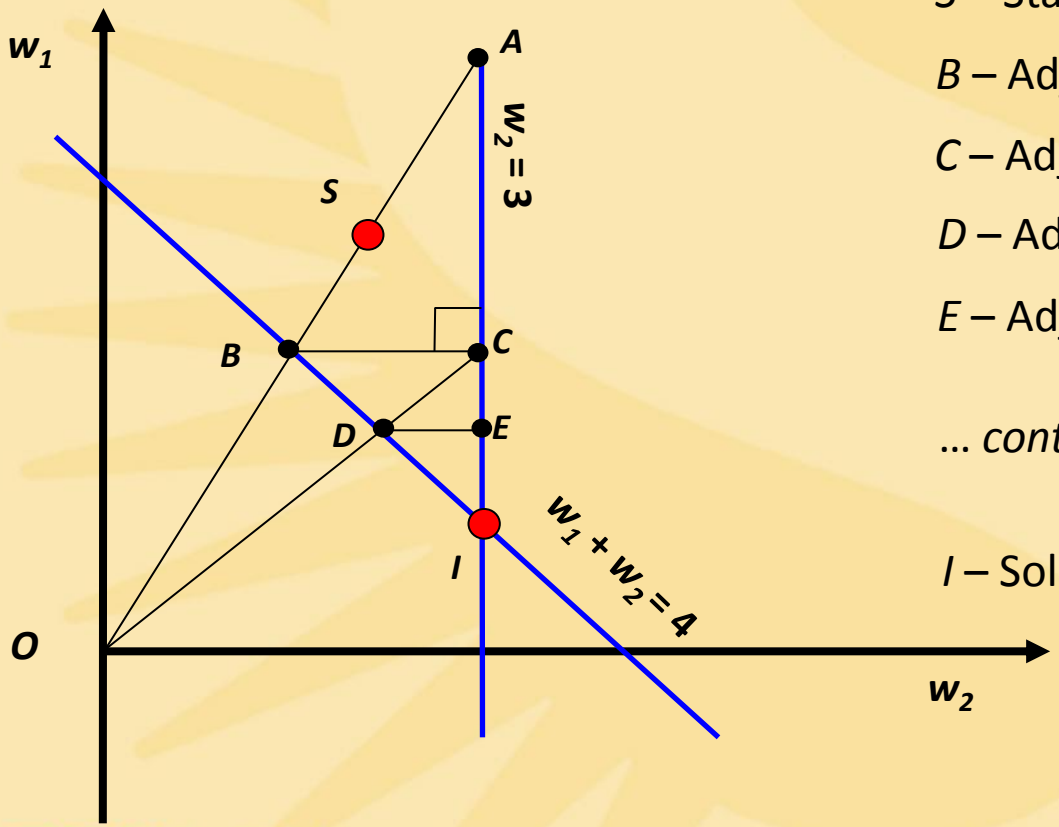
- Weights can be estimated by solving the following system of linear equations

$$w_1 + w_2 = 4$$

$$w_2 = 3$$

# IPU: Geometric Interpretation (continued)

## When solution is **within** the feasible region




*S* – Starting Point

*B* – Adjustment for Household Constraint

*C* – Adjustment for Person Constraint

*D* – Adjustment for Household Constraint

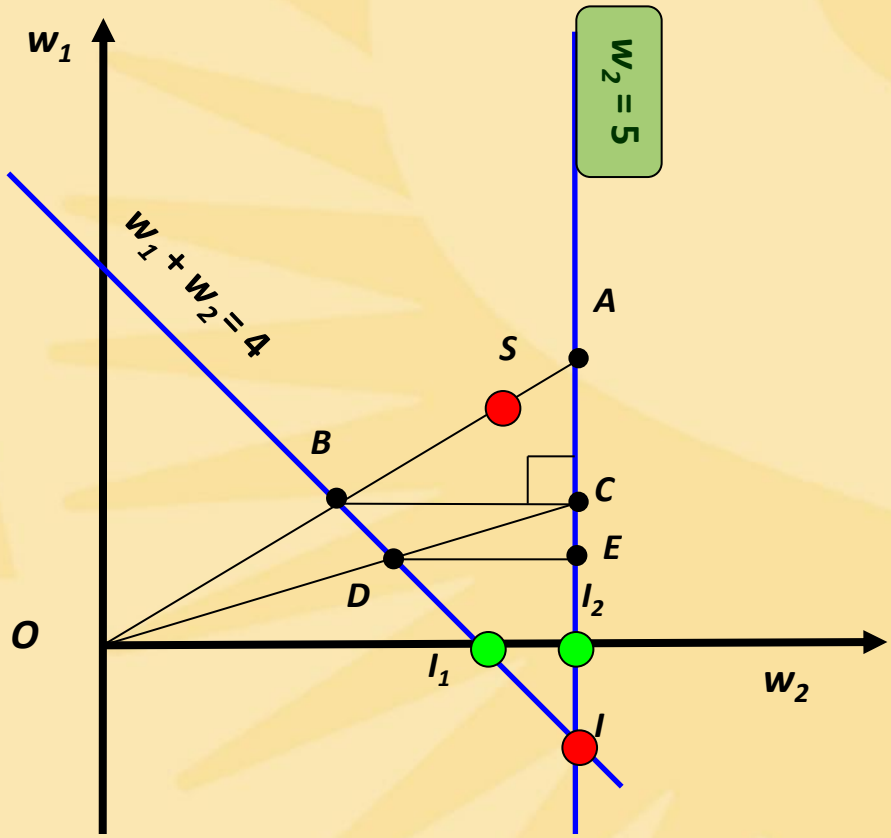
*E* – Adjustment for Person Constraint

*… continue to convergence*

*I* – Solution

# IPU: Geometric Interpretation (continued)

## When solution is **outside** the feasible region




*S* – Starting Point

*B* – Adjustment for household constraint

*C* – Adjustment for person constraint

*D* – Adjustment for household constraint

*E* – Adjustment for person constraint

*… continue to convergence*

*I* – Solution outside feasible region

$I_1$ – Corner solution where household constraint is satisfied

$I_2$ – Corner solution where person constraint is satisfied

# A Test Application of PopGen

- Test area – Maricopa County, Arizona
- Population estimates from Census 2000
  - 3,071,219 individuals
  - 1,133,048 households and 44,689 group quarters
  - 2,090 blockgroups
- Sample household and person data obtained from 2000 PUMS
  - 254,205 individuals
  - 95,066 households
  - 5,489 groupquarters
- Marginal distributions of attributes obtained from 2000 Census Summary Files
- Synthetic population generated at level of blockgroup

# Test Application: Control Variables

## Household Attributes

- Household Type (5 categories)

  1) *Family: Married Couple;* 2) *Family: Male Householder, No Wife;* 3) *Family: Female Householder, No Husband;* 4) *Non-family: Householder Alone;* 5) *Non-family: Householder Not Alone*

- Household Size (7 categories)

  1) *1 Person;* 2) *2 Persons;* 3) *3 Persons;* 4) *4 Persons;* 5) *5 Persons;* 6) *6 Persons;* 7) *7 or more Persons*

- Household Income (8 categories)

  1) *$0 - $14,999;* 2) *$15,000 - $24,999;* 3) *$25,000 - $34,999;* 4) *$35,000 - $44,999;* 5) *$45,000 - $59,999;* 6) *$60,000 - $99,999;* 7) *$100,000 - $149,999;* 8) *Over $150,000*

- Presence of Own Children (2 categories)

  *1) Yes; 2) No*

- ***560 household type constraints***

# Test Application: Control Variables (continued)

## Person Attributes

- Gender (2 categories)

  1) *Male;* 2) *Female*

- Age (10 categories)

  1) *Under 5 years;* 2) *5 to 14 years;* 3) *15 to 24 years;* 4) *25 to 34 years;* 5) *35 to 44 years;* 6) *45 to 54 years;* 7) *55 to 64 years;* 8) *65 to 74 years;* 9) *75 to 84 years;* 10) *85 and more*

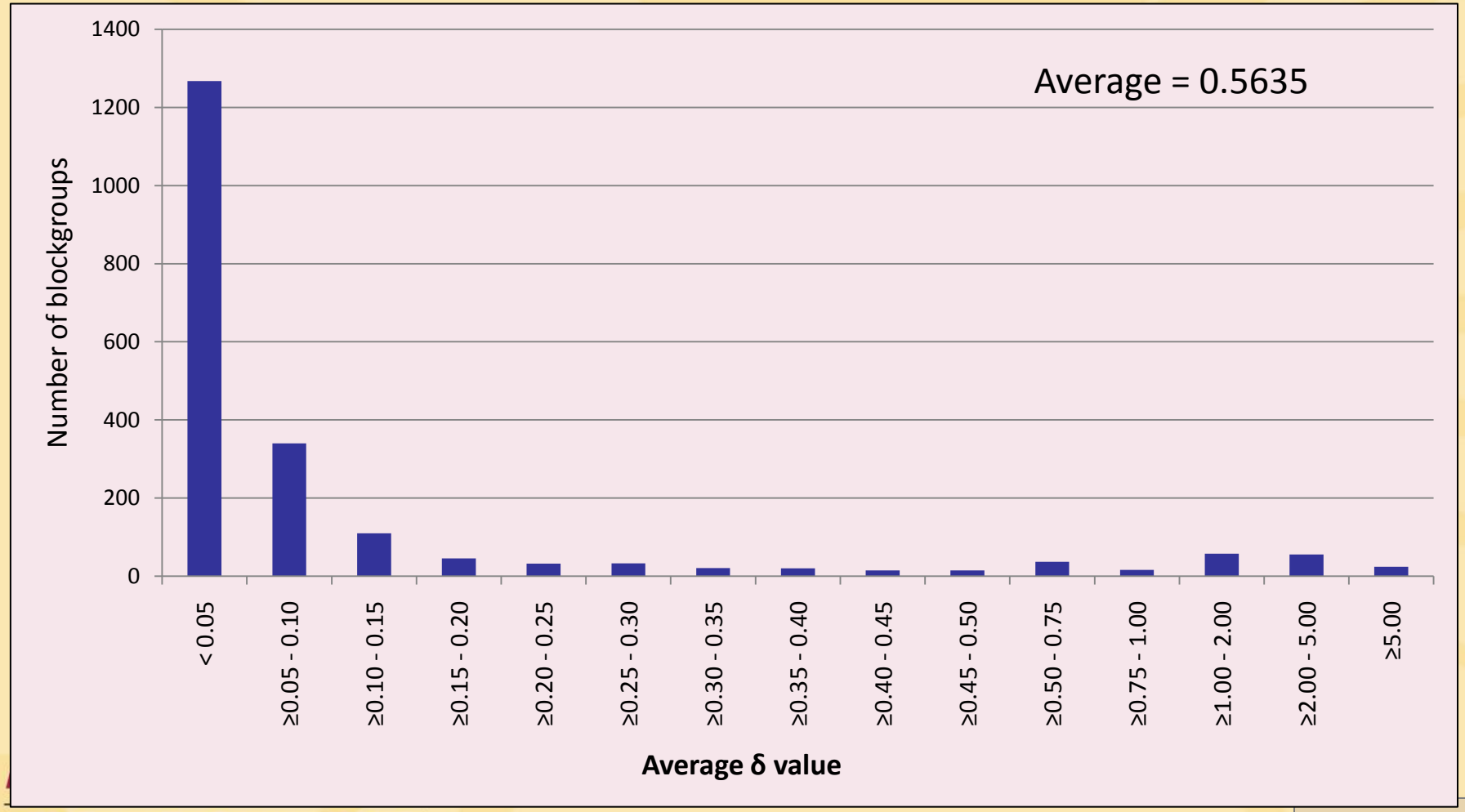
- Ethnicity (7 categories)

  1) *White alone;* 2) *Black or African American alone;* 3) *American Indian and Alaska Native alone;* 4) *Asian alone;* 5) *Native Hawaiian and Other Pacific Islander alone;* 6) *Some other race alone;* 7) *Two or more races*

- ***140 person type constraints***

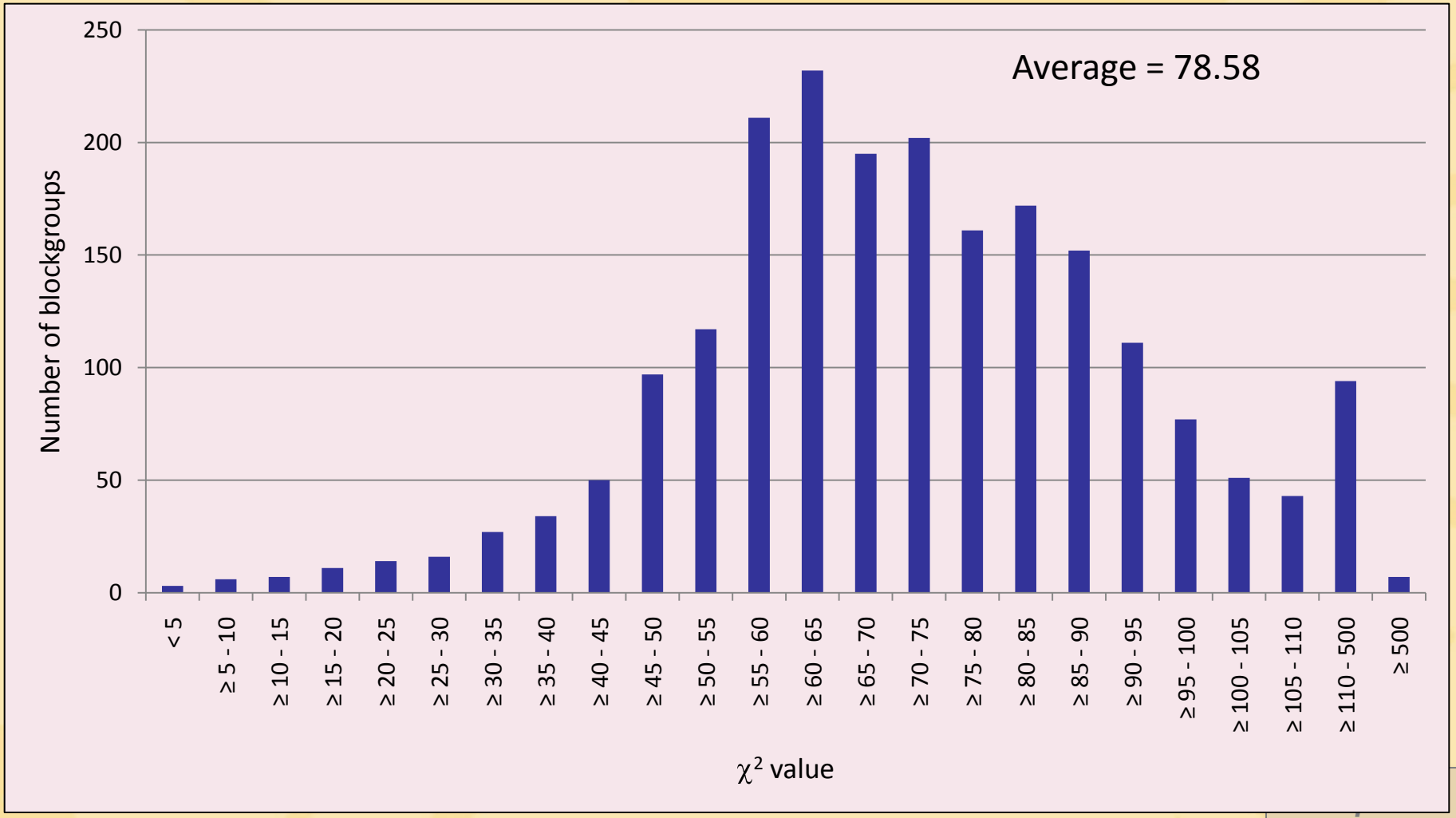
# Test Application: Regional Results

## Distribution of Average $\delta$ Value



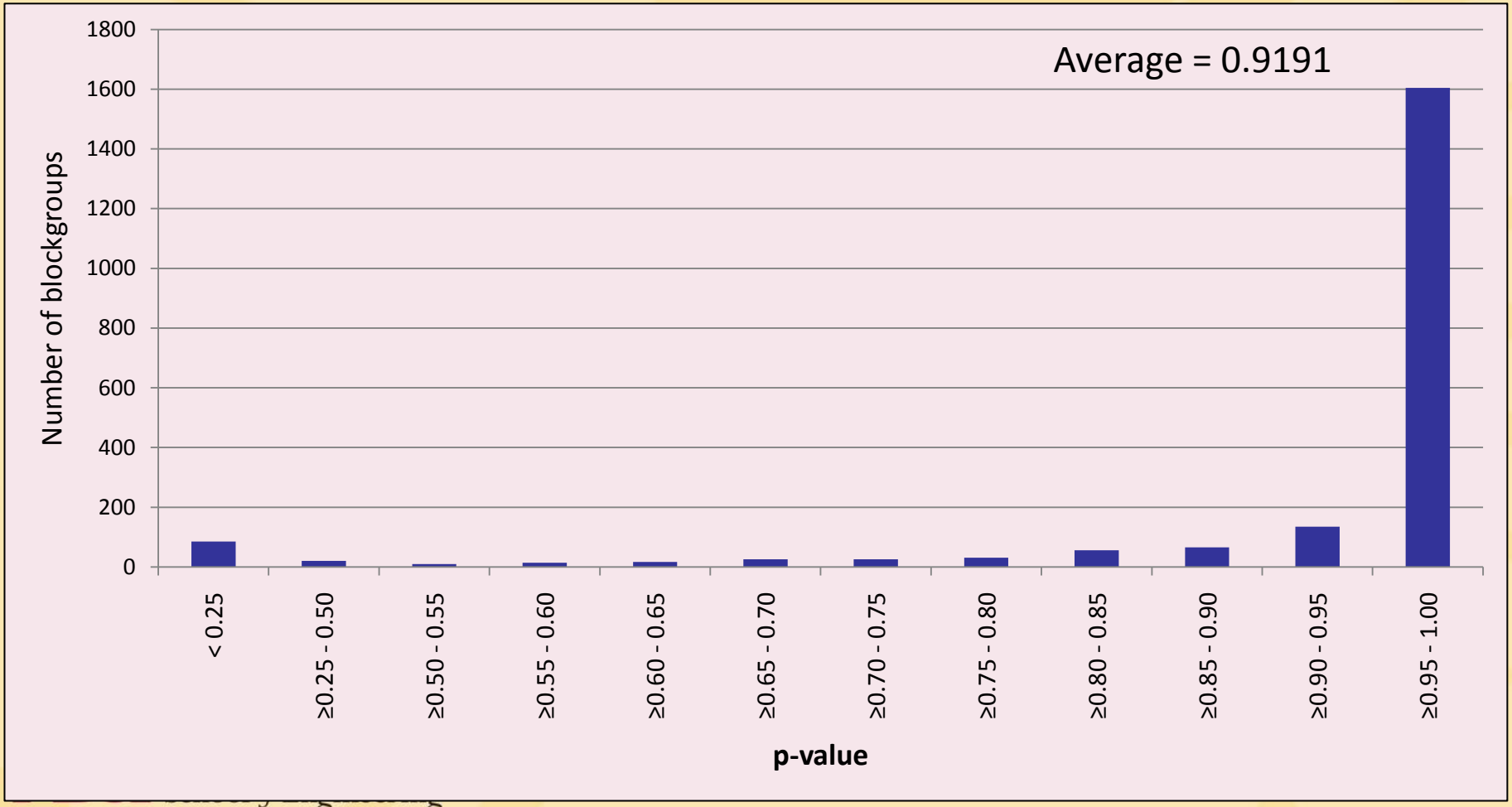
# Test Application: Regional Results (continued)

## Distribution of $\chi^2$-value



# Test Application: Regional Results(continued)

## Distribution of p-value

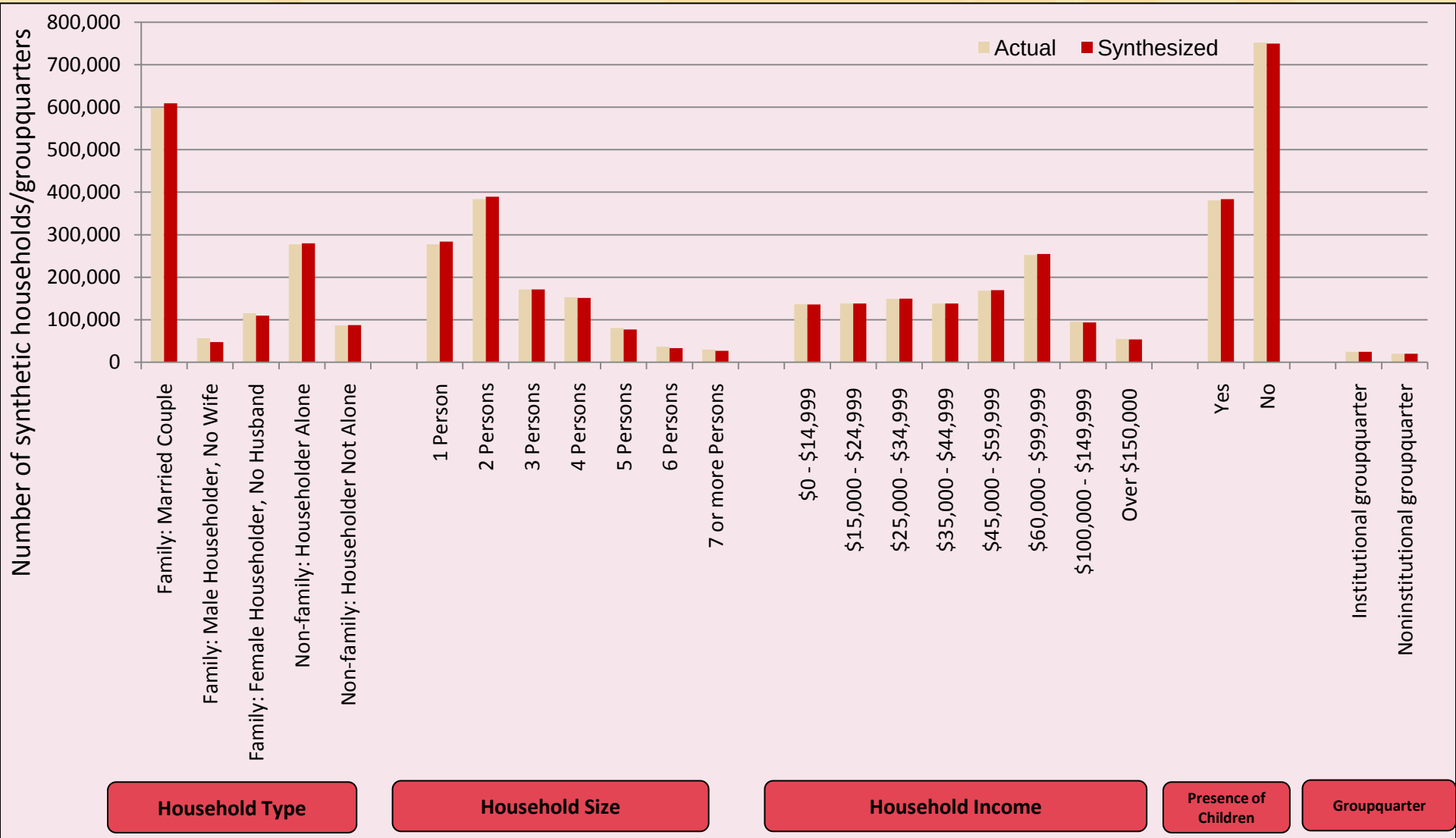


# Test Application: Regional Results(continued)

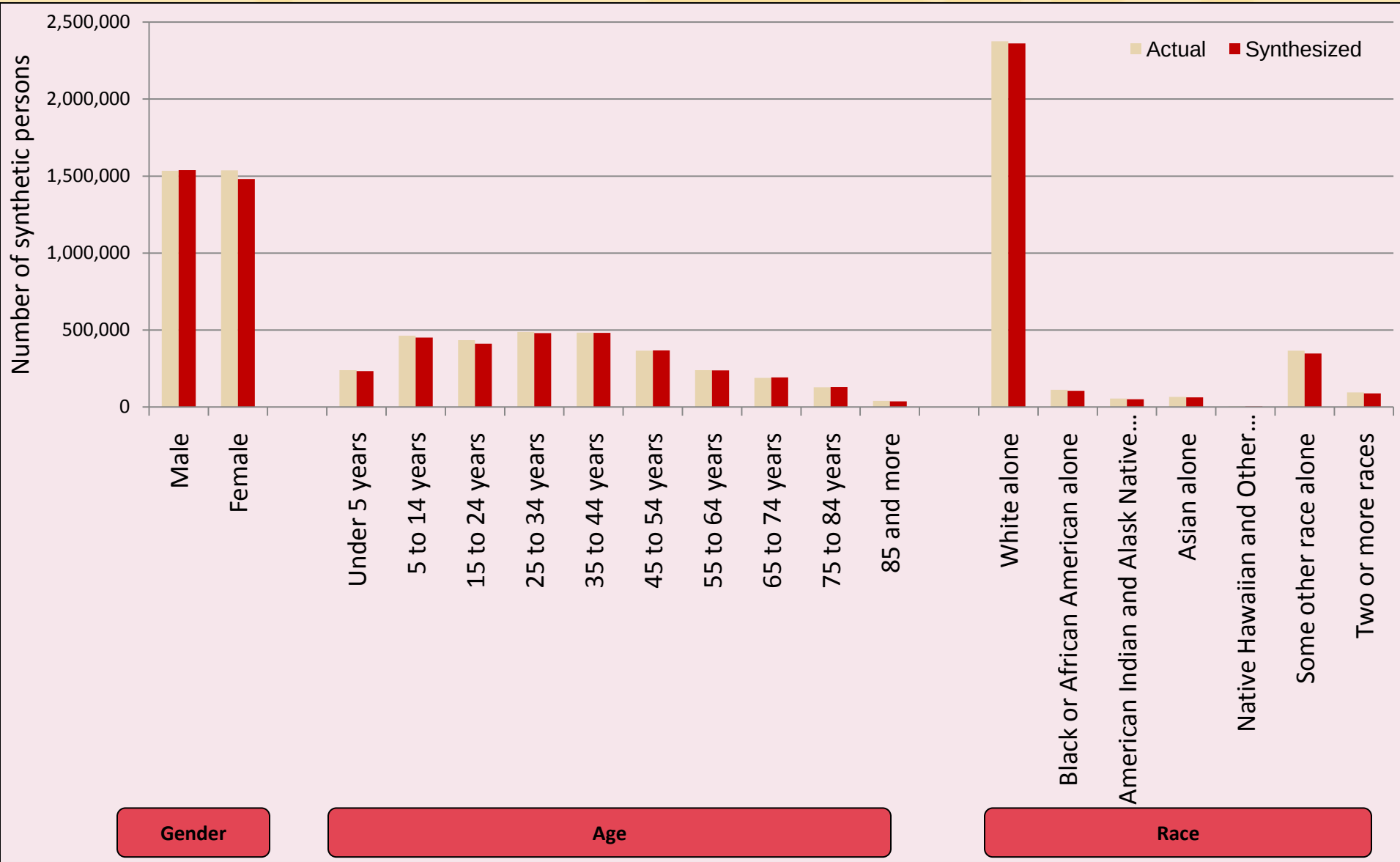
## Comparison of estimated and actual frequencies

| | ***Estimated*** | ***Actual*** |
|---|---|---|
| Households | 1,133,048 | 1,133,048 |
| Groupquarters | 44,689 | 44,689 |
| *Total* | *1,177,737* | *1,177,737* |
| | | |
| Persons | 3,020,695 | 3,072,149 |

# Test Application: Regional Results (continued)



# Test Application: Regional Results (continued)



# Test Application: Regional Results (continued)

- Computational Performance
  - Dell Precision T5400, quad core machine with Intel Xeon Processors and 4 GB of RAM
  - Average **processing** time per blockgroup – 32 seconds
  - Average **run** time per blockgroup using a parallel version of the code – 8 seconds
  - Total **processing** time for 2090 blockgroups – approximately 18 hours and 35 minutes
  - Total **run** time for 2090 blockgroups – approximately 4 hours and 40 minutes

# Test Application: Sample Results

## Results for two illustrative block groups

### Blockgroup A

County – Maricopa

Tract ID – 111602

Blockgroup ID – 5

*Near Perfect Solution Reached*

### Blockgroup B

County – Maricopa

Tract ID – 104203

Blockgroup ID – 2

*Corner Solution Reached*

# Test Application: Sample Results (continued)

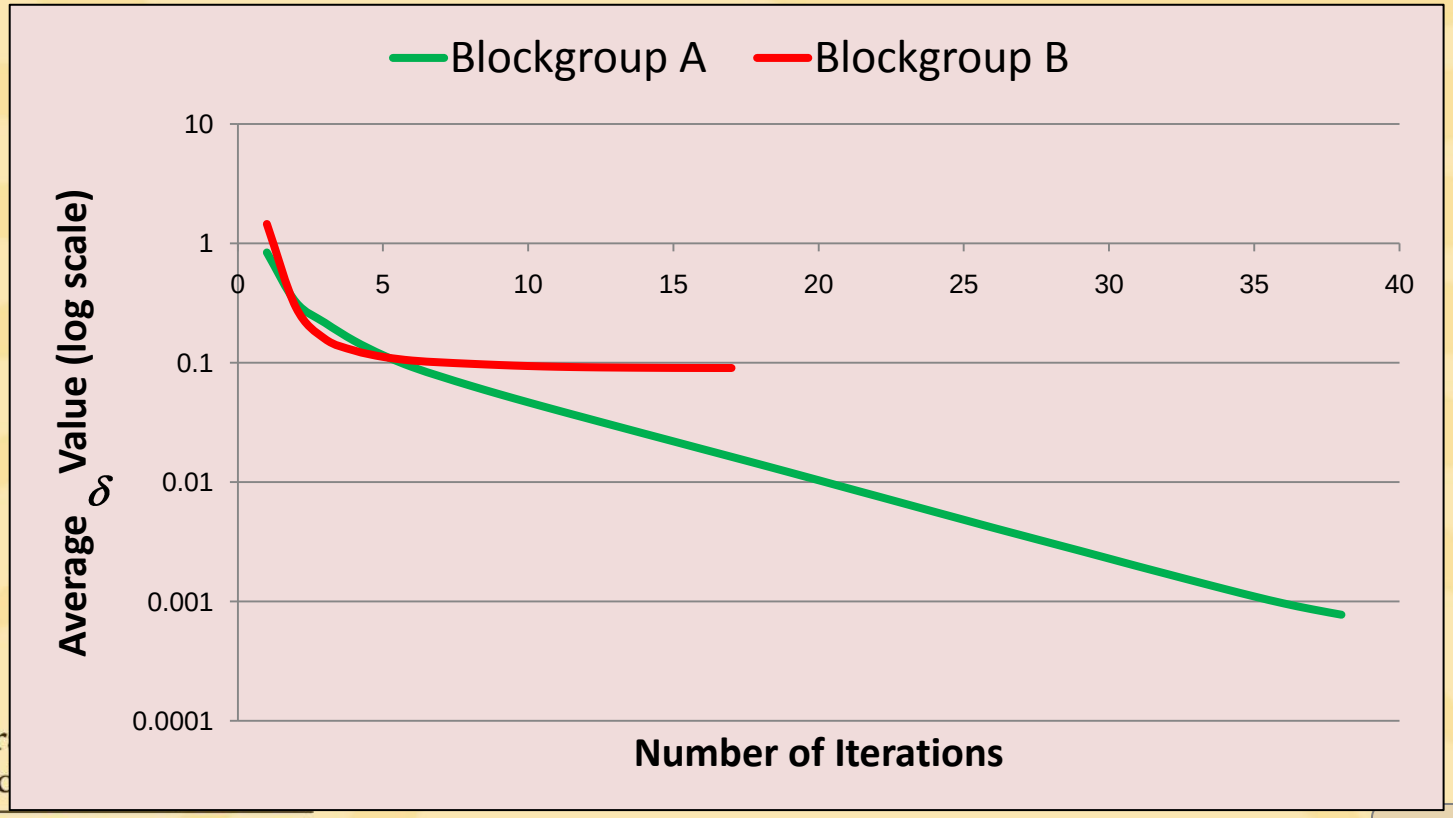
## Reduction in Average Absolute Relative Difference ($\delta$ value)

**Blockgroup A**

$\delta$ 0.8385 → 0.0008 in 38 iterations

**Blockgroup B**

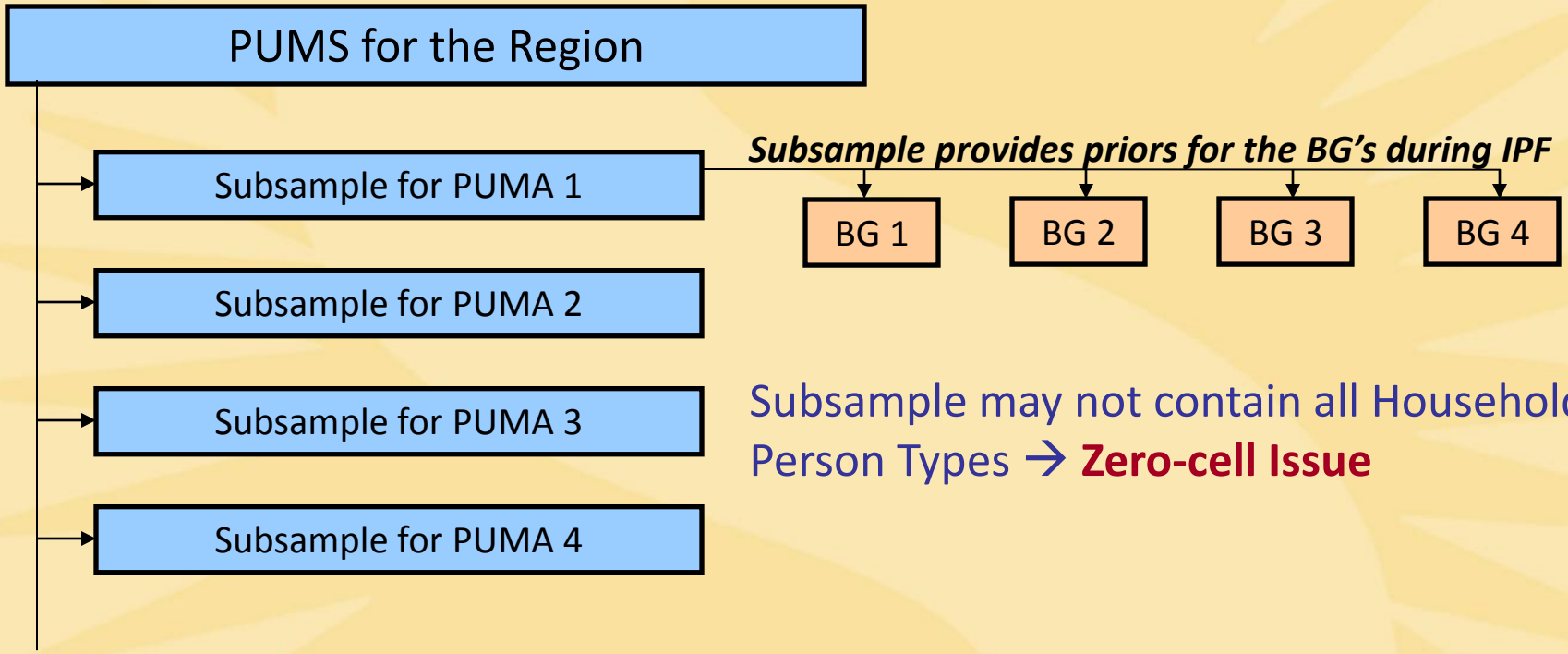
$\delta$ 1.4502 → 0.0903 in 17 iterations



# Small Geographies: Zero-Cell Correction

- Zero-cell Issue
  - The seed matrix from the sub-region (PUMA) to which the small geography belongs does not include infrequent household types
  - IPF for the geography may fail to converge
- Typical Approach
  - Add a small arbitrary number to the zero-cells (Beckman et al 1996)
  - This procedure introduces an arbitrary bias (Guo and Bhat, 2007)
- Solution Implemented in PopGen
  - ***Borrow prior information for the zero cells from the PUMS data for the entire region subject to an upper limit***

# Small Geographies: Zero-Cell Correction (continued)




Subsample may not contain all Household/ Person Types → **Zero-cell Issue**

# Small Geographies: Zero-Cell Correction (continued)

## Priors from PUMA

| | | Household Income | |
|---|---|---|---|
| | | High | Low |
| Household Size Category | 1 | 3 | 0 |
| | 2 | 2 | 4 |
| | 3 or more | 2 | 1 |
| | | | *Total* 12 |

## Priors from Total PUMS

| | | Household Income | |
|---|---|---|---|
| | | High | Low |
| Household Size Category | 1 | 7 | 2 |
| | 2 | 8 | 10 |
| | 3 or more | 3 | 3 |
| | | | *Total* 33 |

## Probabilities from PUMA

| | | Household Income | |
|---|---|---|---|
| | | High | Low |
| Household Size Category | 1 | 0.25 | 0.00 |
| | 2 | 0.17 | 0.33 |
| | 3 or more | 0.17 | 0.08 |

Threshold Probability = 1/12 = 0.083

## Probabilities from Total PUMS

| | | Household Income | |
|---|---|---|---|
| | | High | Low |
| Household Size Category | 1 | 0.21 | 0.06 |
| | 2 | 0.24 | 0.30 |
| | 3 or more | 0.09 | 0.09 |

# Small Geographies: Zero-Cell Correction (continued)

## Zero-cell Adjustment

| | | Household Income | |
|---|---|---|---|
| | | High | Low |
| Household Size Category | 1 | 0.25 | 0.06 |
| | 2 | 0.17 | 0.33 |
| | 3 or more | 0.17 | 0.08 |

Probability sum adds up to more than 1.00 (1.06)
→ adjust probabilities for other cells

## Zero-cell Adjustment

| | | Household Income | |
|---|---|---|---|
| | | High | Low |
| Household Size Category | 1 | 0.25 x 0.94 | 0.06 |
| | 2 | 0.17 x 0.94 | 0.33 x 0.94 |
| | 3 or more | 0.17 x 0.94 | 0.08 x 0.94 |

Adjustment factor = (1.00 - 0.06)

= 0.94

## Adjusted priors

| | | Household Income | |
|---|---|---|---|
| | | High | Low |
| Household Size Category | 1 | 0.23 | 0.06 |
| | 2 | 0.16 | 0.31 |
| | 3 or more | 0.16 | 0.08 |

# Small Geographies: Zero-Marginal Correction

- Issue
  - The marginal values for certain categories of an attribute take a zero value
  - IPF procedure will assign a zero to all household/person type cells that comprise the zero-marginal category
  - As a result the IPU algorithm may fail to proceed
- Solution implemented in PopGen
  - Add a small value (0.001) to the zero-marginal categories
  - IPU algorithm now proceeds to compute weights
  - Effect of this small value on results is negligible

# Small Geographies: Zero-Marginal Correction (continued)

## Iteration 1 of IPU algorithm without correction

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Weights 1 | Weights 2 | Weights 3 | Weights 4 | Weights 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 11.67 | 11.67 | 0.00 | 0.00 | 0.00 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 11.67 | 11.67 | 0.00 | 0.00 | 0.00 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | 11.67 | 11.67 | 0.00 | 0.00 | 0.00 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | 1.00 | 13.00 | 13.00 | 55.00 | 150.00 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| Weighted Sum | | 3 | 5 | 9 | 7 | 7 | | | | | |
| Constraints | | 35 | 65 | 0 | 110 | 150 | 35/3 = 11.67 | 65/5 = 13.00 | 0/111.67 = 0.00 | 110/26 = 4.23 | 150/55 = 2.73 |
| $\delta_0$ | | 0.9143 | 0.9231 | - | 0.9364 | 0.9533 | | | | | |
| Weighted Sum 1 | | *35.00* | 5.00 | 51.67 | 28.33 | 28.33 | | | | | |
| Weighted Sum 2 | | 35.00 | *65.00* | 111.67 | 88.33 | 88.33 | | | | | |
| Weighted Sum 3 | | 0.00 | 13.00 | *0.00* | 26.00 | 13.00 | | | | | |
| Weighted Sum 4 | | 0.00 | 55.00 | 0.00 | *110.00* | 55.00 | | | | | |
| Weighted Sum 5 | | 0.00 | 150.00 | 0.00 | 300.00 | *150.00* | | | | | |
| $\delta_1$ | | 1.0000 | 1.3077 | - | 1.7273 | 0.0000 | | | | | |

**Adjustment**

# Small Geographies: Zero-Marginal Correction (continued)

## Iteration 2 of IPU algorithm without correction

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Weights 1 | Weights 2 | Weights 3 | Weights 4 | Weights 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | | | | | |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | | | | | |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | | | | | |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | | | | | |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | | | | | |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | | | | | |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | | | | | |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | | | | | |
| Weighted Sum | | 3 | 5 | 9 | 7 | 7 | | | | | |
| Constraints | | 35 | 65 | 0 | 110 | 150 | 35/0 = undefined | | | | |
| $\delta_1$ | | 1.0000 | 1.3077 | - | 1.7273 | 0.0000 | | | | | |
| Weighted Sum 1 | | | | | | | | | | | |
| Weighted Sum 2 | | | | | | | | | | | |
| Weighted Sum 3 | | | | | | | | | | | |
| Weighted Sum 4 | | | | | | | | | | | |
| Weighted Sum 5 | | | | | | | | | | | |
| $\delta_2$ | | | | | | | | | | | |

Adjustment

# Small Geographies: Zero-Marginal Correction (continued)

## Iteration 1 of IPU algorithm with correction

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Weights 1 | Weights 2 | Weights 3 | Weights 4 | Weights 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 11.67 | 11.67 | 0.00 | 0.00 | 0.00 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 11.67 | 11.67 | 0.00 | 0.00 | 0.00 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | 11.67 | 11.67 | 0.00 | 0.00 | 0.00 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | 1.00 | 13.00 | 13.00 | 55.00 | 150.00 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 0.00 | 0.00 | 0.00 |
| Weighted Sum | | 3 | 5 | 9 | 7 | 7 | | | | | |
| Constraints | | 35 | 65 | 0.001 | 110 | 150 | | | | | |
| $\delta_0$ | | 0.9143 | 0.9231 | 0.9011 | 0.8923 | 0.9327 | | | | | |
| Weighted Sum 1 | | *35.0000* | 5.0000 | 51.6700 | 28.3300 | 28.3300 | | | | | |
| Weighted Sum 2 | | 35.0000 | *65.0000* | 111.6700 | 88.3300 | 88.3300 | | | | | |
| Weighted Sum 3 | | 0.0003 | 13.0005 | *0.0010* | 26.0006 | 13.0007 | | | | | |
| Weighted Sum 4 | | 0.0010 | 55.0004 | 0.0035 | *110.0000* | 55.0006 | | | | | |
| Weighted Sum 5 | | 0.0019 | 149.9978 | 0.0064 | 299.9944 | *150.0000* | | | | | |
| $\delta_1$ | | 0.9999 | 1.3077 | 5.3619 | 1.7272 | 0.0000 | | | | | |

# Small Geographies: Zero-Marginal Correction (continued)

## Iteration 2 of IPU algorithm with correction

| Household ID | Initial Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Weights 1 | Weights 2 | Weights 3 | Weights 4 | Weights 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 21.83 | 21.83 | 0.00 | 0.00 | 0.00 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 5.16 | 5.16 | 0.00 | 0.00 | 0.00 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | 8.01 | 8.01 | 0.00 | 0.00 | 0.00 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | 150.00 | 65.00 | 65.00 | 55.00 | 150.00 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Weighted Sum | | 3 | 5 | 9 | 7 | 7 | | | | | |
| Constraints | | 35 | 65 | 0.001 | 110 | 150 | | | | | |
| $\delta_1$ | | 0.9999 | 1.3077 | 5.3619 | 1.7272 | 0.0000 | | | | | |
| Weighted Sum 1 | | *35.0000* | 149.9978 | 43.0097 | 329.8319 | 176.9928 | | | | | |
| Weighted Sum 2 | | 35.0000 | *65.0000* | 43.0074 | 159.8380 | 91.9946 | | | | | |
| Weighted Sum 3 | | 0.0008 | 64.9989 | *0.0010* | 129.9984 | 64.9995 | | | | | |
| Weighted Sum 4 | | 0.0007 | 54.9997 | 0.0009 | *110.0000* | 55.0003 | | | | | |
| Weighted Sum 5 | | 0.0017 | 149.9985 | 0.0018 | 299.9983 | *150.0000* | | | | | |
| $\delta_2$ | | 1.0000 | 1.3077 | 0.8139 | 1.7273 | 0.0000 | | | | | |

# Person Total Inconsistency

- Issue
  - The person total range derived from the household size distribution is not consistent with the given person total
  - Results in a corner solution and the estimated weights do not match the person-type constraints
  - Synthetic population generated doesn't match given person totals
- Solution implemented in PopGen
  - Adjust the household marginal distributions

# Person Total Inconsistency (continued)

For household size category 3

- *p_tot_min calculated using a value of 3*
- *p_tot_max calculated using a value of 8*

| hhldsize1 | hhldsize2 | hhldsize3 | p_tot_min | p_tot_max | given_p_tot |
|---|---|---|---|---|---|
| 1026 | 816 | 359 | 3735 | 5530 | 3503 |
| 443 | 539 | 212 | 2157 | 3217 | 3612 |
| 773 | 679 | 235 | 2836 | 4011 | 4321 |
| 323 | 412 | 204 | 1759 | 2779 | 1523 |

# Person Total Inconsistency (continued)

## Step 1: Calculate the person total difference

*Average household size for category 3 – 3.7*

| hhldsize1 | hhldsize2 | hhldsize3 | p_tot_eq | given_p_tot | p_diff |
|---|---|---|---|---|---|
| 1026 | 816 | 359 | 3986.3 | 3503 | -483.3 |
| 443 | 539 | 212 | 2305.4 | 3612 | 1306.6 |
| 773 | 679 | 235 | 3000.5 | 4321 | 1320.5 |
| 323 | 412 | 204 | 1901.8 | 1523 | -378.8 |

# Person Total Inconsistency (continued)

## Step 2: Calculate the corresponding household difference

*hhld_diff = p_diff/(phhldsize1 + phhldsize2 * 2 + phhldsize3 * 3.7)*

| phhldsize1 | phhldsize2 | phhldsize3 | p_diff | hhld_diff |
|---|---|---|---|---|
| 0.47 | 0.37 | 0.16 | -483.30 | -266.85 |
| 0.37 | 0.45 | 0.18 | 1306.60 | 676.71 |
| 0.46 | 0.40 | 0.14 | 1320.50 | 742.44 |
| 0.34 | 0.44 | 0.22 | -378.80 | -187.03 |

# Person Total Inconsistency (continued)

## Step 3: Revise the household marginals

*rhhldsize1 = hhldsize1 + phhldsize1 * hhld_diff*

| rhhldsize1 | rhhldsize2 | rhhldsize3 | rp_tot |
|---|---|---|---|
| 901.61 | 717.07 | 315.47 | 3503.00 |
| 694.07 | 844.48 | 332.15 | 3612.00 |
| 1113.19 | 977.82 | 338.42 | 4321.00 |
| 258.66 | 329.94 | 163.37 | 1523.00 |

# Test Application: Synthetic Population

- Synthetic population generation process can be divided into three steps
  - Estimating whole frequencies
  - Calculating selection probabilities
  - Drawing households

# Test Application: Estimating Frequencies

- IPF-estimated household type constraints provide target frequencies
  - Rounding procedures are employed to convert decimal values to whole frequencies
- Rounding procedures implemented in PopGen
  - Arithmetic Rounding (default)
  - Bucket Rounding
  - Stochastic Rounding

# Test Application: Estimating Frequencies (continued)

- Arithmetic Rounding Procedure
  - Round the decimal frequencies
  - Account for the difference between the rounded frequency sum and the actual frequency sum

# Test Application: Estimating Frequencies (continued)

## Illustration of Arithmetic Rounding Procedure

| Household Type | Frequency | Rounded Frequency | Difference | Ranking to Receive a Household | Adjustment | Adjusted Frequency |
|---|---|---|---|---|---|---|
| 1 | 64.85 | 65 | 0.15 | 16 | | 65 |
| 2 | 12.34 | 12 | -0.34 | 10 | | 12 |
| 3 | 10.36 | 10 | -0.36 | 9 | | 10 |
| 4 | 0.43 | 0 | -0.43 | 5 | 1 | 1 |
| 5 | 0.49 | 0 | -0.49 | 1 | 1 | 1 |
| 6 | 0.47 | 0 | -0.47 | 3 | 1 | 1 |
| 7 | 0.44 | 0 | -0.44 | 4 | 1 | 1 |
| 8 | 0.39 | 0 | -0.39 | 6 | | 0 |
| 9 | 0.48 | 0 | -0.48 | 2 | 1 | 1 |
| 10 | 0.10 | 0 | -0.10 | 15 | | 0 |
| 11 | 0.12 | 0 | -0.12 | 14 | | 0 |
| 12 | 0.20 | 0 | -0.20 | 13 | | 0 |
| 13 | 0.27 | 0 | -0.27 | 12 | | 0 |
| 14 | 0.28 | 0 | -0.28 | 11 | | 0 |
| 15 | 0.38 | 0 | -0.38 | 7 | | 0 |
| 16 | 0.37 | 0 | -0.37 | 8 | | 0 |
| ***Total*** | ***91.97*** | ***87*** | ***-4.97*** | | ***5*** | ***92*** |

# Test Application: Estimating Frequencies (continued)

- Bucket Rounding Procedure
  - The procedure ensures that the rounded frequency sum and the actual frequency sum are the same
  - Keeps track of the accumulated rounding error
  - Accumulated rounding error is used to bias the rounding of the next frequency value

# Test Application: Estimating Frequencies (continued)

## Illustration of Bucket Rounding Procedure

| Household Type | Frequency | Integer Part | Calculations | Accumulated Difference | Adjustment | Adjusted Frequency |
|---|---|---|---|---|---|---|
| 1 | 64.85 | 64 | | 0.85 | 1 | 65 |
| 2 | 12.34 | 12 | *-0.15* + 0.34 | 0.19 | | 12 |
| 3 | 10.36 | 10 | 0.34 + 0.36 | 0.55 | 1 | 11 |
| 4 | 0.43 | 0 | *-0.45* + 0.43 | -0.02 | | 0 |
| 5 | 0.49 | 0 | -0.02 + 0.49 | 0.47 | | 0 |
| 6 | 0.47 | 0 | 0.47 + 0.47 | 0.94 | 1 | 1 |
| 7 | 0.44 | 0 | *-0.06* + 0.44 | 0.38 | | 0 |
| 8 | 0.39 | 0 | 0.38 + 0.39 | 0.77 | 1 | 1 |
| 9 | 0.48 | 0 | *-0.23* + 0.48 | 0.25 | | 0 |
| 10 | 0.10 | 0 | 0.25 + 0.10 | 0.35 | | 0 |
| 11 | 0.12 | 0 | 0.35 + 0.12 | 0.47 | | 0 |
| 12 | 0.20 | 0 | 0.47 + 0.20 | 0.67 | 1 | 1 |
| 13 | 0.27 | 0 | *-0.33* + 0.27 | -0.06 | | 0 |
| 14 | 0.28 | 0 | -0.06 + 0.28 | 0.22 | | 0 |
| 15 | 0.38 | 0 | 0.22 + 0.38 | 0.60 | 1 | 1 |
| 16 | 0.37 | 0 | *-0.40* + 0.37 | -0.03 | | 0 |
| ***Total*** | ***91.97*** | ***92*** | | | | ***92*** |

# Test Application: Estimating Frequencies (continued)

## Stochastic Rounding Procedure

- Frequencies are randomly rounded up or rounded down
- Account for the difference between the rounded frequency sum and the actual frequency sum

1. Consider a household type frequency of 22.41
2. It can be rounded up with a probability of 0.41 and rounded down with a probability of 0.59
3. We randomly draw a number between 0 and 1 to decide which way the frequency gets rounded
   - Say if the random number was 0.20, then $0.00 \leq \mathbf{0.20} \leq 0.41$, so the frequency gets rounded up to 23.00
   - Alternatively if the random number was 0.78, then $0.41 < \mathbf{0.78} \leq 1.00$, so the frequency gets rounded down to 22.00

# Test Application: Estimating Frequencies (continued)

## Illustration of Stochastic Rounding Procedure

| Household Type | Frequency | Rounded Frequency | Difference | Ranking to Receive a Household | Adjustment | Adjusted Frequency |
|---|---|---|---|---|---|---|
| 1 | 64.85 | 64 | -0.85 | 1 | 1 | 65 |
| 2 | 12.34 | 12 | -0.34 | 8 | | 12 |
| 3 | 10.36 | 10 | -0.36 | 7 | | 10 |
| 4 | 0.43 | 0 | -0.43 | 4 | | 0 |
| 5 | 0.49 | 0 | -0.49 | 2 | 1 | 1 |
| 6 | 0.47 | 1 | 0.53 | 13 | | 1 |
| 7 | 0.44 | 1 | 0.56 | 14 | | 1 |
| 8 | 0.39 | 0 | -0.39 | 5 | | 0 |
| 9 | 0.48 | 0 | -0.48 | 3 | | 0 |
| 10 | 0.10 | 0 | -0.10 | 12 | | 0 |
| 11 | 0.12 | 0 | -0.12 | 11 | | 0 |
| 12 | 0.20 | 1 | 0.80 | 16 | | 1 |
| 13 | 0.27 | 0 | -0.27 | 10 | | 0 |
| 14 | 0.28 | 0 | -0.28 | 9 | | 0 |
| 15 | 0.38 | 1 | 0.62 | 15 | | 1 |
| 16 | 0.37 | 0 | -0.37 | 6 | | 0 |
| ***Total*** | ***91.97*** | ***90*** | ***-1.97*** | | ***2*** | ***92*** |

# Test Application: Selection Probabilities

- Synthetic households are drawn probabilistically based on IPU-estimated weights
- Selection probabilities are estimated for each household type that needs to be synthesized
- No additional adjustments to match person constraints are needed
- The individuals from the synthetic households comprise the synthetic population

# Test Application: Selection Probabilities (continued)

| Household ID | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Final Weights | Household Type 1 | | Household Type 2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Cumulative | | Cumulative | |
| | | | | | | | Sum | Probability | Sum | Probability |
| 1 | 1 | 0 | 1 | 1 | 1 | 1.36 | 1.36 | 0.0389 | - | - |
| 2 | 1 | 0 | 1 | 0 | 1 | 25.66 | 27.02 | 0.7720 | - | - |
| 3 | 1 | 0 | 2 | 1 | 0 | 7.98 | 35.00 | 1.0000 | - | - |
| 4 | 0 | 1 | 1 | 0 | 2 | 27.79 | - | - | 27.79 | 0.4276 |
| 5 | 0 | 1 | 0 | 2 | 1 | 18.45 | - | - | 46.24 | 0.7115 |
| 6 | 0 | 1 | 1 | 1 | 0 | 8.64 | - | - | 54.88 | 0.8444 |
| 7 | 0 | 1 | 2 | 1 | 2 | 1.47 | - | - | 56.35 | 0.8671 |
| 8 | 0 | 1 | 1 | 1 | 0 | 8.64 | - | - | 64.99 | 1.0000 |

# Test Application: Drawing Households

- Rounded frequencies and the selection probabilities from earlier steps are used to generate a synthetic population
- For each household type, we use the corresponding selection probabilities to draw households
- The persons in the drawn households comprise the synthetic population for the target year
- As the drawing procedure is probabilistic, the fit of the synthetic population is checked
- The drawing procedure is repeated until a synthetic population with the best fit is obtained

# Test Application: Drawing Households (continued)

| Household ID | Household Type 1 | | Household Type 2 | |
|---|---|---|---|---|
| | Cumulative Sum | Probability | Cumulative Sum | Probability |
| 1 | 1.36 | 0.0389 | - | - |
| 2 | 27.02 | 0.7720 | - | - |
| 3 | 35.00 | 1.0000 | - | - |
| 4 | - | - | 27.79 | 0.4276 |
| 5 | - | - | 46.24 | 0.7115 |
| 6 | - | - | 54.88 | 0.8444 |
| 7 | - | - | 56.35 | 0.8671 |
| 8 | - | - | 64.99 | 1.0000 |
| **Frequency** | **35** | | **65** | |

1. Consider Household Type 1
2. Generate a random number between 0 and 1, e.g. 0.23
3. **0.0389** < *0.23* < **0.7720**
4. Household ID – 2 is added to the synthetic population
5. The process is repeated until 35 households of Household Type 1 are included
6. The process is repeated for Household Type 2

# Test Application: Synthetic Population

- $\chi^2$ goodness-of-fit statistic
  - A goodness-of-fit measure to check match against person-level distributions
  - The corresponding p-value represents the level of confidence at which the synthetic population matches the given constraints
  - A synthetic population is drawn repeatedly until a desired p-value is achieved or a maximum number of draws is reached
  - Maximum number of draws is user specified and dependent on geographic context

$$\chi^2 = \sum_j \left[ \frac{(n_j - c_j)^2}{c_j} \right]$$

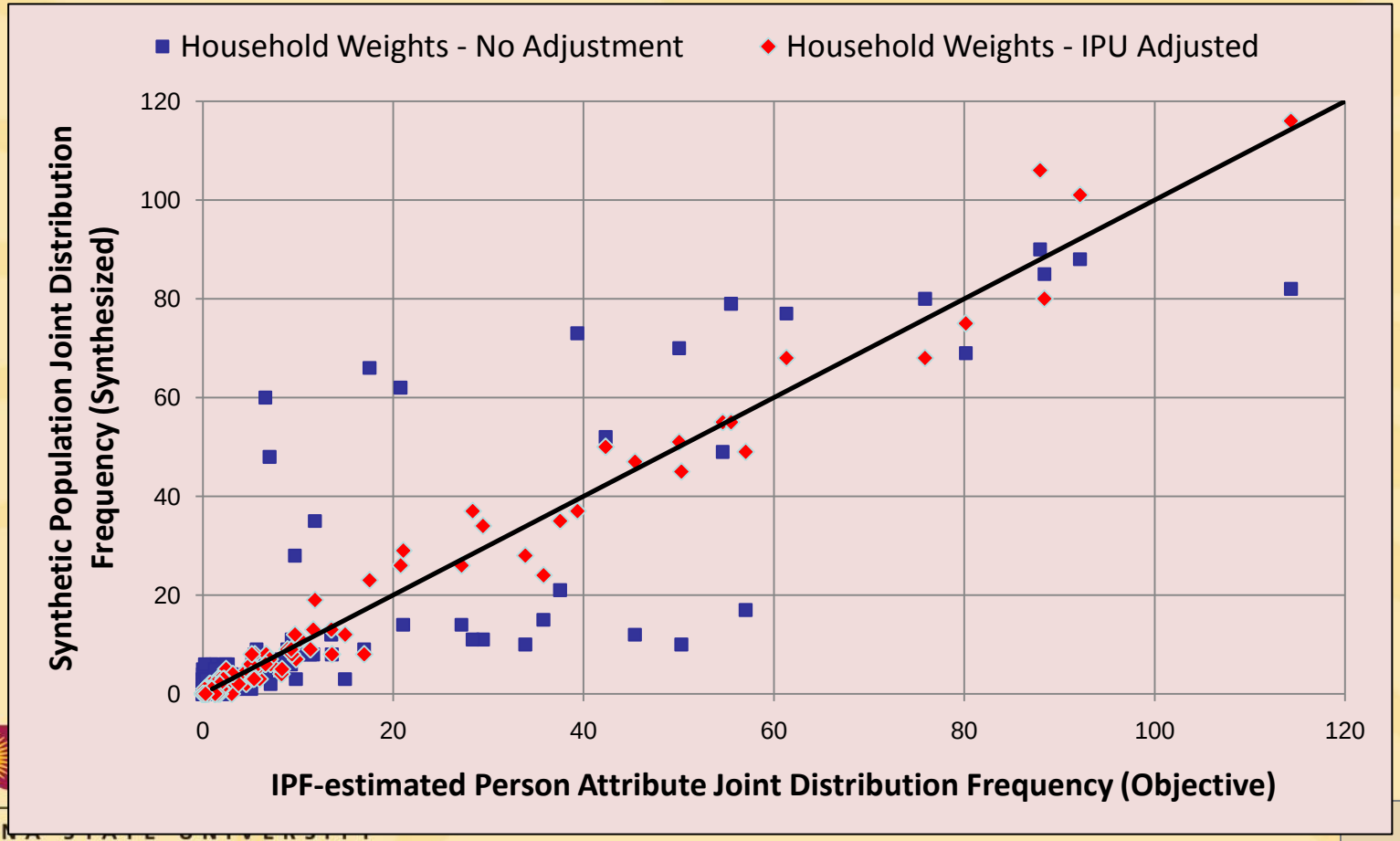
$n_j$ = frequency of synthetic persons of the $j^{th}$ person-type

$c_j$ = the $j^{th}$ IPF-estimated person-type constraint

# Test Application: Performance

Blockgroup A

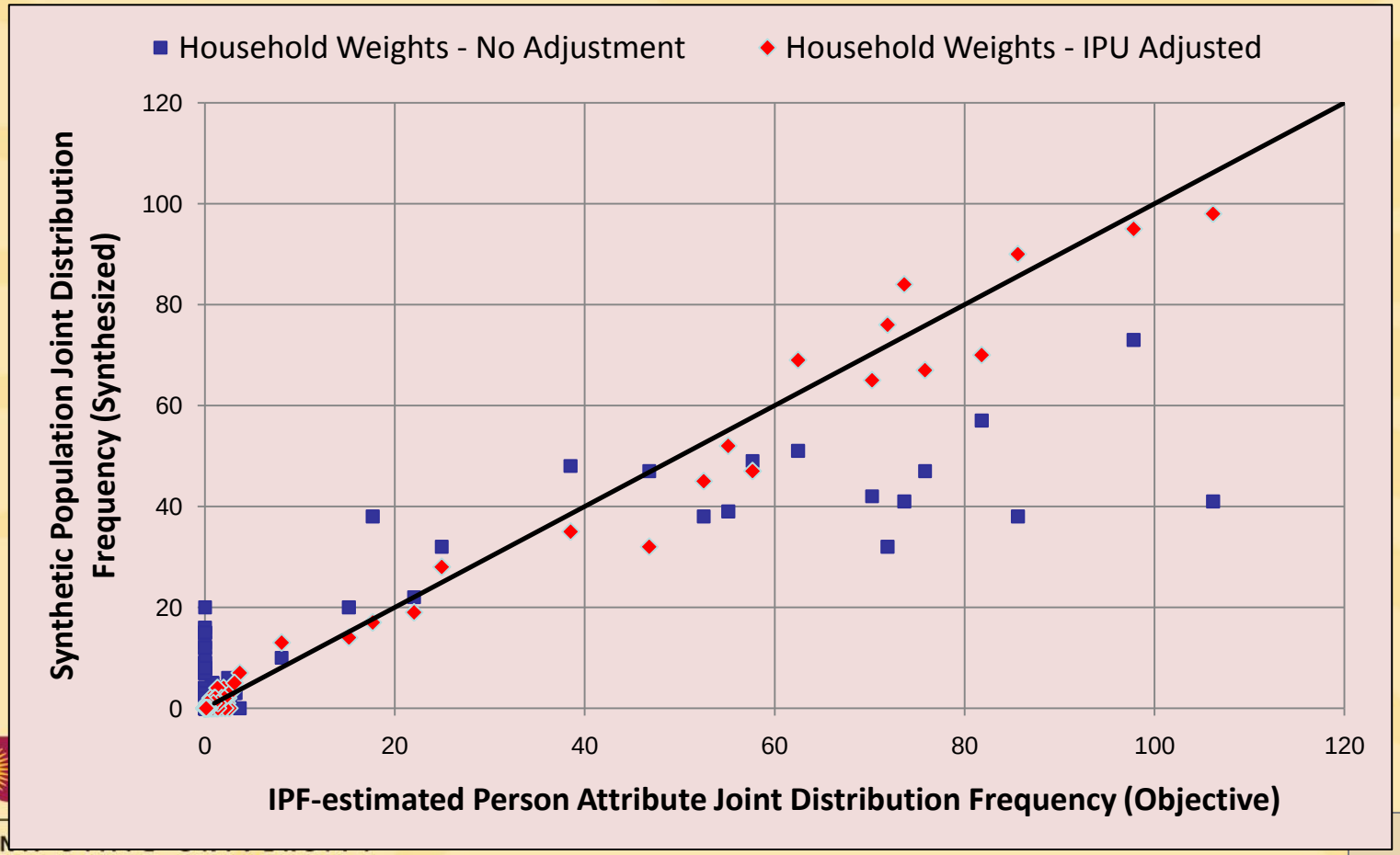
$\chi^2$ = 93.8, df = 120, p-value = 0.9632



# Test Application: Performance (continued)

Blockgroup B

$\chi^2$ = 61.9, df = 108, p-value = 0.9999



# Summary and Conclusions

- State of practice moving towards disaggregate microsimulation modeling of travel demand
- Need synthetic population to implement microsimulation models
- Standard IPF-based procedures for synthetic population generation generally do not control for both household- and person-control variables
- PopGen incorporates new IPU algorithm based on concept of redistributing household weights to reflect differences in household composition
- Test application shows procedure is practical, computationally feasible, and provides a synthetic population that is more representative of the true population

# Development of PopGen Software

- Features of the package
  - A stand-alone application
  - Graphic User Interface to enhance user-friendliness
  - Data downloading, processing and editing capabilities
  - Modify marginal distributions to match person totals more closely
  - Synthesis using classic and IPU approach
  - Interface for visualizing and exporting the results

# Development of PopGen Software (continued)

- PopGen 1.0 was released on July 15, 2009
- PopGen 1.1 was released on November 15, 2009

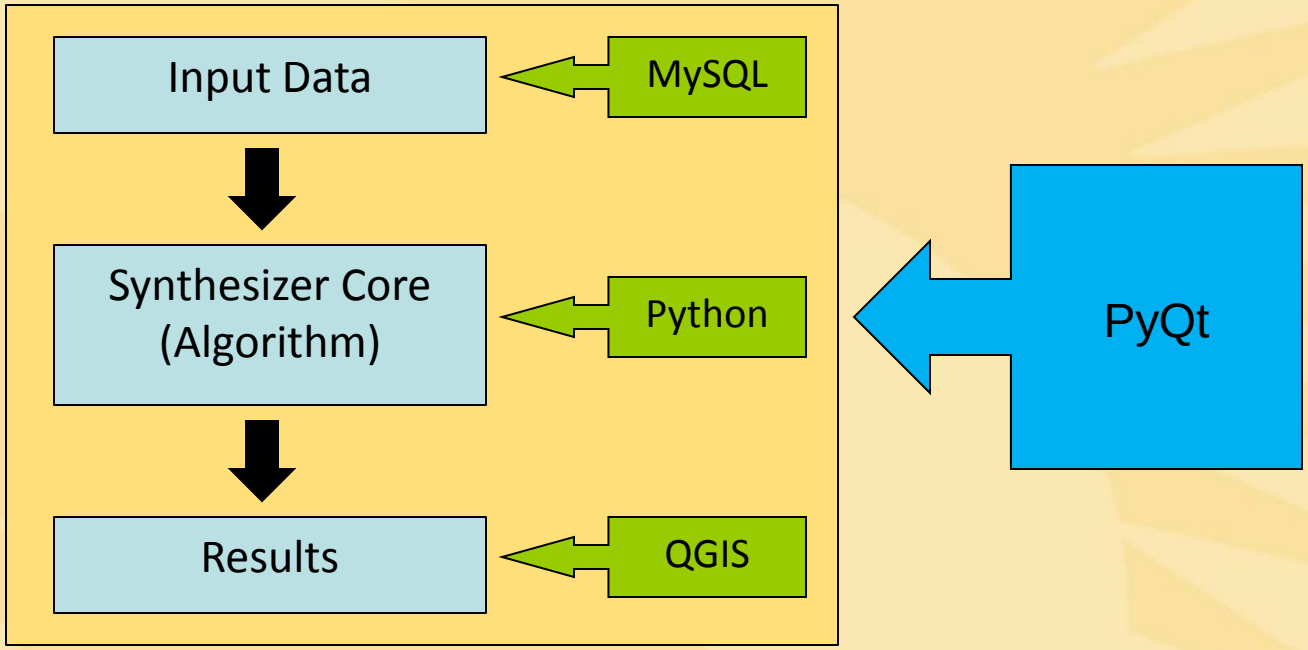
Main Website:

http://urbanmodel.asu.edu/popgen.html

Wiki site:

http://simtravel.wikispaces.asu.edu/Population+Synthesis

# PopGen: Open Source Framework



# Population Evolution Model

- Evolve the base year synthetic population to obtain population for a future year
- A composite set of models to capture all the evolutionary processes
  - Migration of households in and out of a region
  - Person-level life cycle events
  - Household-level changes over time

# Population Evolution Model

## Household Migration Models

- Emigration Model: Rate-based probability model which deletes households in excess in a future year
- Immigration Model: Rate-based probability model which adds households of types that are deficient in the future year

# Population Evolution Model

## Person Evolution Models

- Aging Model: Deterministic model that increments the age of a person by one year
- Fertility Model: Rate-based fertility model for women predicting the birth of a child based on age, race etc.
- Mortality Model: Mortality rate-based model predicting the death of a person based on age, gender etc.
- Educational Attainment Model: Logit-based model predicting if a person will be in school in year t controlling for age, schooling completed, status in year t-1 etc.

# Population Evolution Model

## Person Evolution Models (continued)

- Occupation Choice Model: Multinomial-logit model to predict occupation choice based on age, experience, wage etc. Includes a non-work choice.
- Wage Model: This model would set the wages to clear the job market in each occupation choice. It would be used by the occupation choice model.
- Mobility Options Model: Rate-based model for predicting driver-license and or transit pass holder status.

# Population Evolution Model

## Household Evolution Models

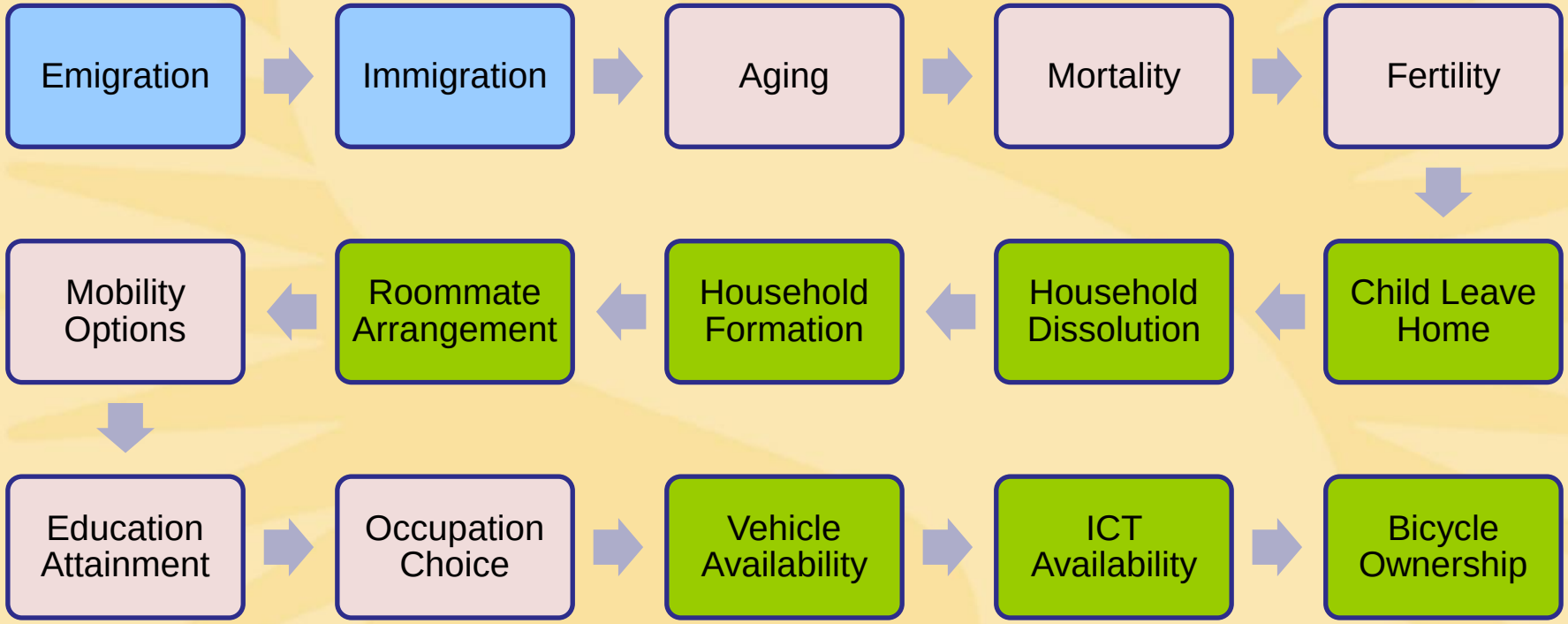
- Child Leaving Home Model: Model based on rates specific to age and gender. A new household is created and assigned to the person.
- HH Separation Model: Predicts the probability that a household with two or more adults chooses to separate. Children, if present, are allocated to a primary residence.
- Household Formation Model: This model predicts the event of persons of different genders combining to form a household. Need to develop "scoring" system that will "match" individuals.

# Population Evolution Model

## Household Evolution Models (continued)

- Roommate Model: Probability that two persons of the same gender will cohabitate is predicted. Need a "scoring" system that will "match" individuals together.
- Auto Availability Model: This ordered-probit/ multinomial logit model would predict number of vehicles by type for a household.
- Bicycle Ownership Model: Number of bicycles present in the household is predicted.
- ICT Availability Model: Predicts cell phones, computers, and internet connectivity for a household. Use market penetration statistics to determine ICT availability.

# Population Evolution Model



# Population Evolution Model

- Challenges
  - Data availability
  - Model estimation
  - Reconciling household interactions and dependencies
  - Modeling simultaneous choices, e.g., Education and Occupation choices
  - Endogeneity of choices, e.g., auto ownership and residential/ workplace location choices (typically in land use model)
  - Overall sequencing of events

# Questions

# ?

# Session 2 – Installation

PopGen Training Workshop

May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Session 2 – Installation

- In this session, you will learn how to install the PopGen software and the supplementary software required to run PopGen
- The supplementary software needs to be installed prior to launching PopGen on your system

# Installing Supplementary Software

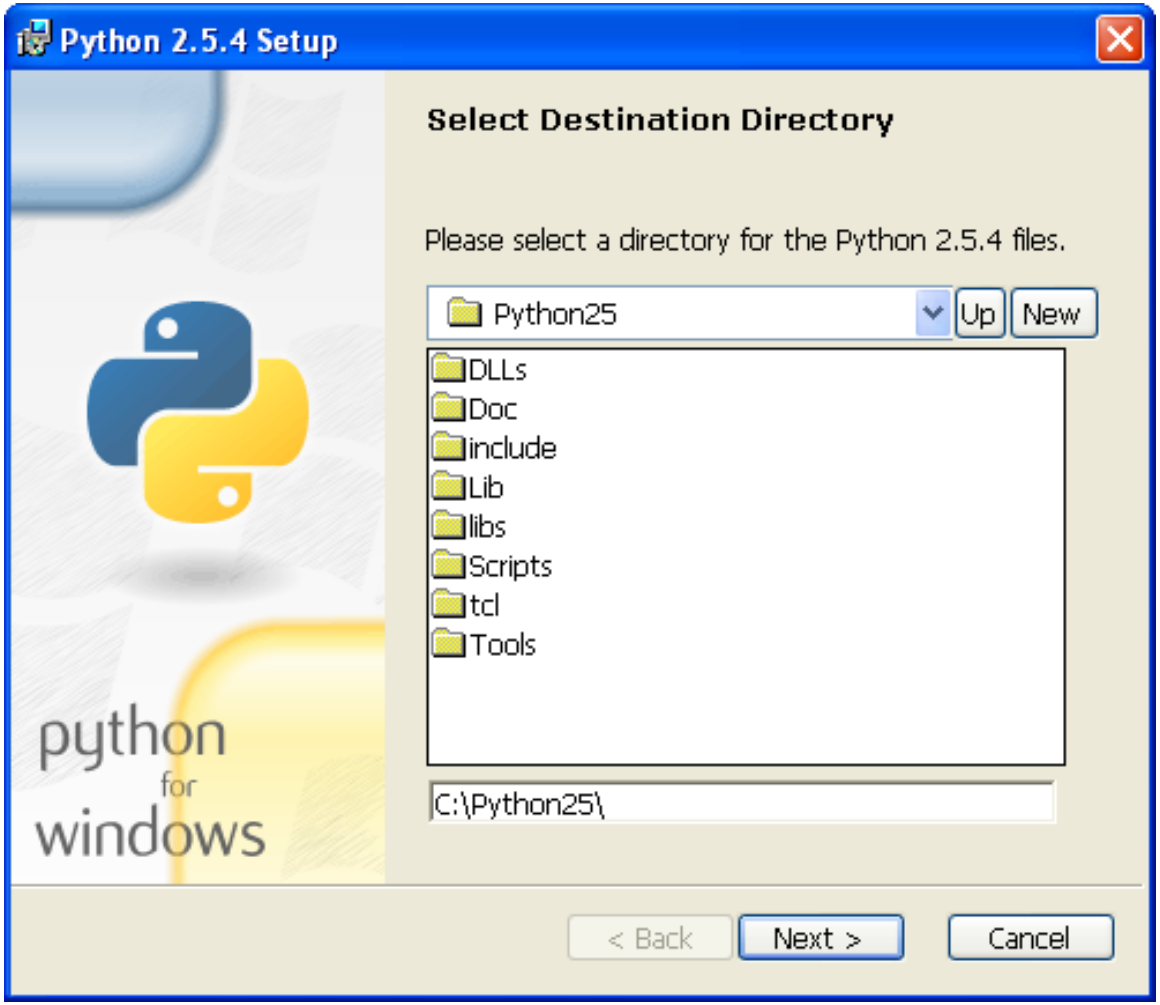
- Before installing PopGen 1.1 on your computer, additional software packages need to be installed
- Each software package should be installed to the directory “C:\<filename>”

  (**Example: C:\MySQL)**
- Unless otherwise stated, choose the **typical** or **default** installation setting during setup

# Python 2.5.4 for Windows

- Go to URL:
- http://www.python.org/ftp/python/2.5.4/python-2.5.4.msi
- Choose **Run** to launch the installation wizard
- Choose the file directory location **C:\Python25**
- Click **Next** and leave all other options on their default settings
- Install Python 2.5.4 on your system

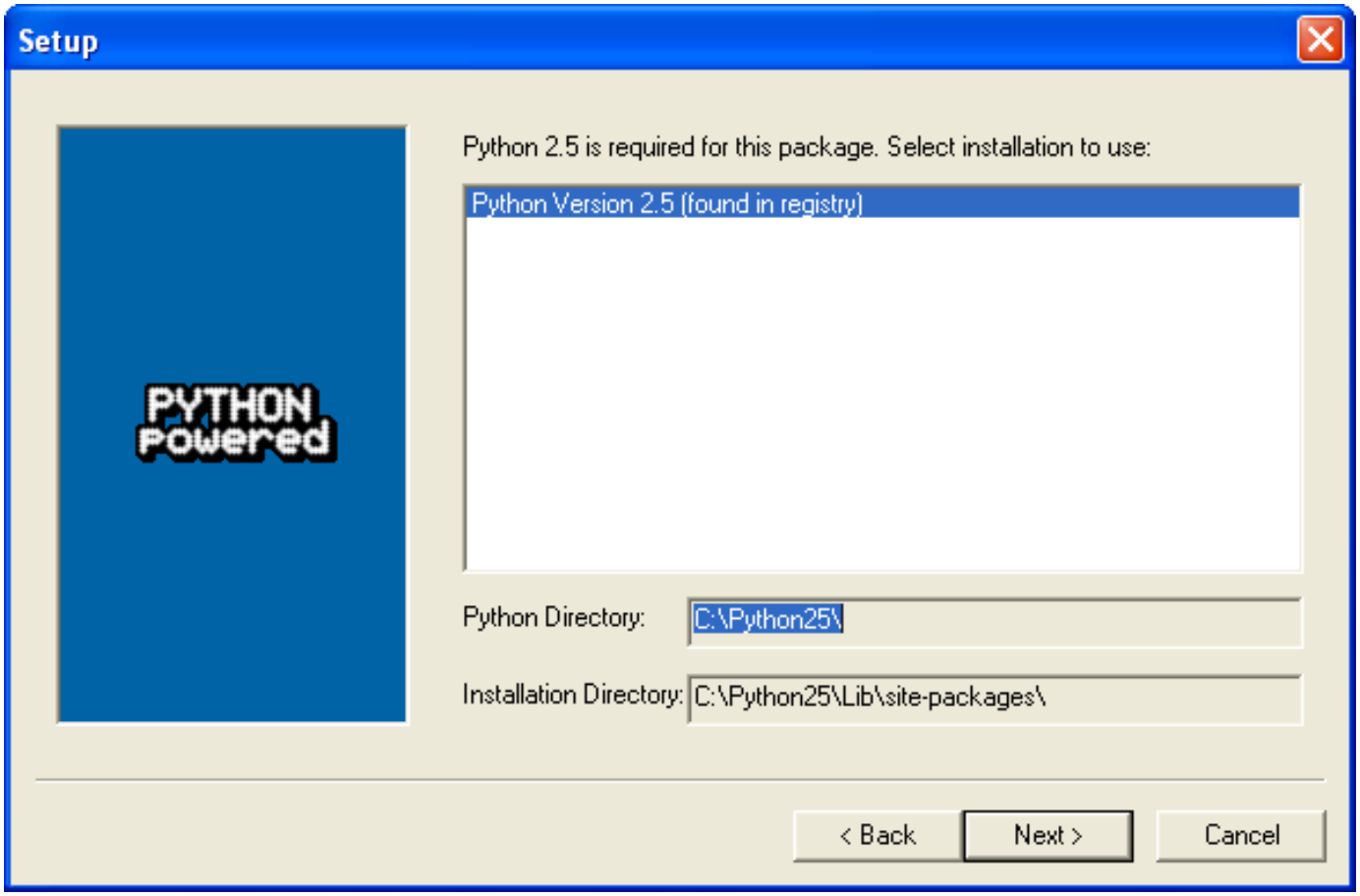
# Python 2.5.4 for Windows



# NumPy 1.4.1

- Go to URL: http://sourceforge.net/projects/numpy/files/NumPy/1.4.1/numpy-1.4.1-win32-superpack-python2.5.exe/download
- If the file does not download immediately, select "Please use this direct link" at the top
- Choose the file directory location **C:\Python25**
- Click **Next** and leave all other options on their default settings
- Install NumPy 1.4.1 on your system

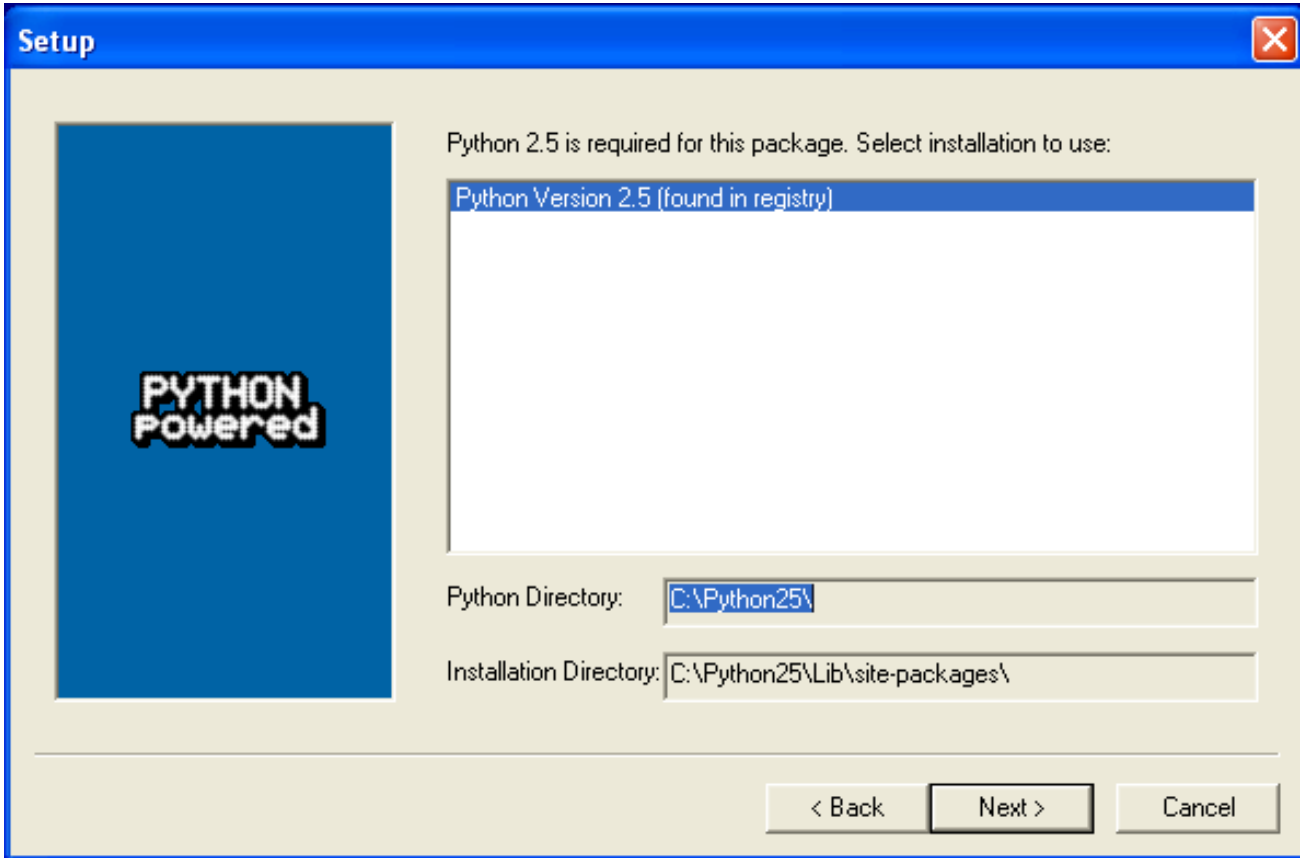
# NumPy 1.4.1


Setup
PYTHON
Powered
Python 2.5 is required for this package. Select installation to use:
Python Version 2.5 (found in registry)
Python Directory:
C:\Python25\
Installation Directory:
C:\Python25\Lib\site-packages\
< Back
Next >
Cancel

# SciPy 0.7.2

- Go to URL: http://sourceforge.net/projects/scipy/files/scipy/0.7.2/scipy-0.7.2-win32-superpack-python2.5.exe/download
- If the file does not download immediately, select "Please use this direct link" at the top
- Choose the file directory location **C:\Python25**
- Click **Next** and leave all other options on their default settings
- Install SciPy 0.7.2 on your system

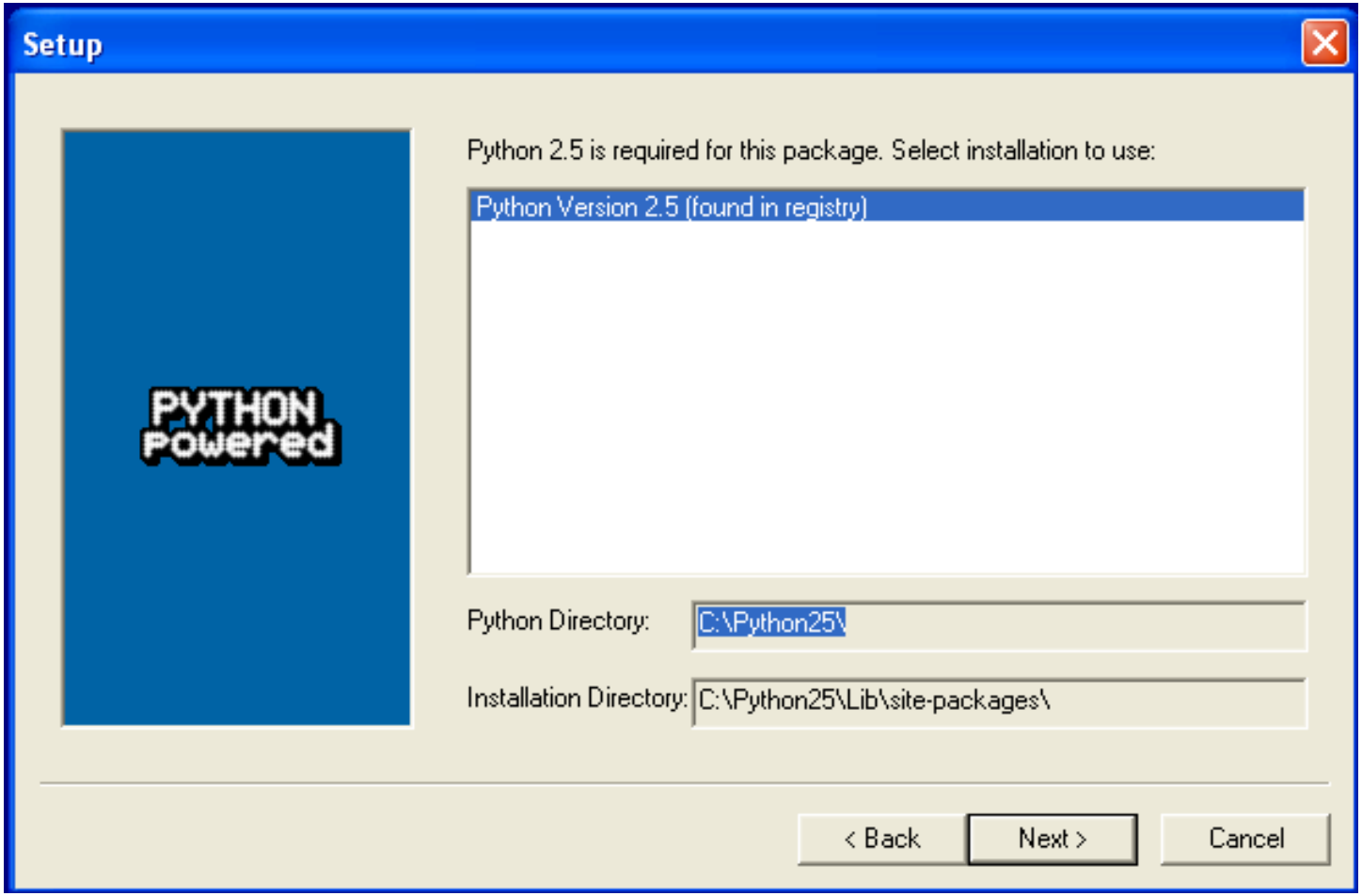
# SciPy 0.7.2


Setup
PYTHON
powered
Python 2.5 is required for this package. Select installation to use:
Python Version 2.5 (found in registry)
Python Directory:
C:\Python25\
Installation Directory:
C:\Python25\Lib\site-packages\
< Back
Next >
Cancel

# MySQL Python 1.2.2

- Go to URL: http://sourceforge.net/project/downloading.php?group_id=22307&filename=MySQL-python-1.2.2.win32-py2.5.exe&a=85924604
- If the file does not download immediately, select "Please use this direct link" at the top
- Choose the file directory location **C:\Python25**
- Click **Next** and leave all other options on their default settings
- Install MySQL Python 1.2.2 on your system

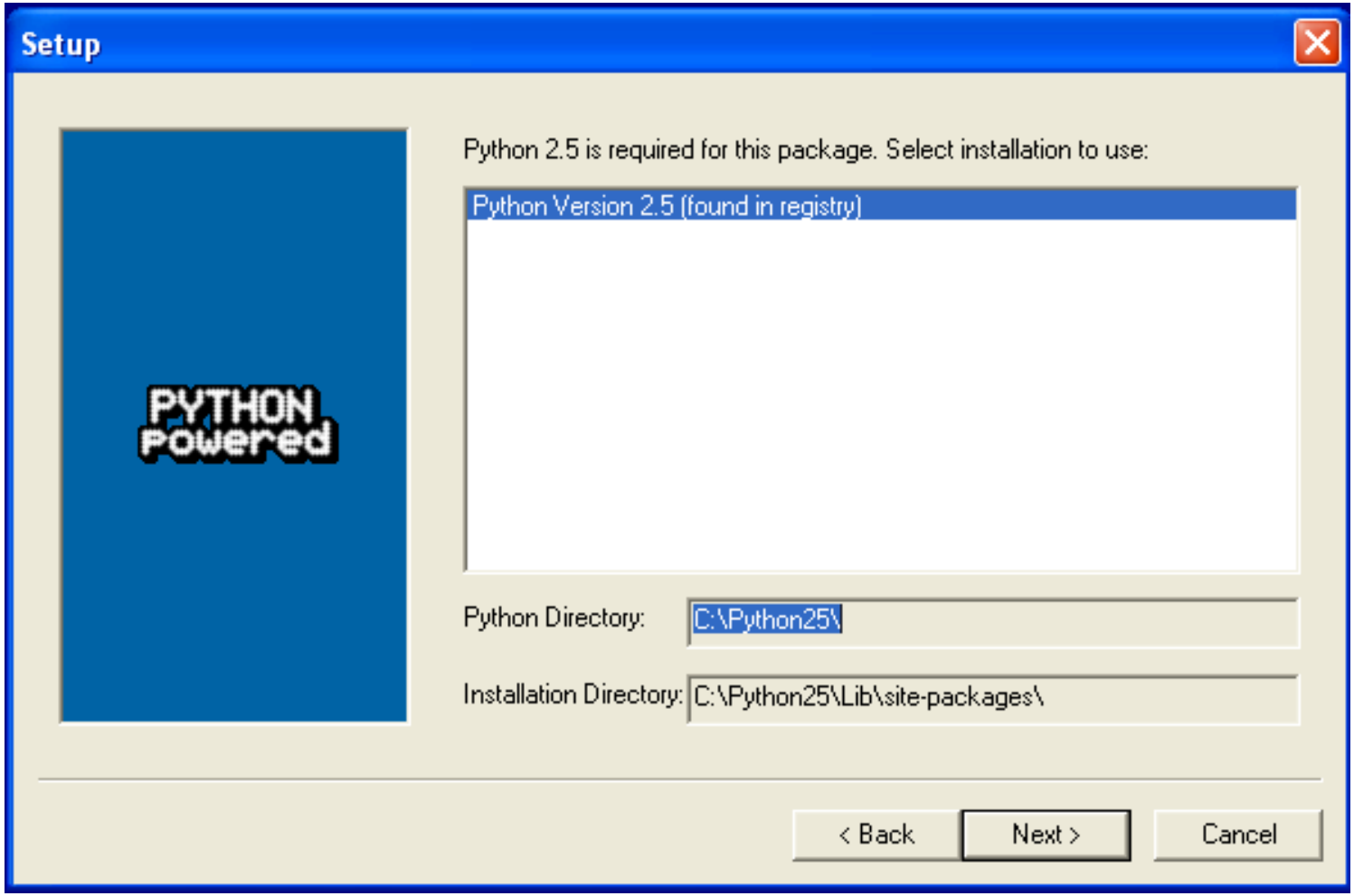
# MySQL Python 1.2.2



# Matplotlib 0.99.1.1

- Go to URL: http://sourceforge.net/projects/matplotlib/files/matplotlib/matplotlib-0.99.1/matplotlib-0.99.1.win32-py2.5.exe/download
- If the file does not download immediately, select "Please use this direct link" at the top
- Choose the file directory location **C:\Python25**
- Click **Next** and leave all other options on their default settings
- Install Matplotlib 0.99.1.1 on your system

# Matplotlib 0.99.1.1

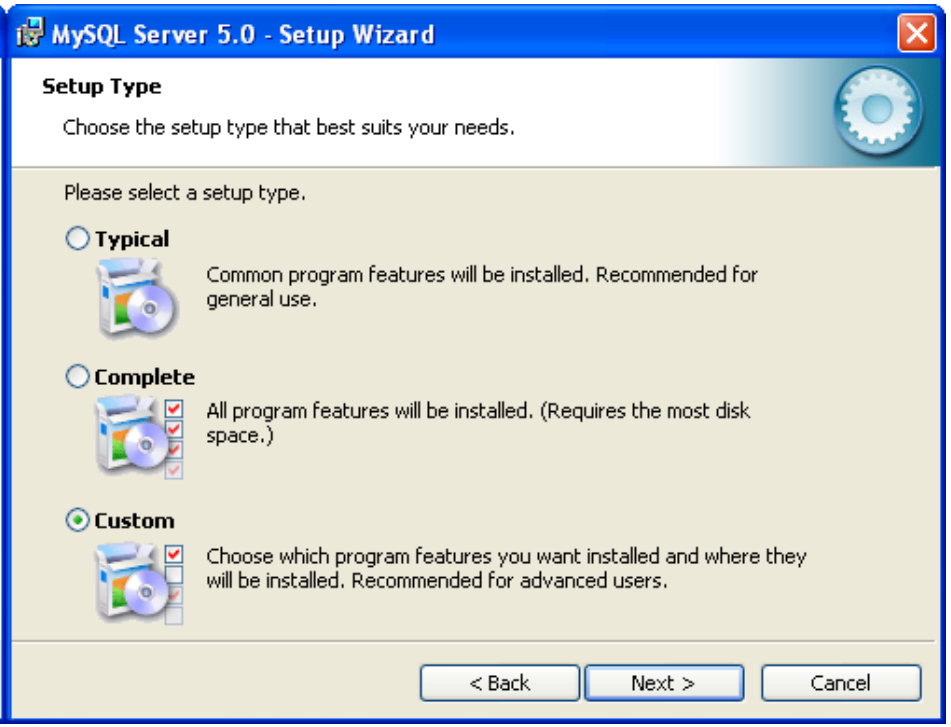


# MySQL 5.1

- Go to URL: http://dev.mysql.com/get/Downloads/MySQL-5.1/mysql-essential-5.1.46-win32.msi/from/http://mysql.mirrors.hoobly.com/
- Choose **Run** to start the installation wizard

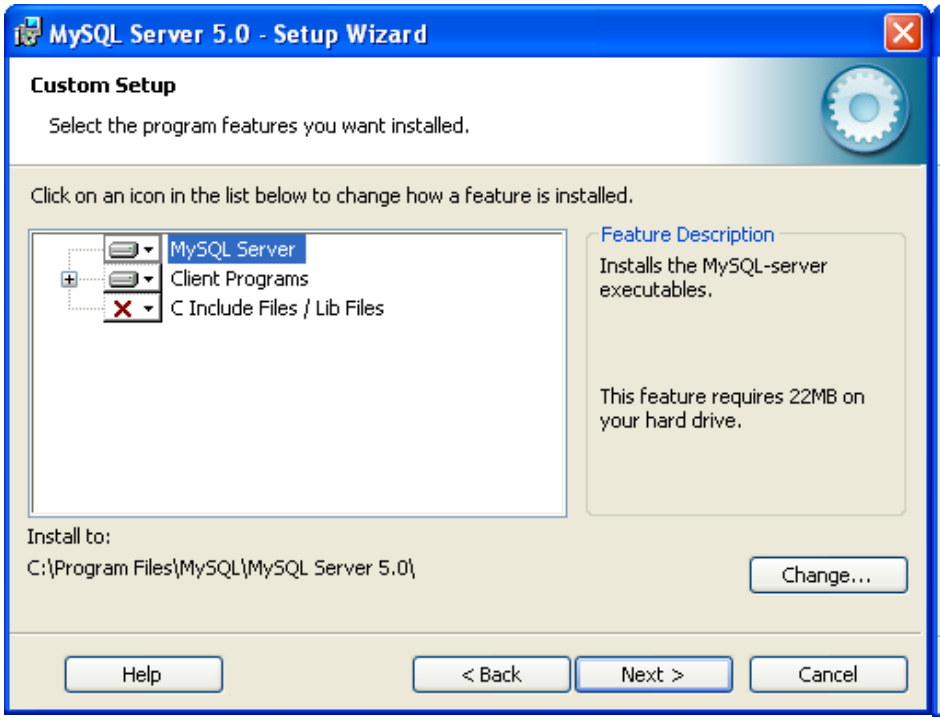
# MySQL 5.1

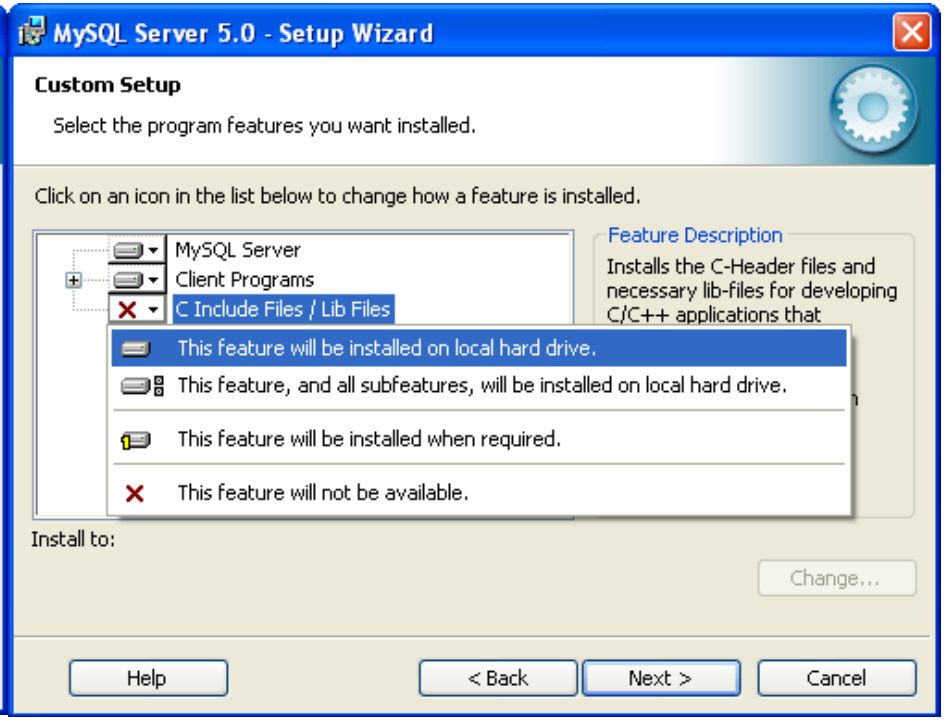
- When prompted to select a setup type, be sure to select **Custom**
- Click **Next** to continue



# MySQL 5.1

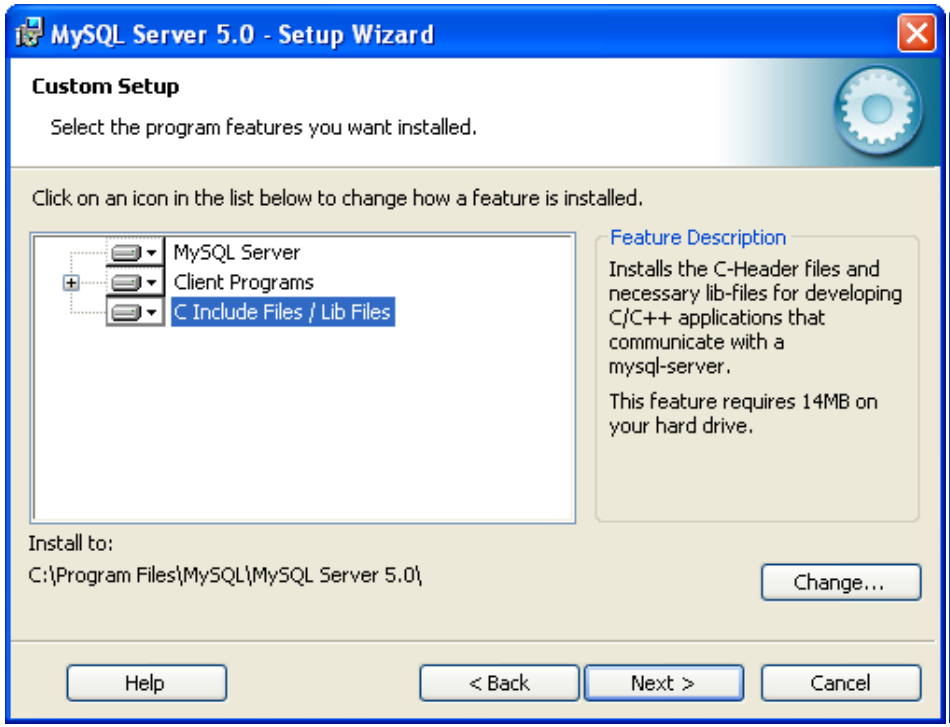
- Make sure that the feature **C Include Files/Lib Files** is installed by changing the option to “this feature will be installed on local hard drive”

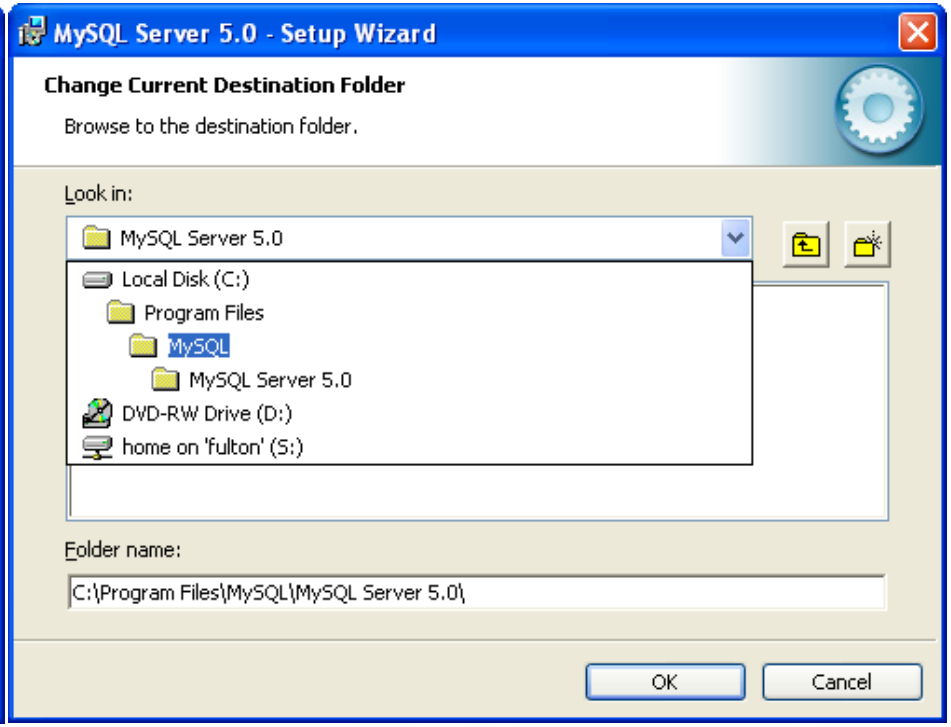


# MySQL 5.1

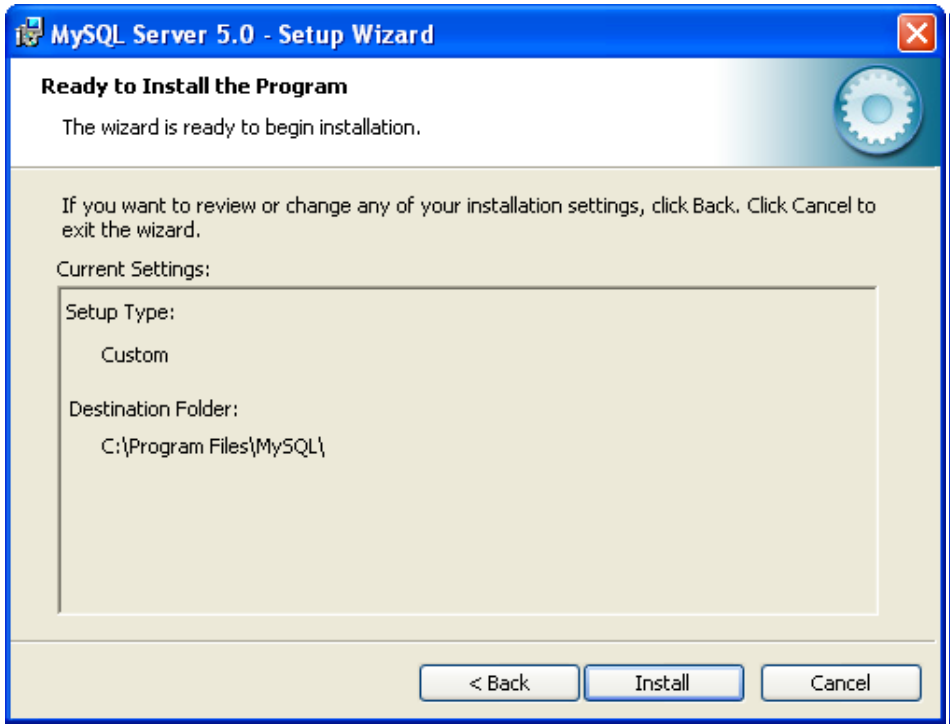
- Change the “Install to” location to **C:\MySQL** by clicking on the **Change** button near the bottom of the Custom Setup screen

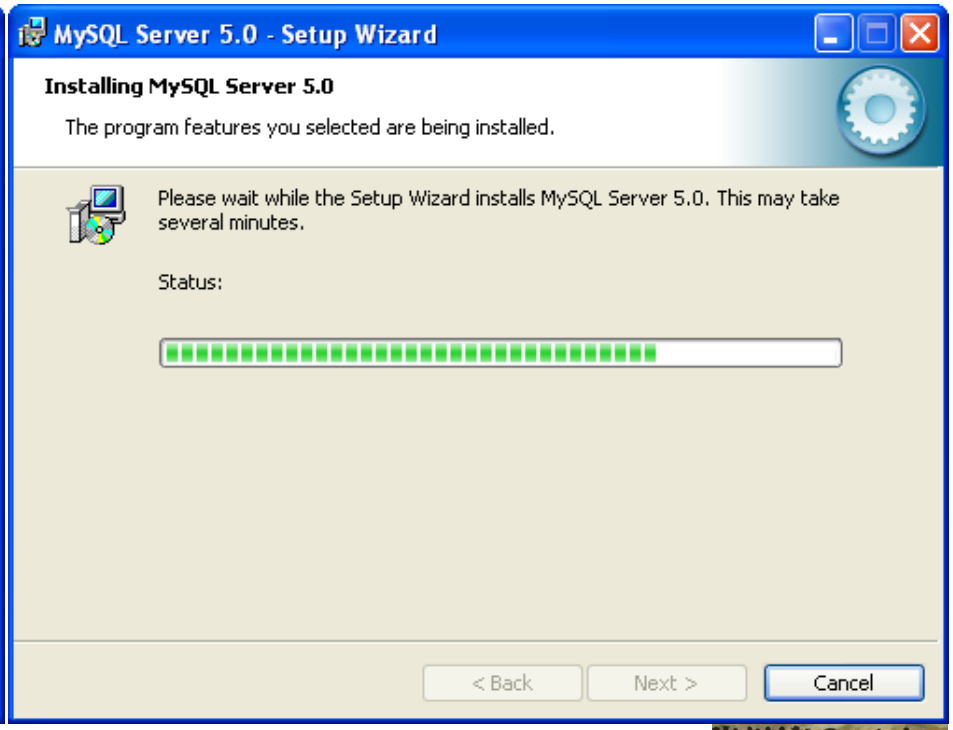


# MySQL 5.1

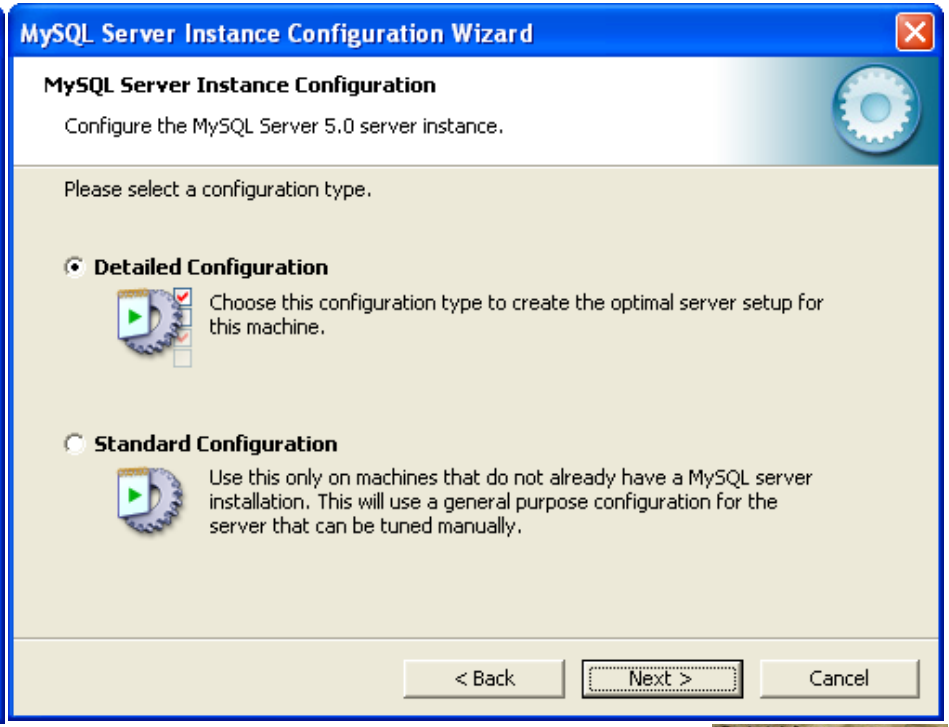
- Choose **Install** to begin the installation
- When installation is almost complete, an advertisement will appear; just click through this



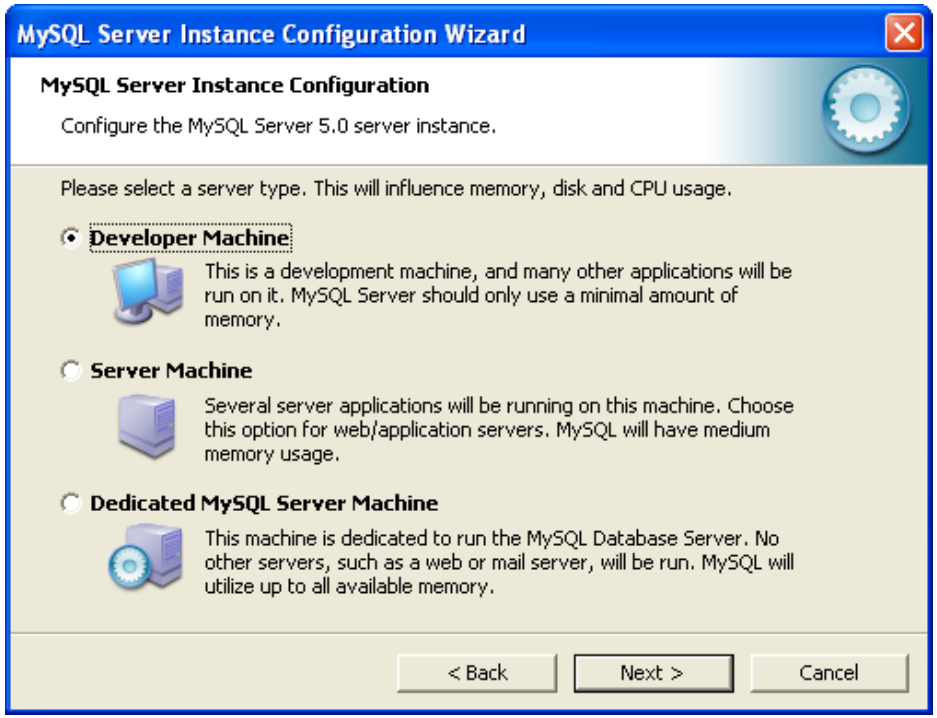
# MySQL 5.1

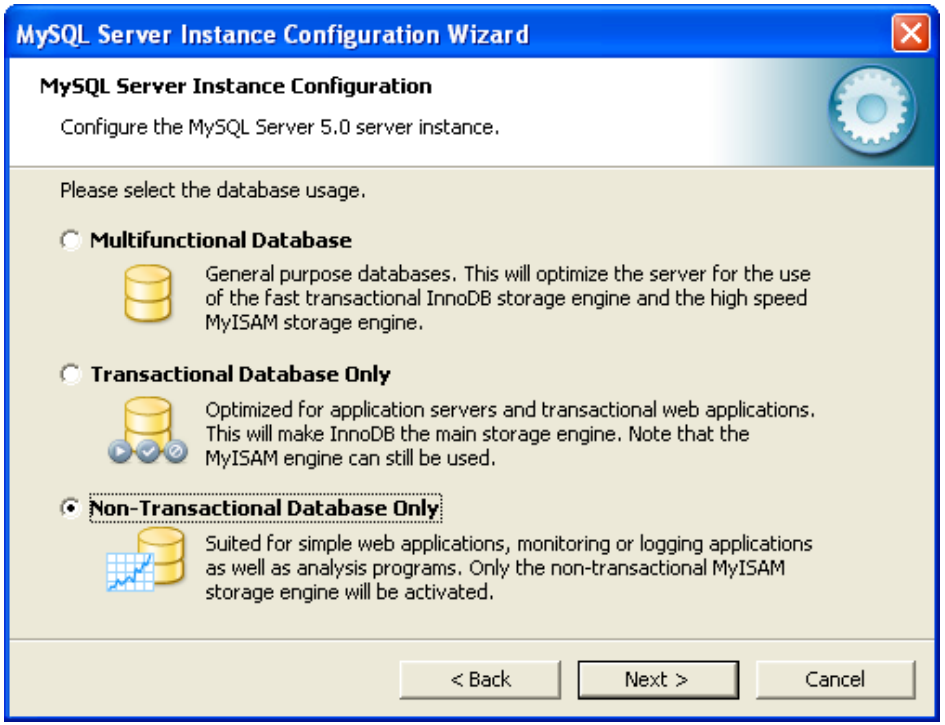
- Once the installation has finished, a Wizard Completed screen will appear prompting you to configure the MySQL Server
- Check the box to **Configure the MySQL Server now**
- On the next screen, choose the **Detailed Configuration** option



# MySQL 5.1

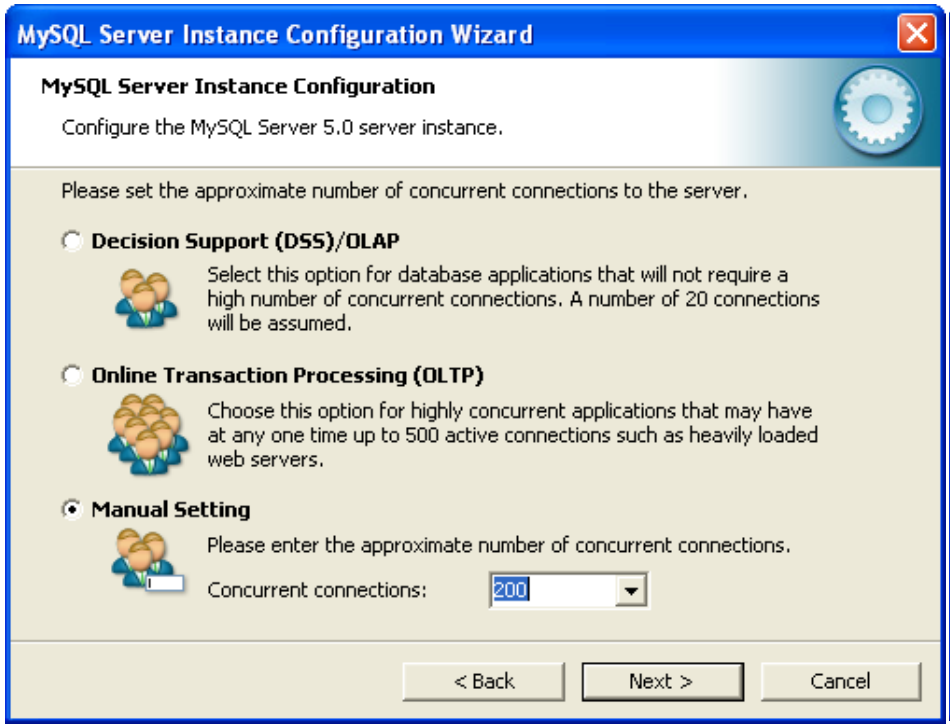
- On the second screen, leave the **Developer Machine** option selected
- On the third screen, choose **Non-Transactional Database Only**
- Click **Next** to continue

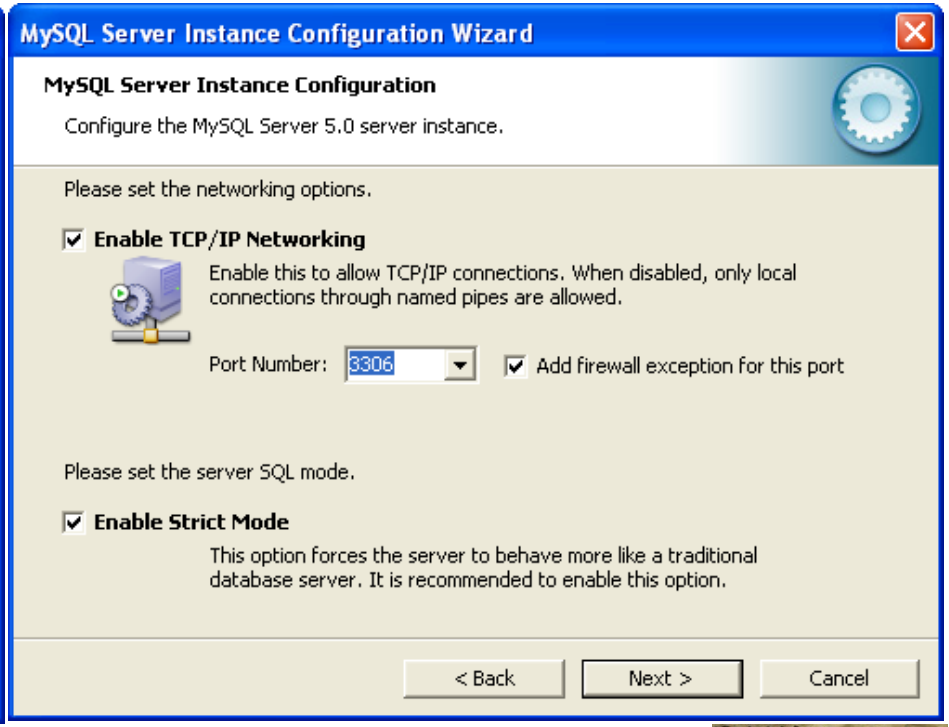


# MySQL 5.1

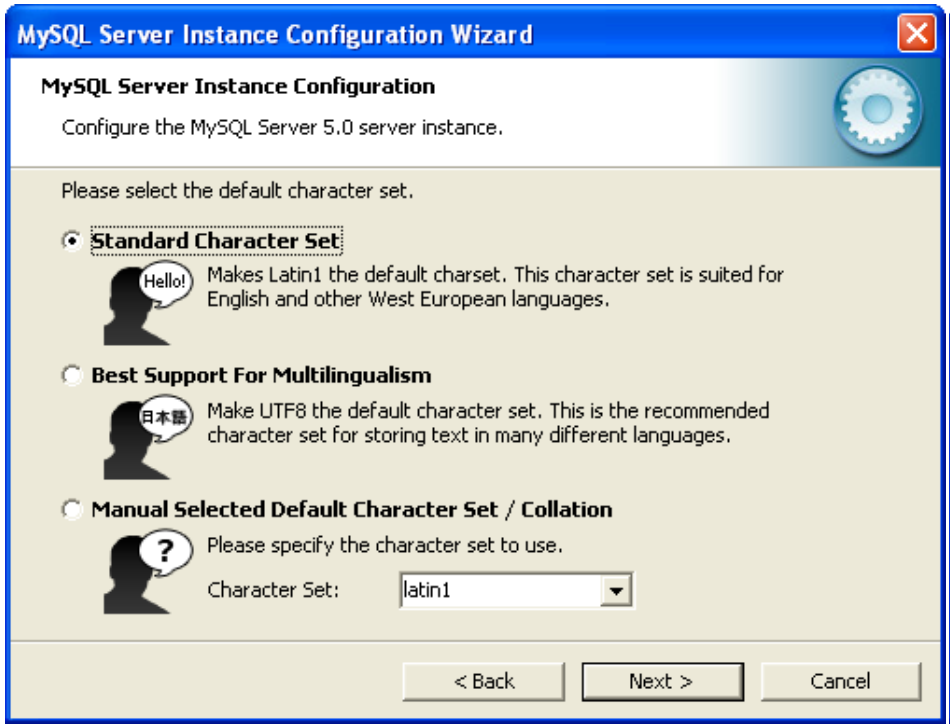
- On the fourth screen, choose **Manual Setting** and change the Concurrent connections to 200
- On the fifth and sixth screens, use the default settings



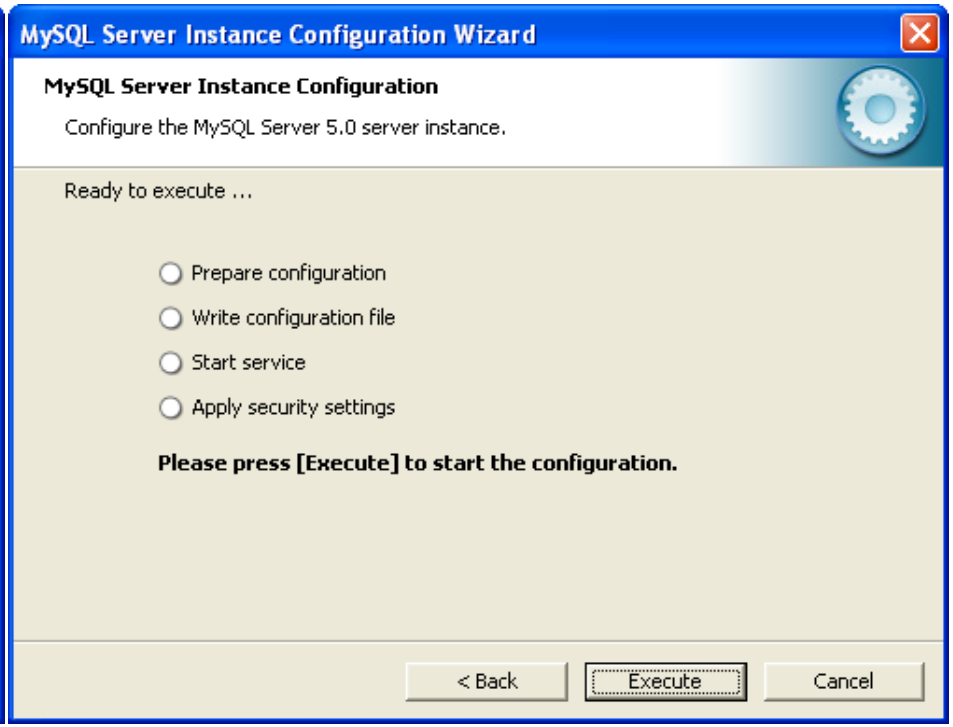
# MySQL 5.1

- On the seventh screen, check the box **Include Bin Directory in Windows PATH**
- Click **Next** to continue



# MySQL 5.1

- On the eighth screen, you will need to setup a MySQL password
- Choose any password you like
- Click **Next** to continue and then click **Execute** to complete the MySQL configuration



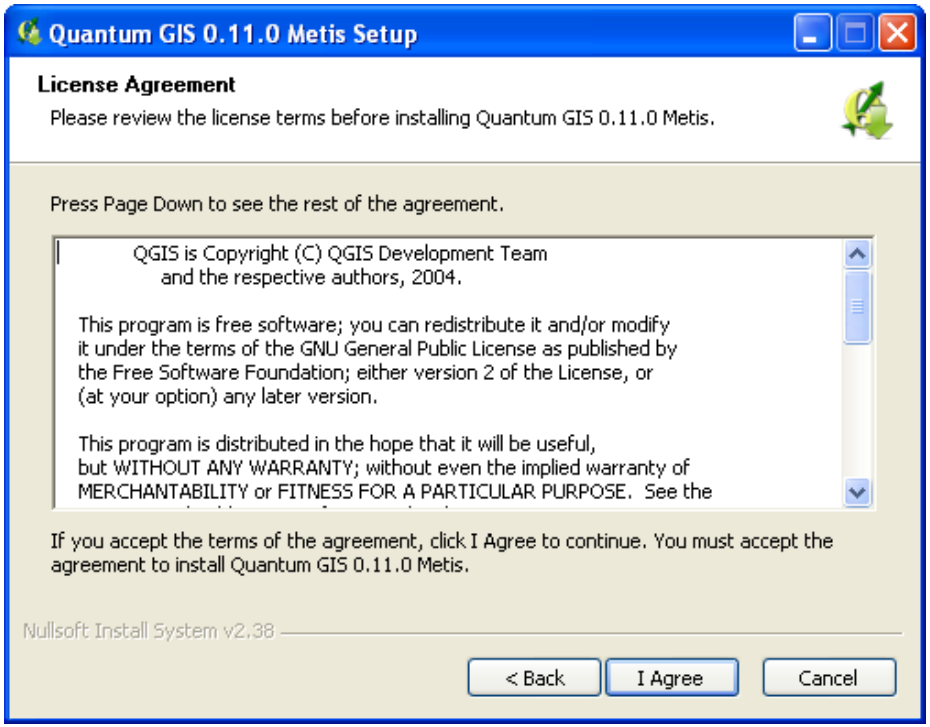
# PyQt 4.7.3-1

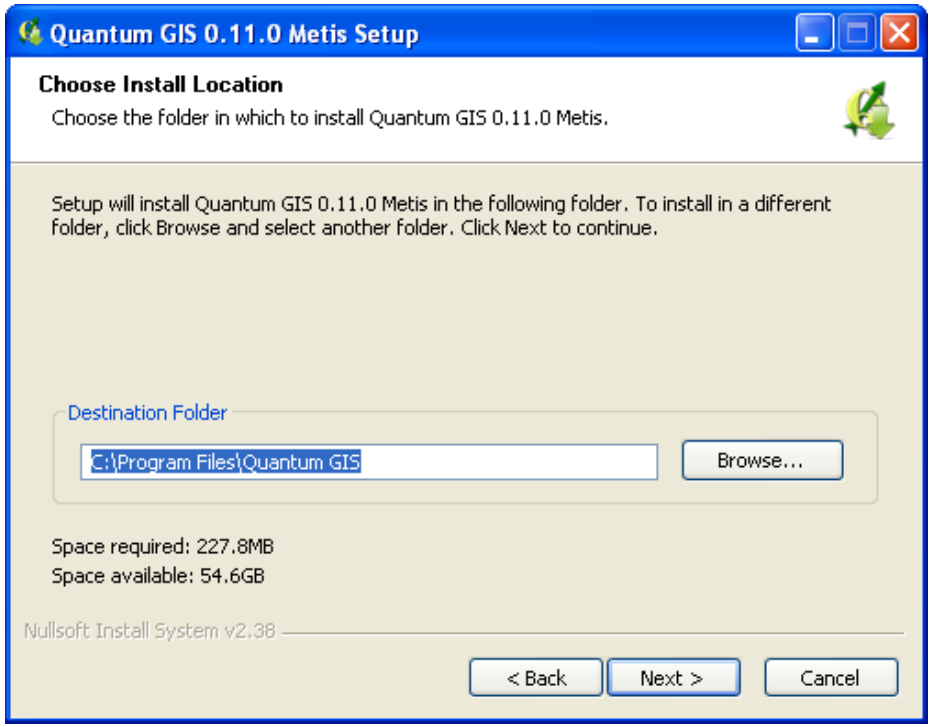
- Go to URL: http://www.riverbankcomputing.co.uk/static/Downloads/PyQt4/PyQt-Py2.5-gpl-4.7.3-1.exe
- Choose Run to start the installation wizard
- Choose the file directory location **C:\Python25**
- Click **Next** and leave all other options on their default settings
- Install PyQt 4.7.3 on your system

# Quantum GIS 0.11.0

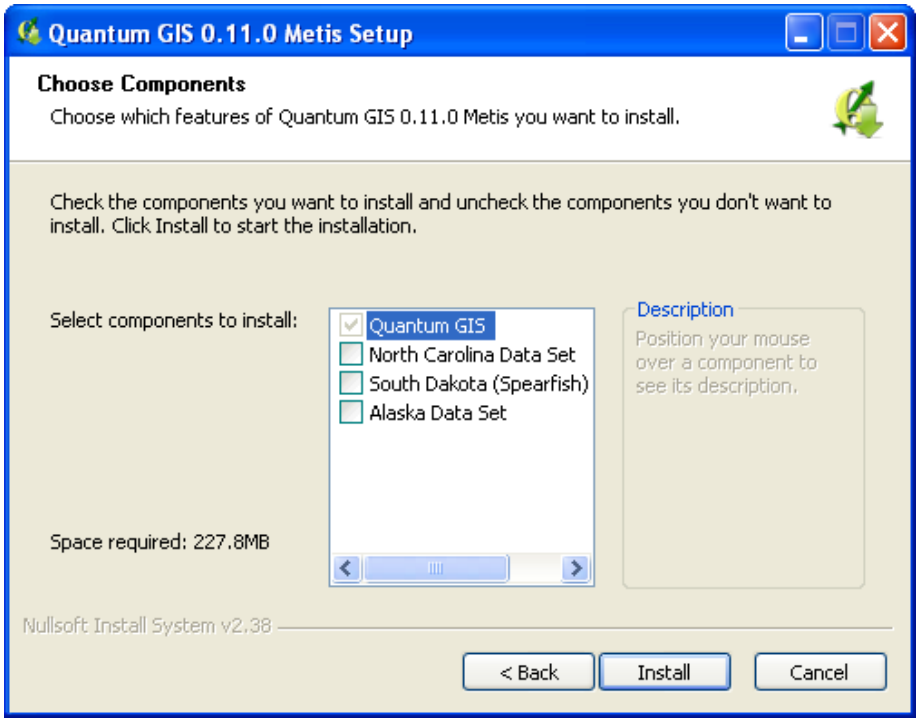
- Go to URL: http://download.osgeo.org/qgis/win32/QGIS-0.11.0-2-Setup.exe
- Choose **Run** to start the installation wizard
- Choose the file directory location **C:\QGIS**
- Click **Next** and leave all other options on their default settings
- Install Quantum GIS on your system

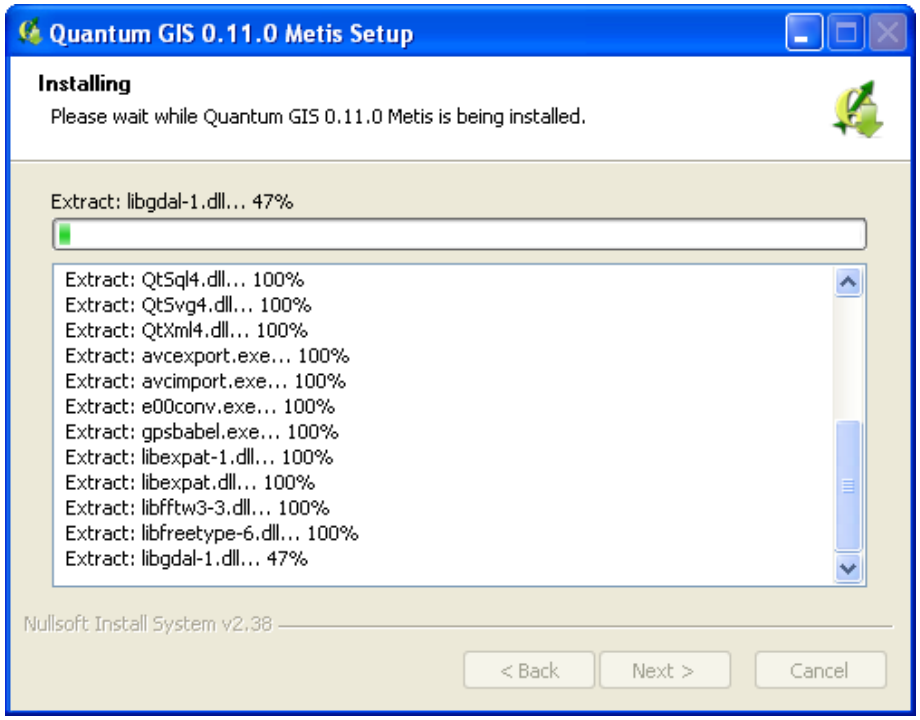
# Quantum GIS 0.11.0



# Quantum GIS 0.11.0

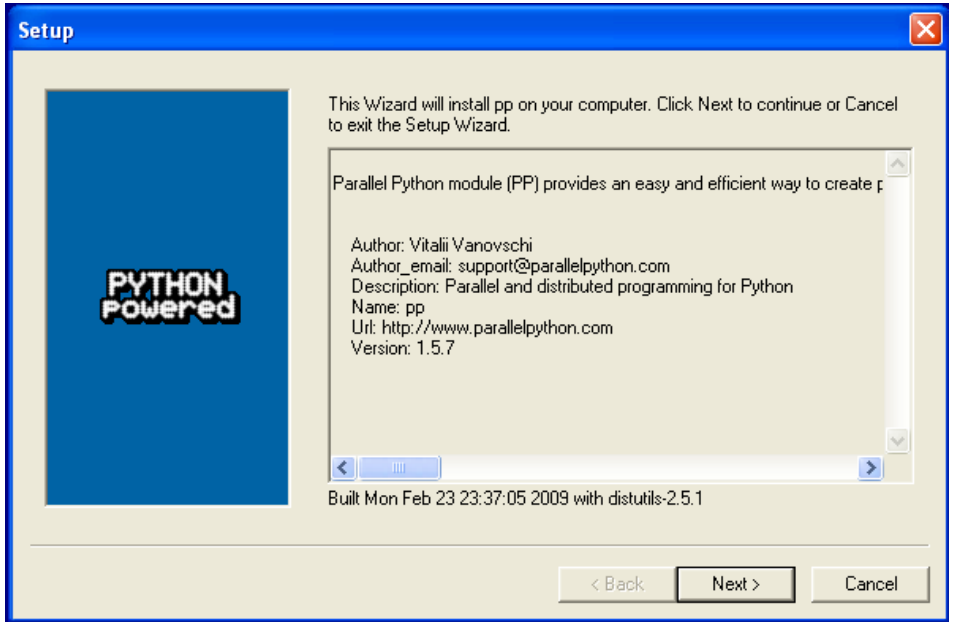


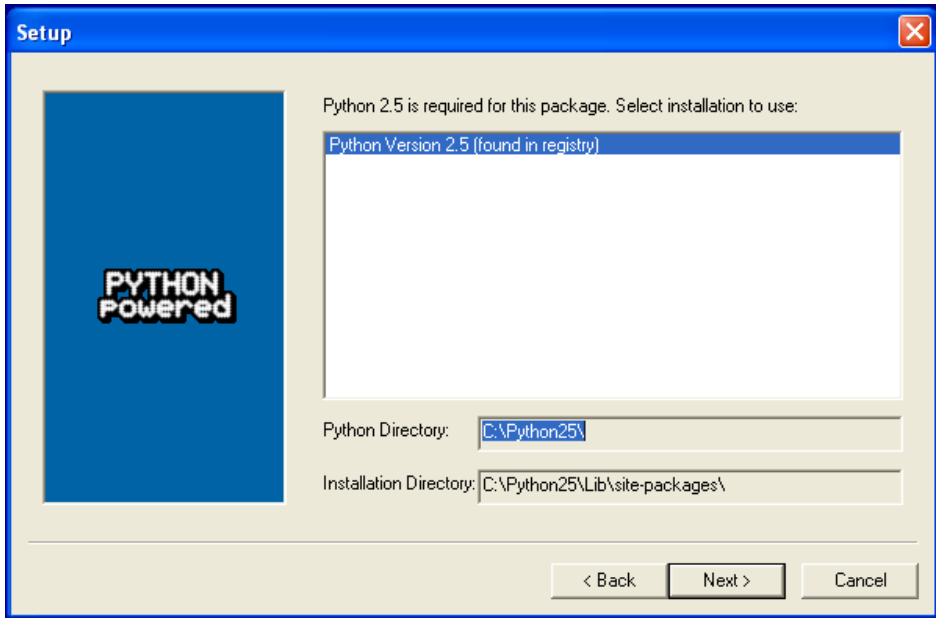



# Parallel Python 1.5.7

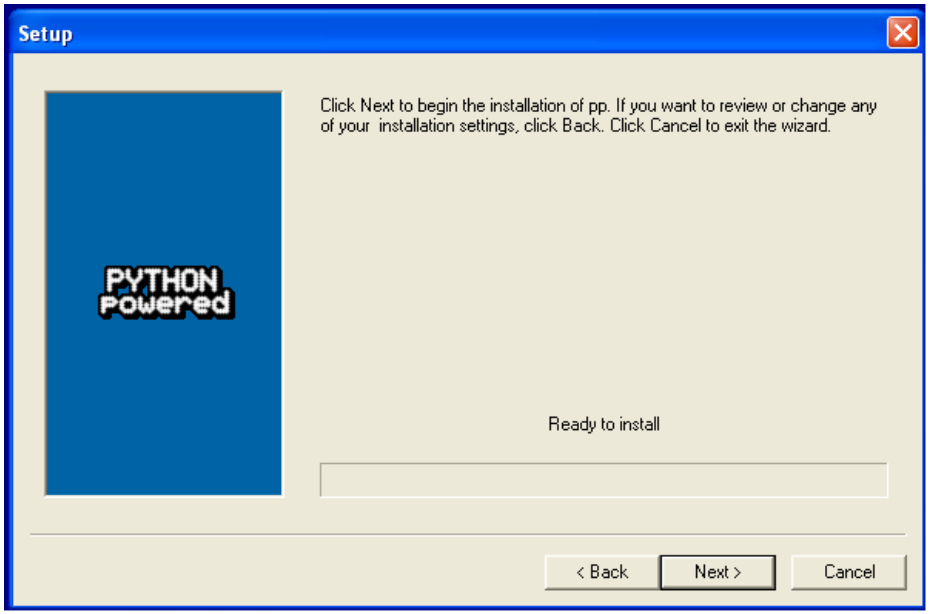
- Go to URL: http://www.parallelpython.com/downloads/pp/pp-1.5.7.exe
- Choose **Run** to start the installation wizard
- Choose the file directory location **C:\Python25**
- Click **Next** and leave all other options on their default settings
- Install Parallel Python 1.5.7 on your system

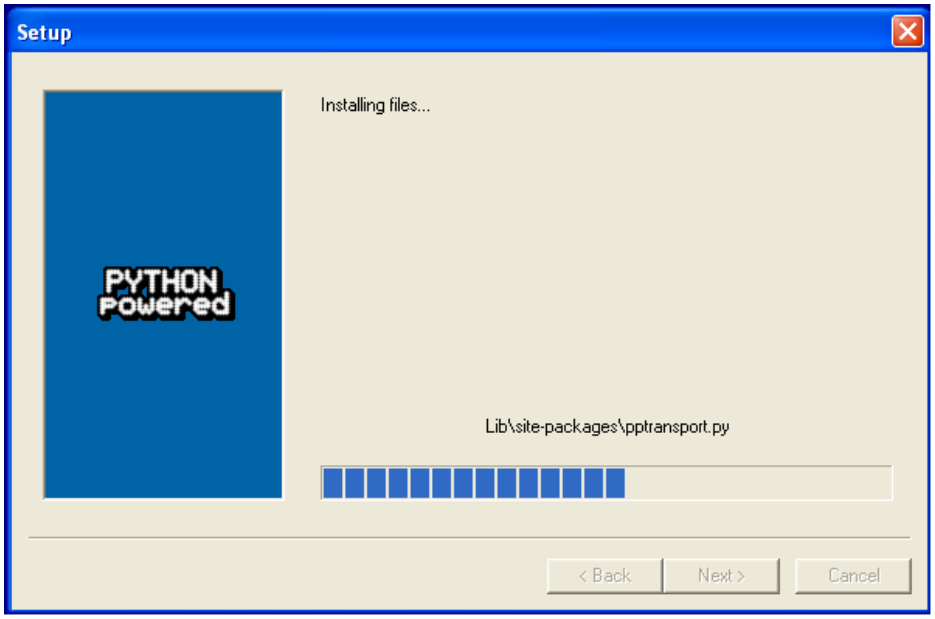
# Parallel Python 1.5.7



# Parallel Python 1.5.7



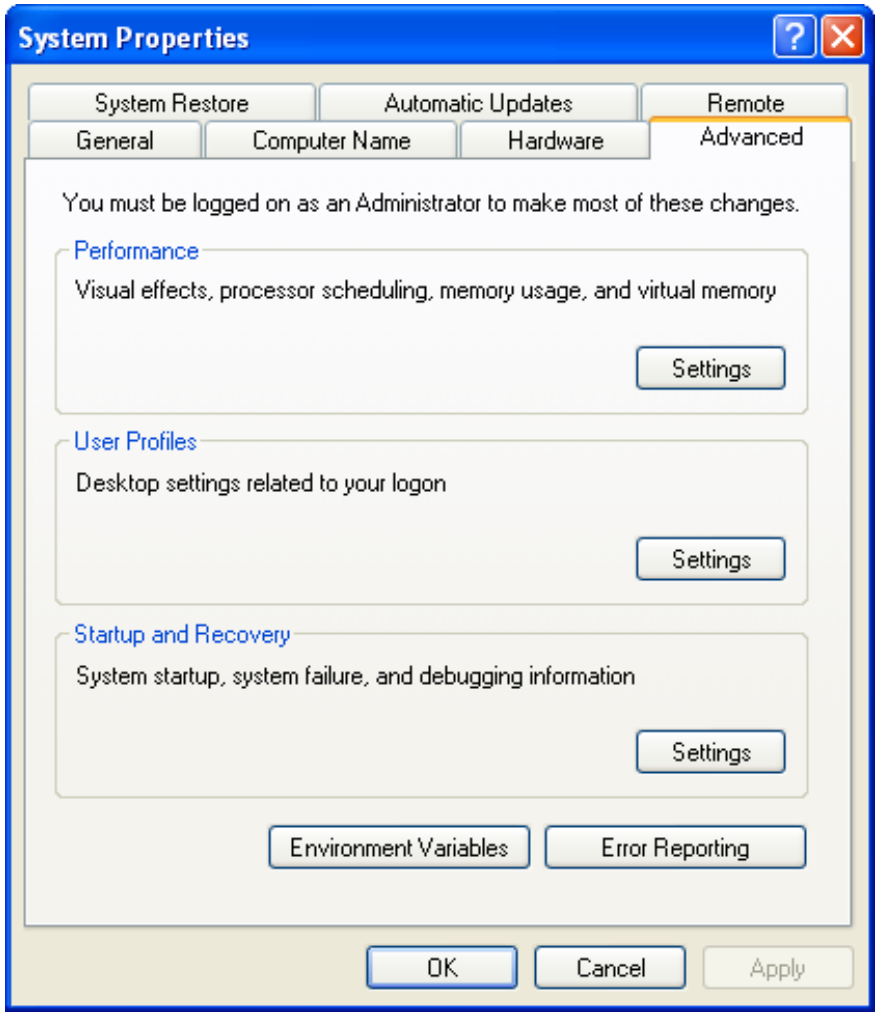
# Download PopGen Software

- Go to URL: http://urbanmodel.asu.edu/popgen/code.html
- Download and extract the PopGen program code titled **Code**
- Store the files on your computer's hard drive (Preferably as C:\PopGen)
- Before running PopGen, the environment variables need to be set on the system

# Environment Variables

- Right-click on **My Computer** and select **Properties**
- Click the **Advanced** tab and then select **Environment Variables** at the bottom
- Under **System Variables** find the variable named **Path**
- With the **Path** variable still selected, click **Edit**

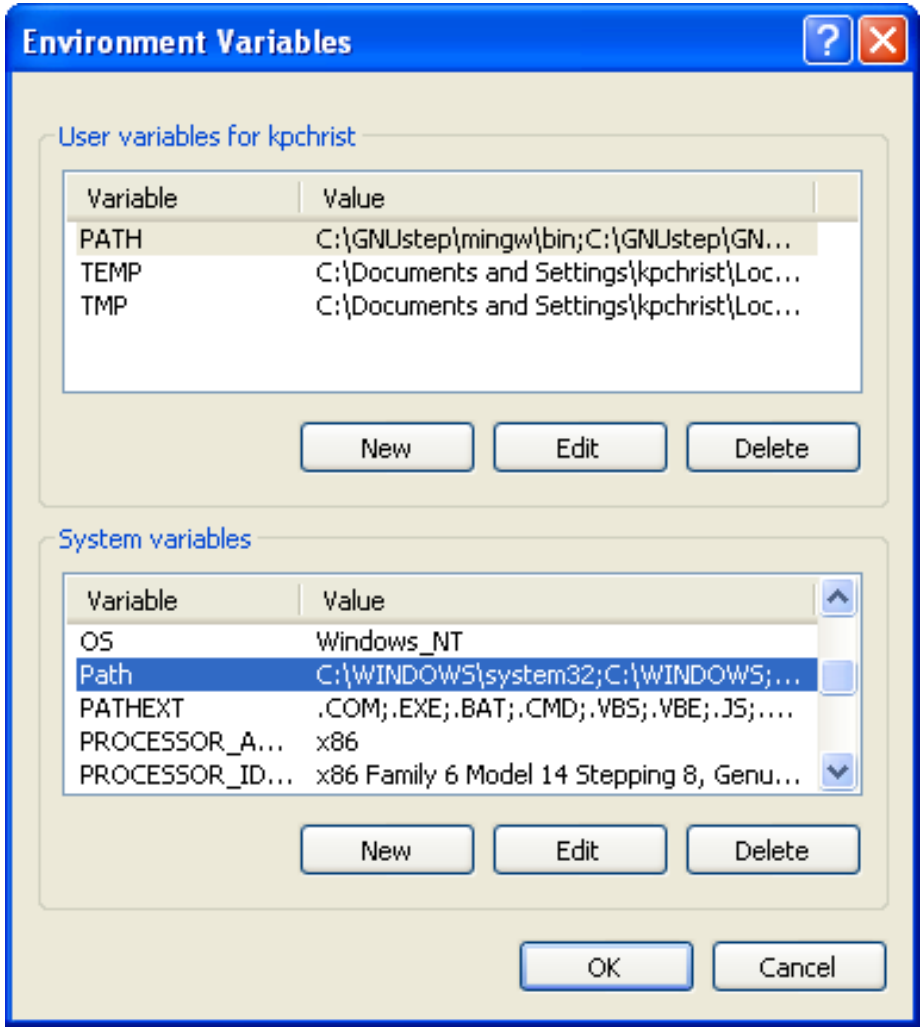
# Environment Variables

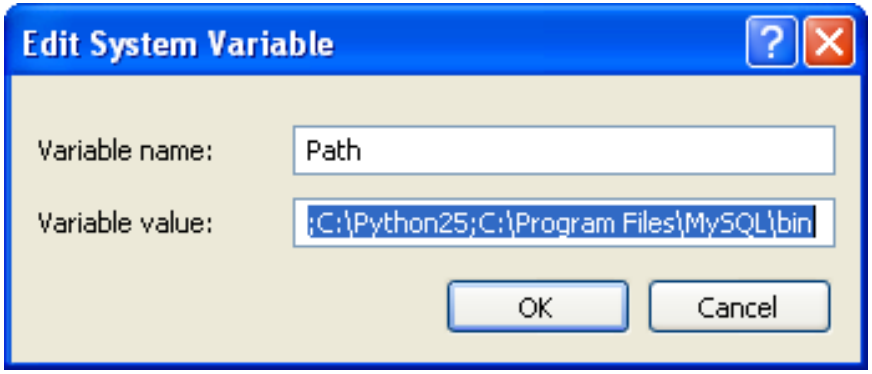


# Environment Variables

- Add the following locations to the **Variable Value**, separating each with a semicolon:
  1. C:\qgis
  2. C:\Python25
- Once these have been entered, click **OK**

# Environment Variables

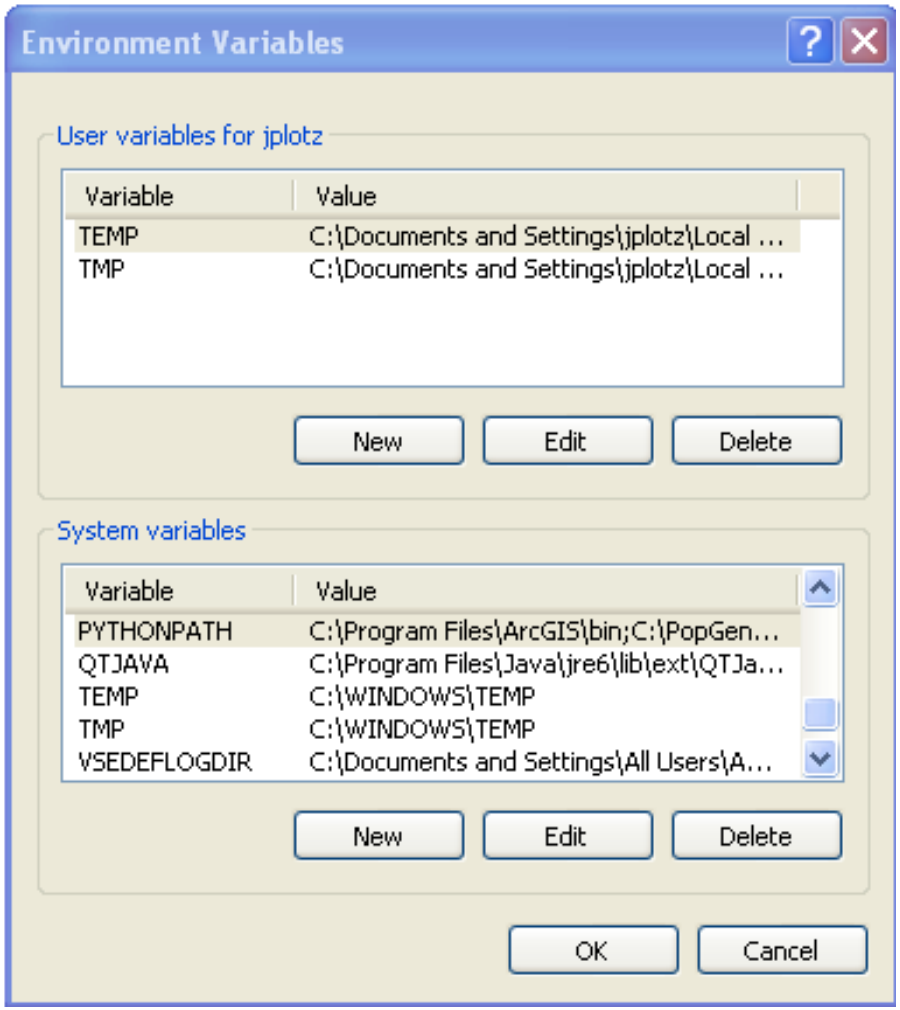


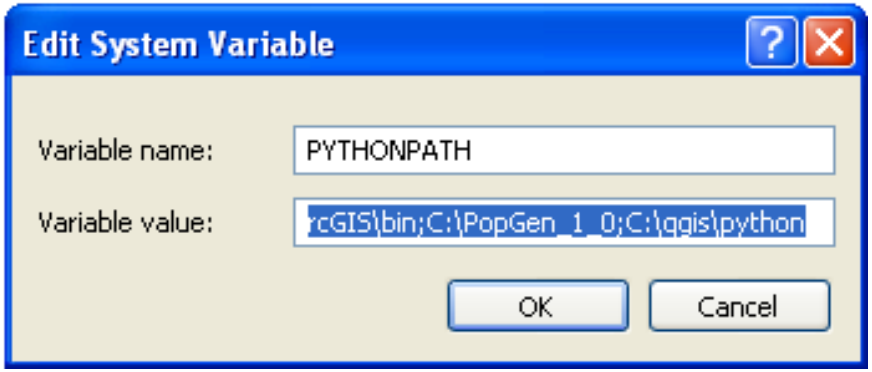



# Environment Variables

- We also need to edit the **PYTHONPATH** variable
- Under **System Variables** find the variable named **PYTHONPATH**
- Click **PYTHONPATH** to highlight it in the list
- With the **PYTHONPATH** variable highlighted, click **Edit**
- Add the following locations to the **Variable Value,** separating each with a semicolon:
  1. C:\PopGen
  2. C:\qgis\python

# Environment Variables



# Starting PopGen

- With all the necessary software installed on the system, we can now launch the PopGen program
- Find the **PopGen.bat** file located in the **gui** folder of **C:\PopGen**
- Double-click the batch file **PopGen.bat** to initialize the application

# Session 3 – Data Structures

PopGen Training Workshop

May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Data Structures

- When providing their own data, users need to follow certain data structure formats
- In this session, you will learn the appropriate formats for specifying your own data in PopGen

# Data Structures

- All the input files have some common features
  - First row contains the variable names
  - Second row contains the variable types
    - Integers – bigint
    - Floating point values – double
    - Strings - text
  - First few columns in any table are compulsory followed by optional columns
  - In the sample file, categories for population attributes should:
    - Start with 1 and subsequent categories are in increments of 1
  - TAZ's should be treated as blockgroups when setting up data

# Data Structure for Household Sample File

- The file provides the household attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno
  - Optional fields – household attributes

| state | pumano | hhid | serialno | <householdvariable1> | <householdvariable2> | <householdvariable3> |
|---|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |

# Data Structure for Groupquarter Sample File

- The file provides the groupquarter attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno
  - Optional fields – groupquarter attributes

| state | pumano | hhid | serialno | <groupquartervariable1> | <groupquartervariable2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Person Sample File

- The file provides the person attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno, pnum
  - Optional fields – person attributes

| state | pumano | hhid | serialno | pnum | <personvariable1> | <personvariable2> |
|---|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |

# Data Structure for Houshold Marginals File

- The file provides the household marginal distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – household marginals distributions

| state | county | tract | bg | <householdvariable1category1> | <householdvariable1category2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Groupquarter Marginals File

- The file provides the groupquarter marginal distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – groupquarter marginal distributions

| state | county | tract | bg | <groupquartervariable1category1> | <groupquartervariable1category2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Person Marginals File

- The file provides the person marginal distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – person marginal distributions

| state | county | tract | bg | <personvariable1category1> | <personvariable1category2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Geographic Correspondence File

- The file provides the correspondence between the geography and the PUMA to which the geography belongs
- Data structure
  - Compulsory fields –county, tract, bg, state, pumano, stateabb, countyname

| county | tract | bg | state | pumano | stateabb | countyname |
|---|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |

# Data Structure for Synthetic Housing File

- The file lists frequencies and ID's of the sample *households/groupquarters* selected in the synthetic population
- Data structure
  - Compulsory fields – state, county, tract, bg, hhid, serialno, frequency, hhuniqueid

| state | county | tract | bg | hhid | serialno | frequency | hhuniqueid |
|---|---|---|---|---|---|---|---|
| value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value |

# Data Structure for Synthetic Person File

- The file lists frequencies and ID's of the sample *persons* selected in the synthetic population
- Data structure
  - Compulsory fields – state, county, tract, bg, hhid, serialno, pnum, frequency, personuniqueid

| state | county | tract | bg | hhid | serialno | pnum | frequency | personuniqueid |
|---|---|---|---|---|---|---|---|---|
| value | value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value | value |

# Resolution

- If you wish to synthesize population at the resolution of
  - County
    - Set  tract and bg columns to 0 in the marginals tables
  - Tract
    - Set bg column to 0 in the marginals table
  - Blockgroup
    - No change
  - TAZ
    - Treat TAZ's as blockgroup

# Data Inconsistencies

- Avoiding data inconsistencies when providing your own data can prevent crashes and improve fit of the synthetic population
- Data Inconsistency Types
  - Person Total Inconsistency: Person total derived from household size distribution is not consistent with the given person total
  - Marginals Distribution Totals: The population totals derived from household/groupquarter/person variables of interest are not consistent

# Data Inconsistencies

- Data Inconsistency Types (continued)
  - Sample Serial ID's
    - Sample serial numbers are not consistent across input files
  - Geographic Correspondence
    - More than one PUMA entry for a particular geography

# Data Structures

- When providing their own data, users need to follow certain data structure formats
- In this session, you will learn the appropriate formats for specifying your own data in PopGen

# Data Structures

- All the input files have some common features
  - First row contains the variable names
  - Second row contains the variable types
    - Integers – bigint
    - Floating point values – double
    - Strings - text
  - First few columns in any table are compulsory followed by optional columns
  - In the sample file, categories for population attributes should:
    - Start with 1 and subsequent categories are in increments of 1
  - TAZ's should be treated as blockgroups when setting up data

# Data Structure for Household Sample File

- The file provides the household attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno
  - Optional fields – household attributes

| state | pumano | hhid | serialno | <householdvariable1> | <householdvariable2> | <householdvariable3> |
|---|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |

# Data Structure for Groupquarter Sample File

- The file provides the groupquarter attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno
  - Optional fields – groupquarter attributes

| state | pumano | hhid | serialno | <groupquartervariable1> | <groupquartervariable2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Person Sample File

- The file provides the person attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno, pnum
  - Optional fields – person attributes

| state | pumano | hhid | serialno | pnum | <personvariable1> | <personvariable2> |
|---|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |

# Data Structure for Houshold Marginals File

- The file provides the household marginal distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – household marginals distributions

| state | county | tract | bg | <householdvariable1category1> | <householdvariable1category2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Groupquarter Marginals File

- The file provides the groupquarter marginals distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – groupquarter marginal distributions

| state | county | tract | bg | <groupquartervariable1category1> | <groupquartervariable1category2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Person Marginals File

- The file provides the person marginal distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – person marginal distributions

| state | county | tract | bg | <personvariable1category1> | <personvariable1category2> |
|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |
| value | value | value | value | value | value |

# Data Structure for Geographic Correspondence File

- The file provides the correspondence between the geography and the PUMA to which the geography belongs
- Data structure
  - Compulsory fields –county, tract, bg, state, pumano, stateabb, countyname

| county | tract | bg | state | pumano | stateabb | countyname |
|---|---|---|---|---|---|---|
| <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> | <variable type> |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |
| value | value | value | value | value | value | value |

# Data Structure for Synthetic Housing File

- The file lists frequencies and ID's of the sample *households/groupquarters* selected in the synthetic population
- Data structure
  - Compulsory fields – state, county, tract, bg, hhid, serialno, frequency, hhuniqueid

| state | county | tract | bg | hhid | serialno | frequency | hhuniqueid |
|---|---|---|---|---|---|---|---|
| value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value |

# Data Structure for Synthetic Person File

- The file lists frequencies and ID's of the sample *persons* selected in the synthetic population
- Data structure
  - Compulsory fields – state, county, tract, bg, hhid, serialno, pnum, frequency, personuniqueid

| state | county | tract | bg | hhid | serialno | pnum | frequency | personuniqueid |
|---|---|---|---|---|---|---|---|---|
| value | value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value | value |
| value | value | value | value | value | value | value | value | value |

# Resolution

- If you wish to synthesize population at the resolution of
  - County
    - Set  tract and bg columns to 0 in the marginals tables
  - Tract
    - Set bg column to 0 in the marginals table
  - Blockgroup
    - No change
  - TAZ
    - Treat TAZ's as blockgroup

# Data Inconsistencies

- Avoiding data inconsistencies when providing your own data can prevent crashes and improve fit of the synthetic population
- Data Inconsistency Types
  - Person Total Inconsistency: Person total derived from household size distribution is not consistent with the given person total
  - Marginals Distribution Totals: The population totals derived from household/groupquarter/person variables of interest are not consistent

# Data Inconsistencies

- Data Inconsistency Types (continued)
  - Sample Serial ID's
    - Sample serial numbers are not consistent across input files
  - Geographic Correspondence
    - More than one PUMA entry for a particular geography

# Session 4 – Project Setup using Census Data

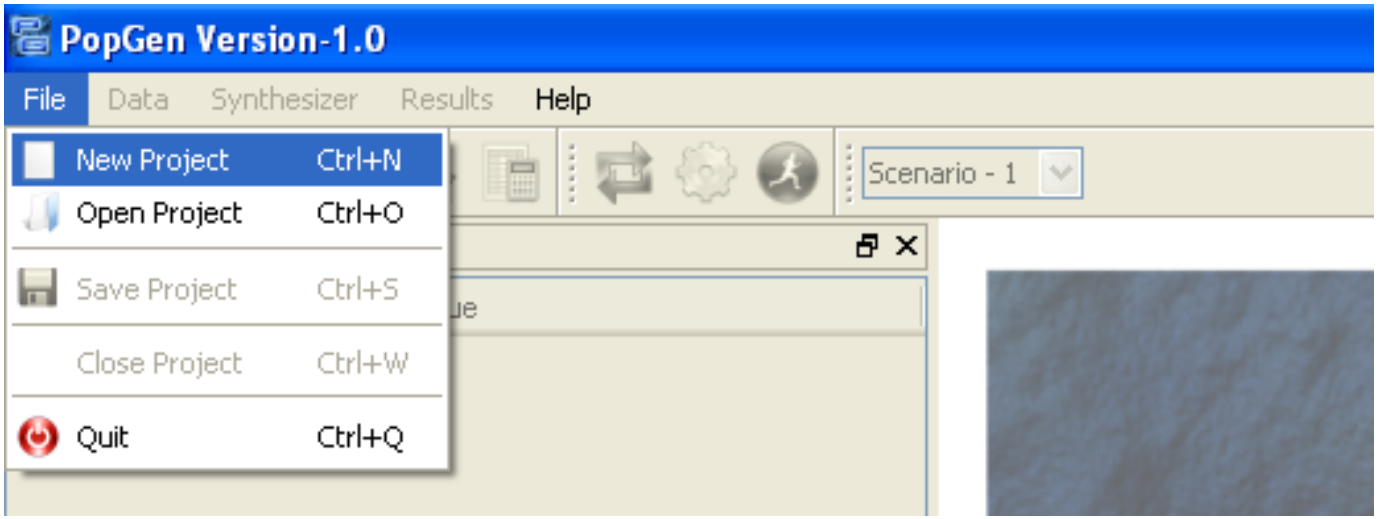
PopGen Training Workshop

May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Create a New Project

- To create a new project, go to **File** ➡ **New Project** on the menu bar at the top or by clicking the **New Project** icon at the top left of the page



# Project Setup Wizard

- In order to create a project and eventually run the population synthesizer, there are 6 steps which must be completed
- The first step is selecting the geographic region you will be using for your population synthesis

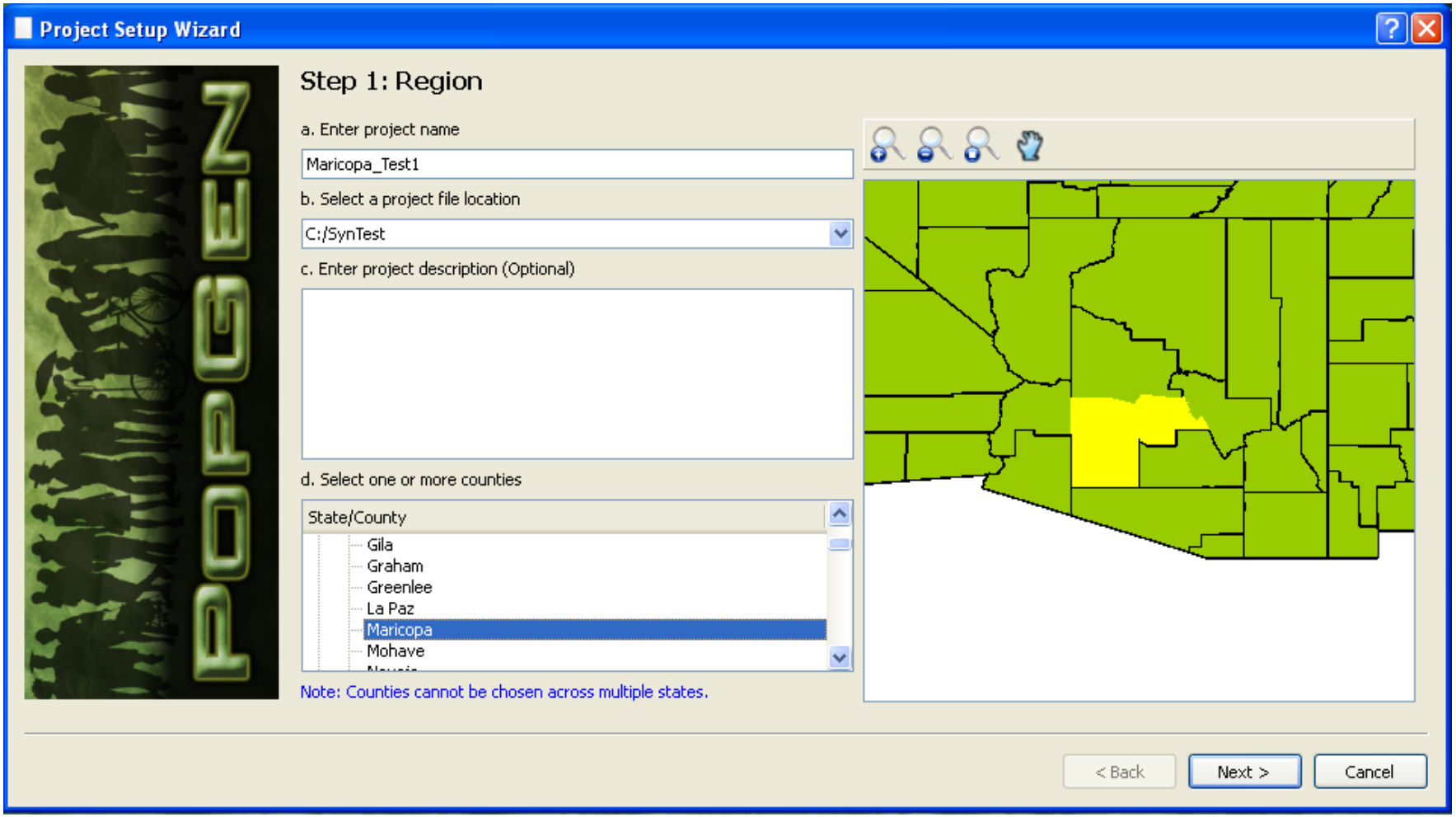
# Step 1 - Region

- Choose a name for your project

  (For this exercise, just name it **Maricopa_Test**)
- Select a project file location on your computer
- Enter a project description (Optional)
- A list of all US counties is displayed beside a map
- Go to **Arizona** and select **Maricopa** as the county

*Note*: If you want to select multiple counties, hold the Ctrl button while selecting. However, counties cannot be selected across multiple states.

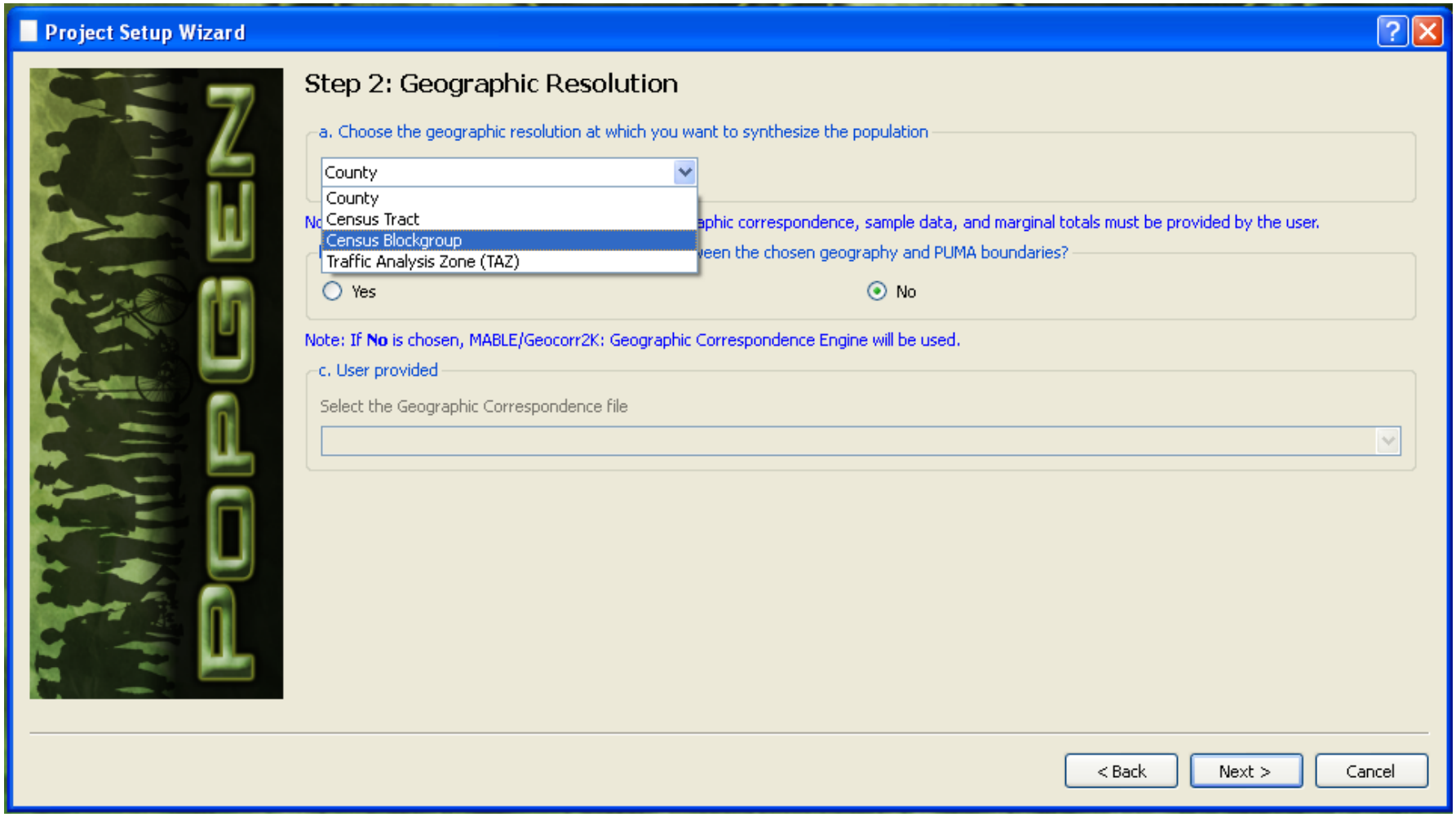
# Step 1 - Region



# Step 2 – Geographic Resolution

- Choose **Census Blockgroup** as the geographic resolution for this exercise
- The standard options are census tract and blockgroup levels
- Any other level (such as a TAZ) may also be used as long as correspondence between that geography and PUMA (from where the population will be sampled) is provided
- Choose **No** for step **2b**, since we will be using the default Census 2000 data for this exercise

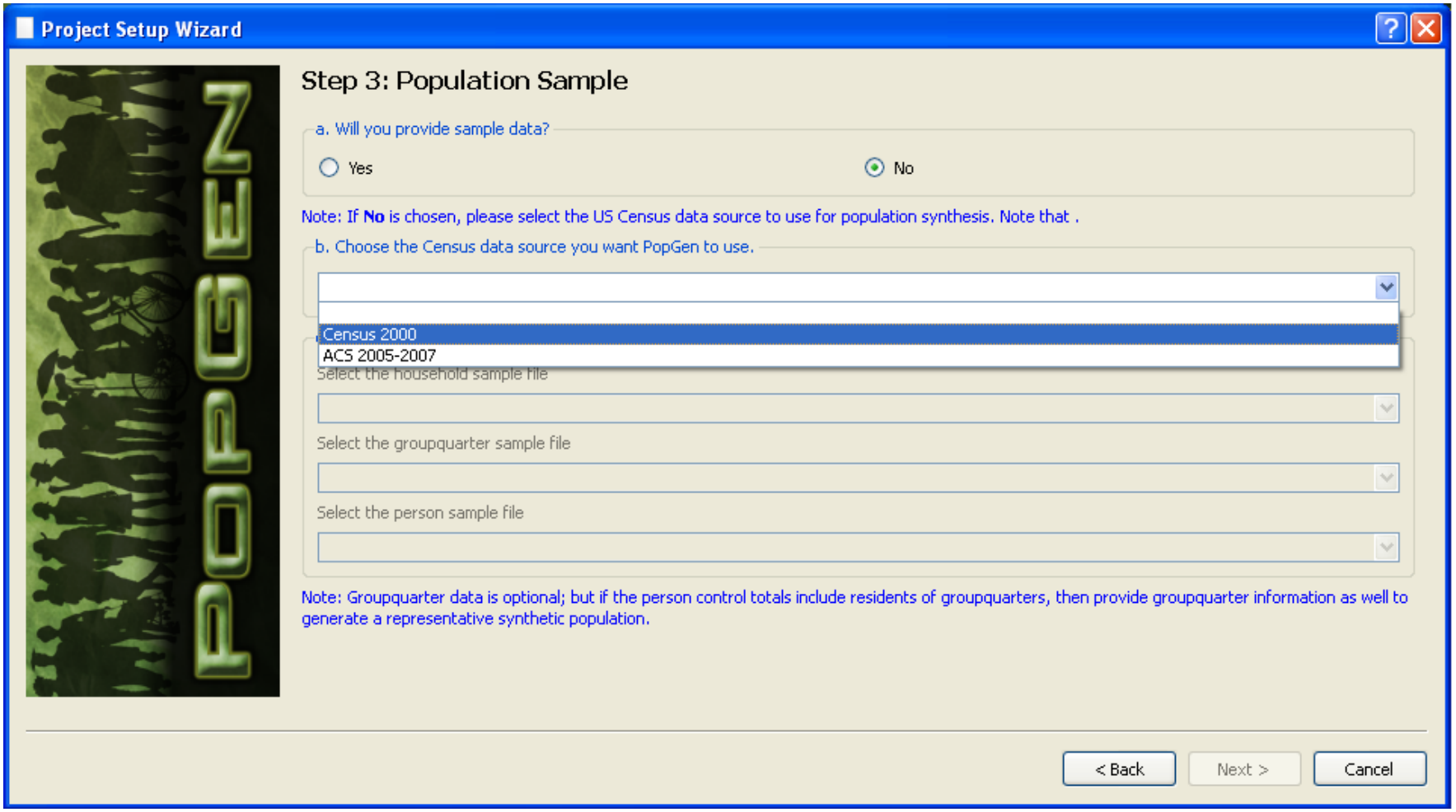
# Step 2 – Geographic Resolution



# Step 3 – Population Sample

- Choose **No** for step **3a**, since we will be using the default sample data
- In step **3b**, choose **Census 2000** for the data source in the drop down menu
- Click **Next** to continue the setup process

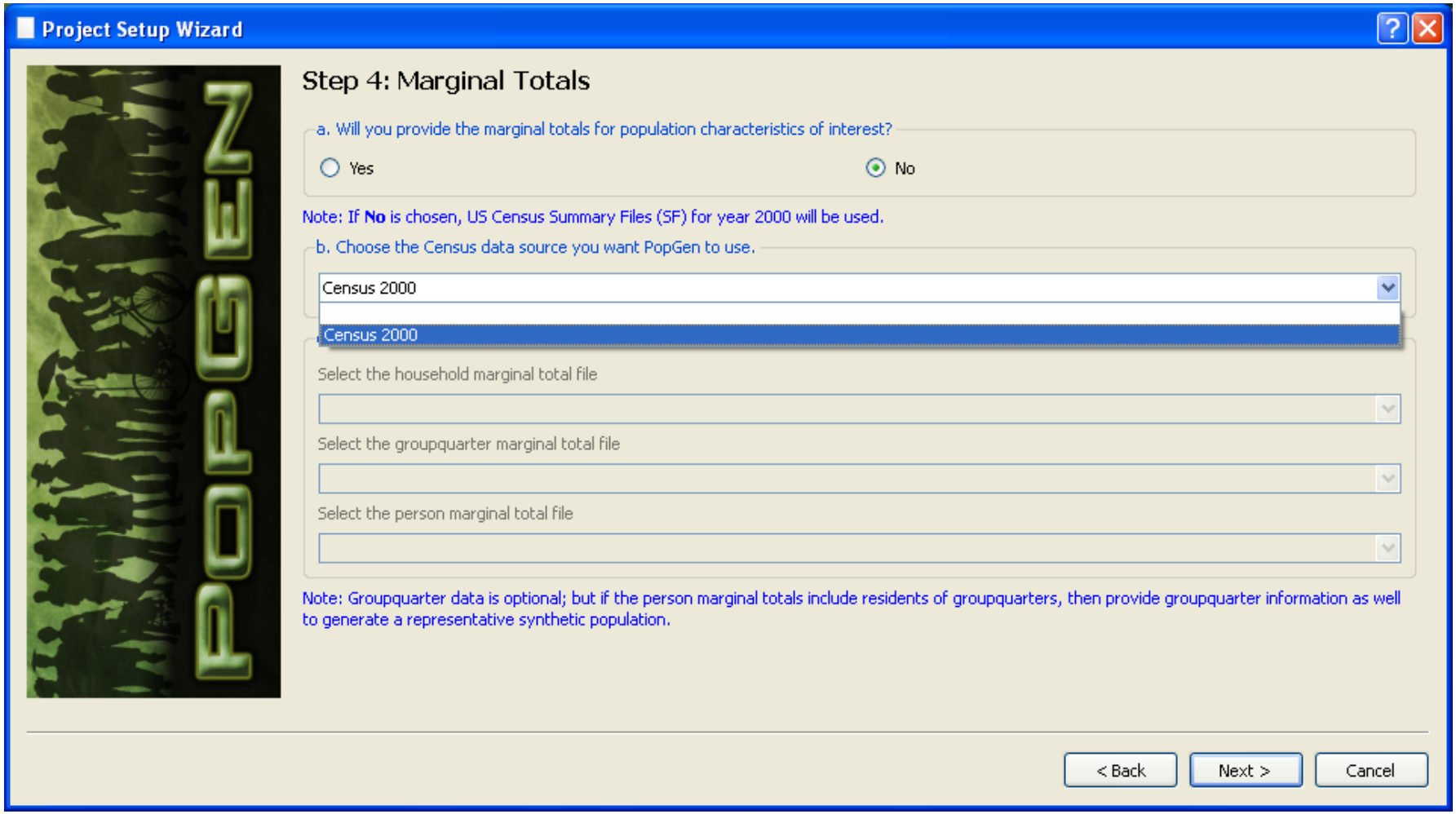
# Step 3 – Population Sample



# Step 4 – Marginal Totals

- Choose **No** for step **4a**, since we will be using the default Census 2000 files
- Choose **Census 2000** in the drop-down menu for step **4b**
- Click **Next** to continue the setup process

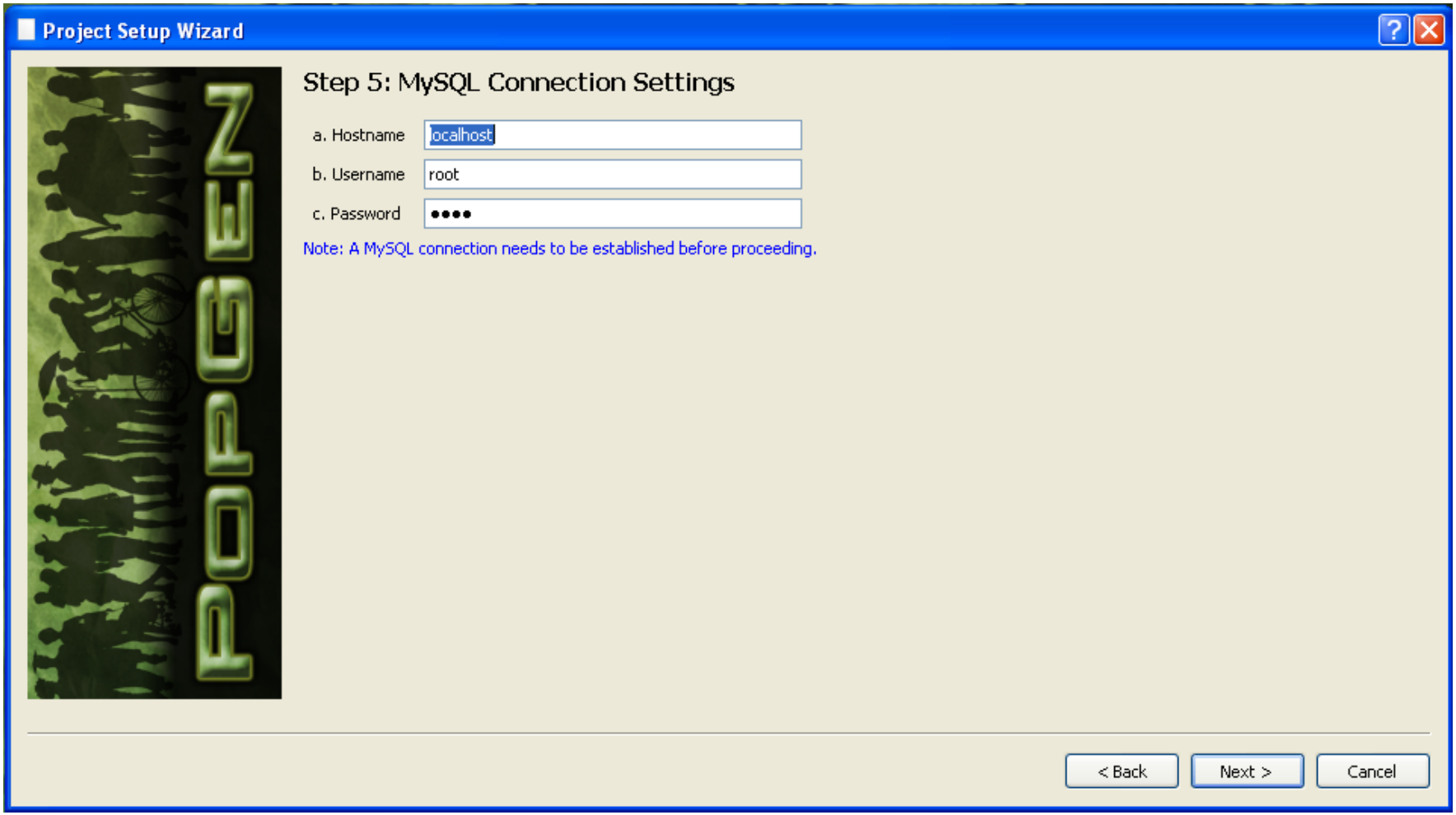
# Step 4 – Marginal Totals



# Step 5 – MySQL Connection Settings

- Choose a Hostname or leave the default as **localhost** for this exercise
- Choose a username or leave the default as **root**
- Choose a password or leave the default as **1234**
- Click **Next** to continue

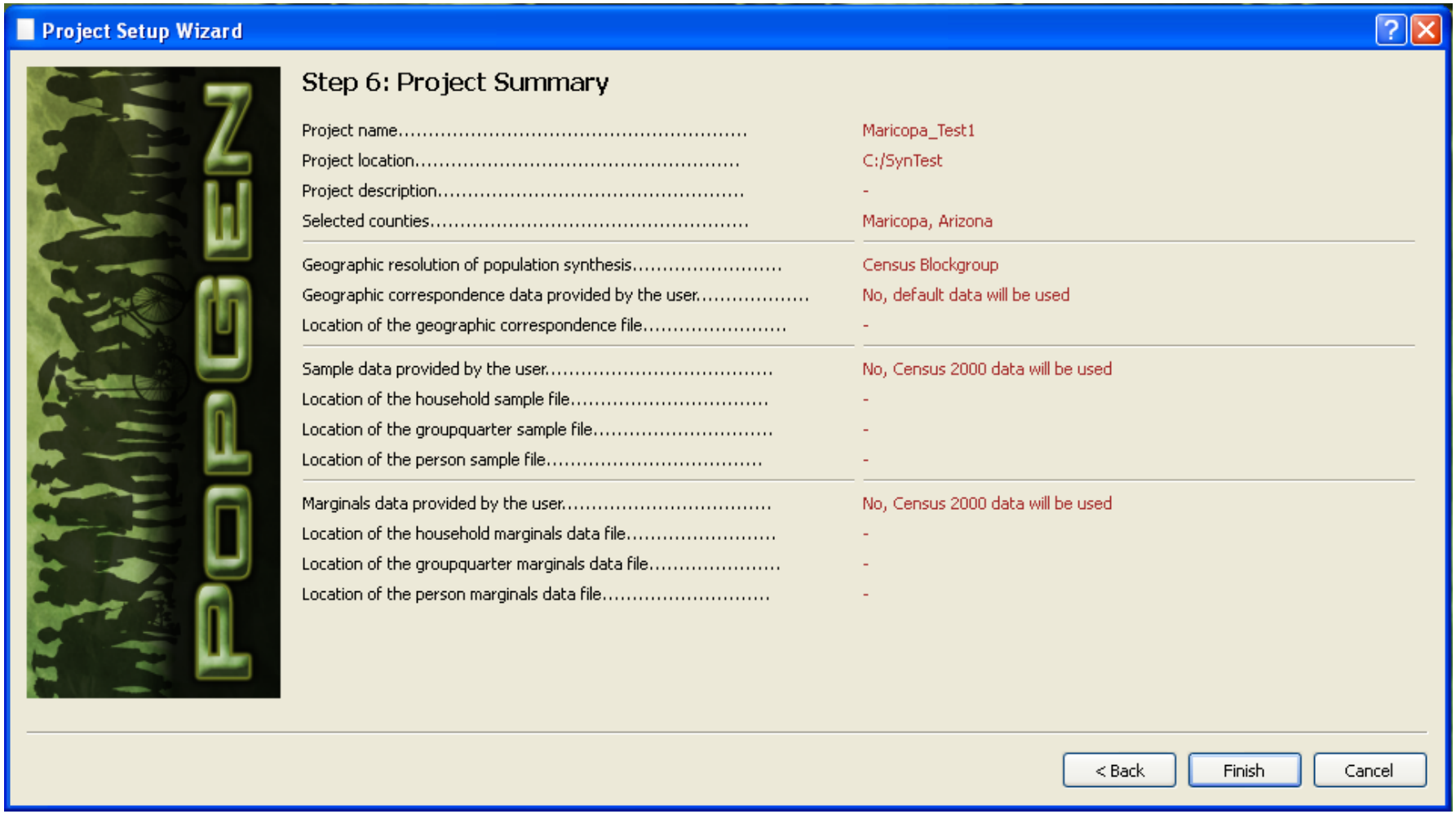
# Step 5 – MySQL Connection Settings



# Step 6 – Project Summary

- A window showing all the previous steps with the chosen options specified is displayed
- Review and finalize that everything is correct
- Click **Finish** to complete the project setup process

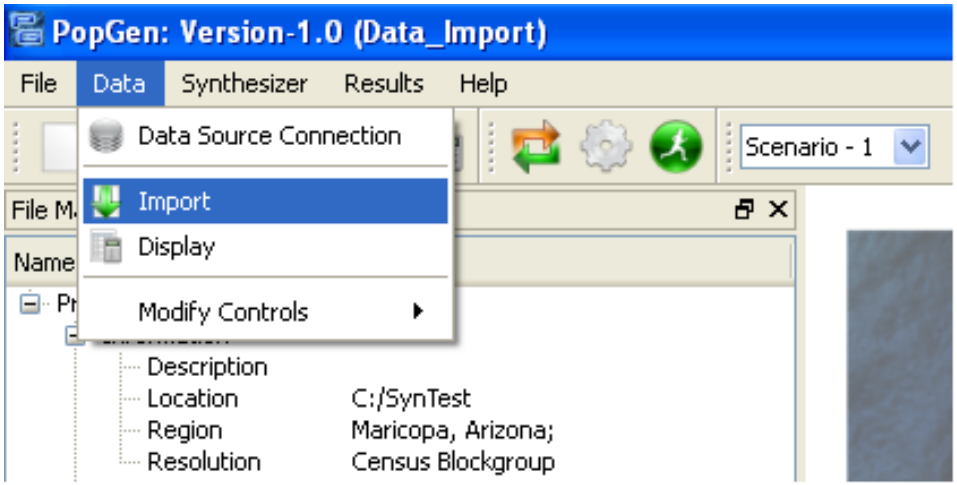
# Step 6 – Project Summary

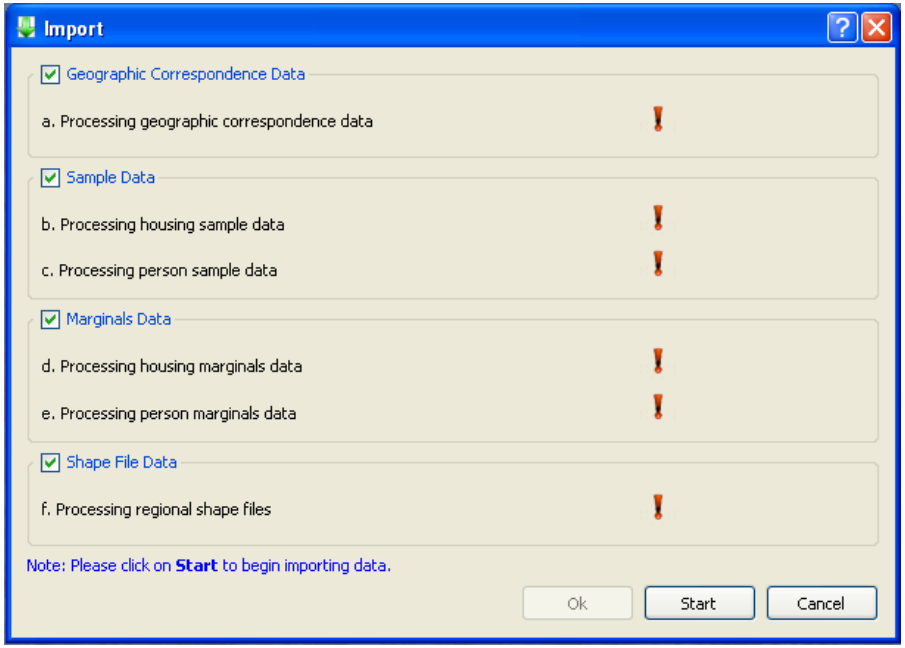


# Data Import

- The data import process has 6 steps:
  1) Processing Geographic Correspondence data
  2) Processing Housing PUMS/ Sample data
  3) Processing Person PUMS/ Sample data
  4) Processing Housing Summary data
  5) Processing Person Summary data
  6) Processing Region's Shape file
- Go to **Data** → **Import** at the top of the screen to start importing data or click the **Data Import icon** in the toolbar

# Data Import

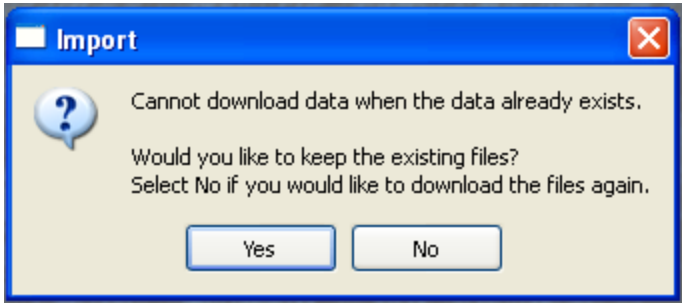

PopGen: Version-1.0 (Data_Import)
File
Data
Synthesizer
Results
Help
Data Source Connection
Import
Display
Modify Controls
Scenario - 1
Name
Description
Location
C:/SynTest
Region
Maricopa, Arizona;
Resolution
Census Blockgroup


Import
Geographic Correspondence Data
a. Processing geographic correspondence data
Sample Data
b. Processing housing sample data
c. Processing person sample data
Marginals Data
d. Processing housing marginals data
e. Processing person marginals data
Shape File Data
f. Processing regional shape files
Note: Please click on Start to begin importing data.
Ok
Start
Cancel

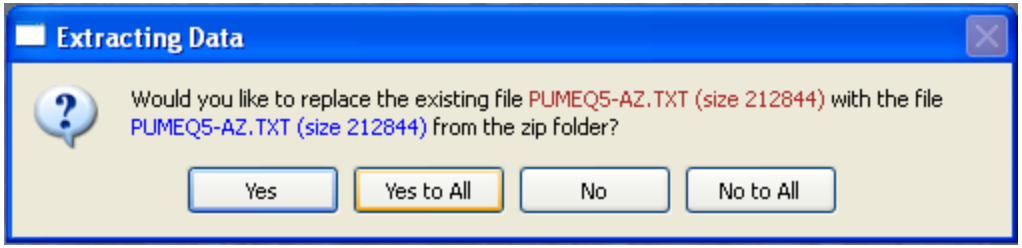
# Data Import

- Click **Start** to begin the import process
- The first 5 steps downloads the census files, parses them, and then uploads them into the MySQL database as tables
- While importing, you will be prompted to overwrite any files if they already exist and also for the selection of the specific variables to be imported into the database
- The final step downloads and extracts the shape files containing the boundaries of the geographies selected

# Data Import




- If you are prompted to keep data files that already exist, you may choose to download the data again (Choose No), or keep the existing data (Choose Yes) that has already been downloaded

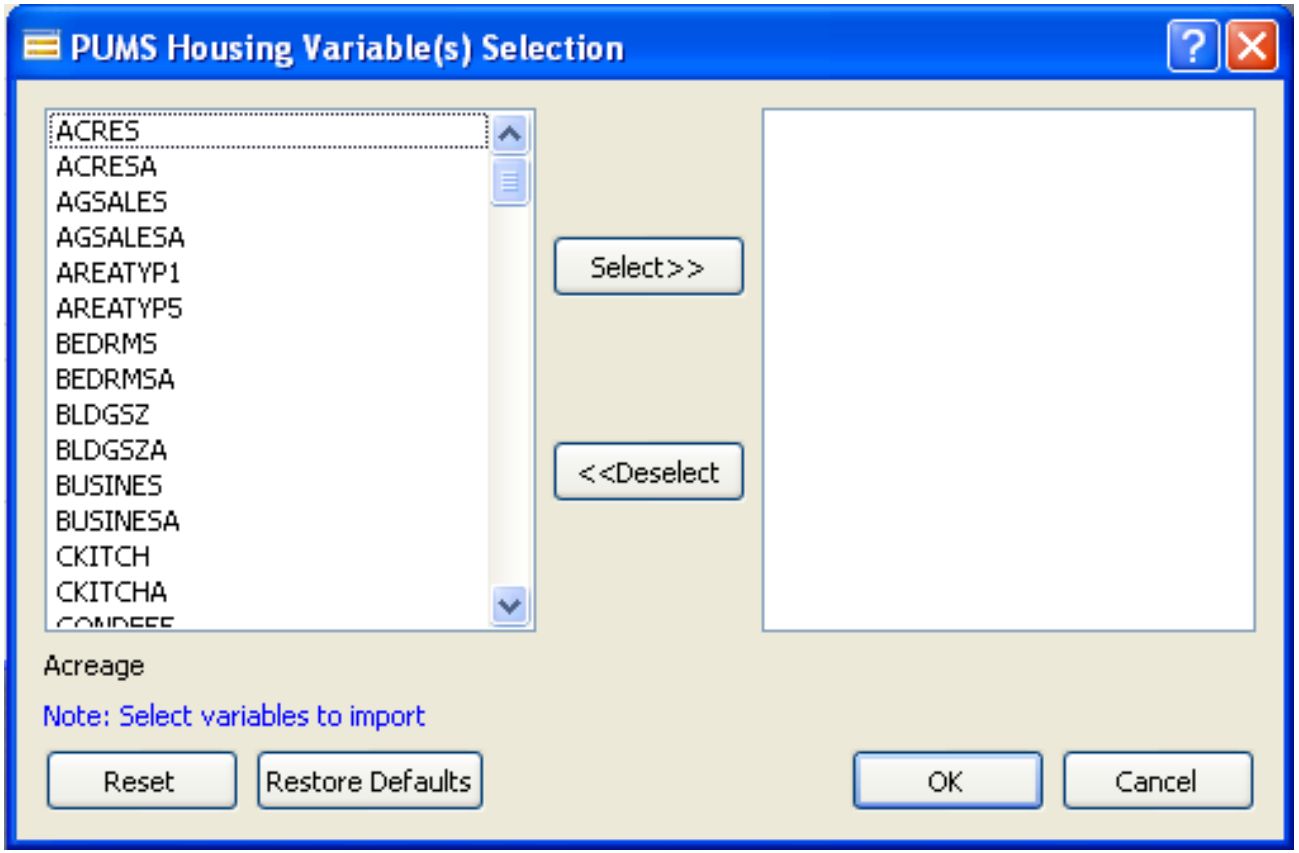



- If you are prompted to replace existing files, you may choose to overwrite them all (No to All) or replace them (Yes to All)

# Data Import

- While importing the data, there will be a series of variable selection windows that will appear
- For each one of these windows, select the **Restore Defaults** option at the bottom left corner of the window
- This option will automatically select the default variables from the master list of variables

# Data Import


PUMS Housing Variable(s) Selection
ACRES
ACRESA
AGSALES
AGSALESA
AREATYP1
AREATYP5
BEDRMS
BEDRMSA
BLDGSZ
BLDGSZA
BUSINES
BUSINESA
CKITCH
CKITCHA
Select>>
<<Deselect
Acreage
Note: Select variables to import
Reset
Restore Defaults
OK
Cancel

# Data Import

- When the Data Import process is complete, click **OK** to continue
- The data that has been imported will now appear in the **File Manager** on the left hand side of the screen

# Session 5 – Project Setup using User Provided Data

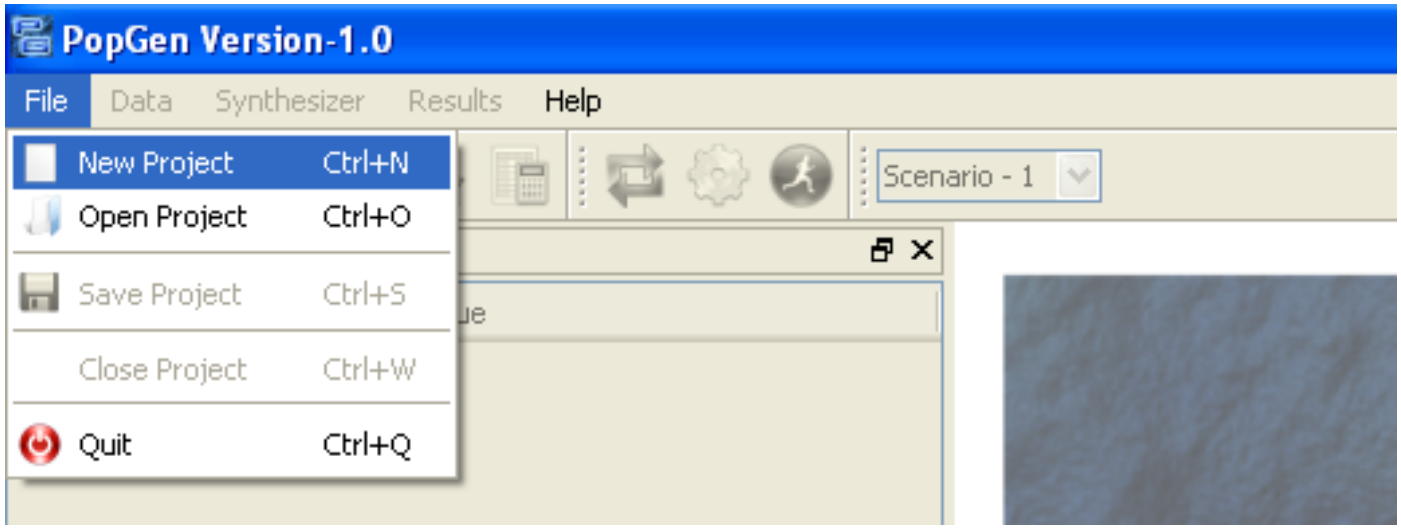
PopGen Training Workshop

May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Create a New Project

- To create a new project, go to **File** ➡ **New Project** on the menu bar at the top or by clicking the **New Project** icon at the top left of the page



# Project Setup Wizard

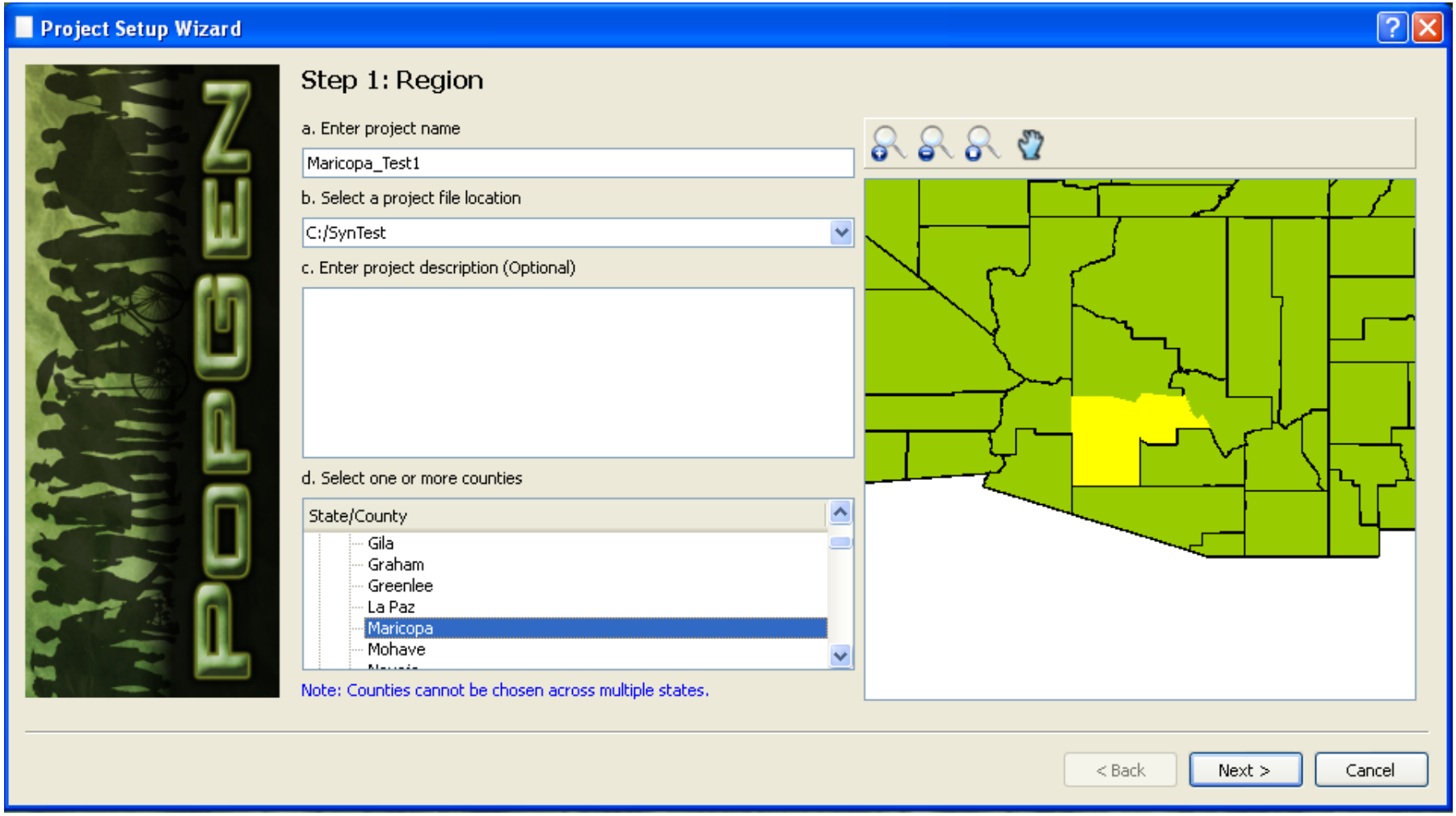
- In order to create a project and eventually run the population synthesizer, there are 6 steps which must be completed
- The first step is selecting the geographic region you will be using for your population synthesis

# Step 1 - Region

- Choose a name for your project
  (For example: **Maricopa_User**)
- Select a project file location on your computer
- Enter a project description (Optional)
- Select the county of your choice

*Note*: If you want to select multiple counties, hold the Ctrl button while selecting. However, counties cannot be selected across multiple states.

# Step 1 - Region



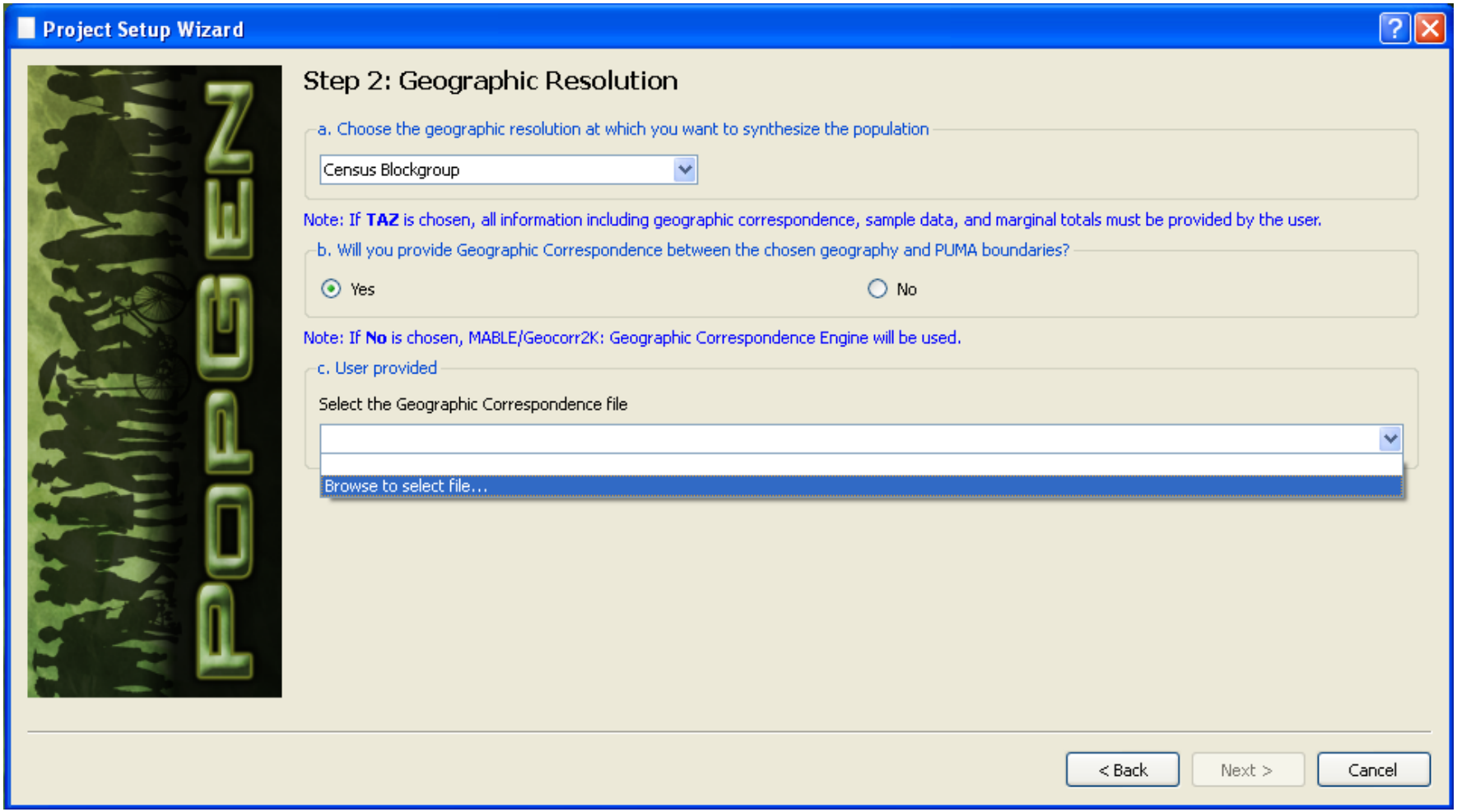
# Step 2 – Geographic Resolution

- Choose **Census Blockgroup** as the geographic resolution for this exercise
- The standard options are census tract and blockgroup levels
- Any other level (such as a TAZ) may also be used as long as correspondence between that geography and PUMA (from where the population will be sampled) is provided

# Step 2 – Geographic Resolution

- Choose **Yes** for step **2b**, since we will be using our own data
- In the drop down menu, go to **Browse to select file**
- Find the appropriate geographic correspondence file for the data you will be using and select it
- Click **Next** to continue

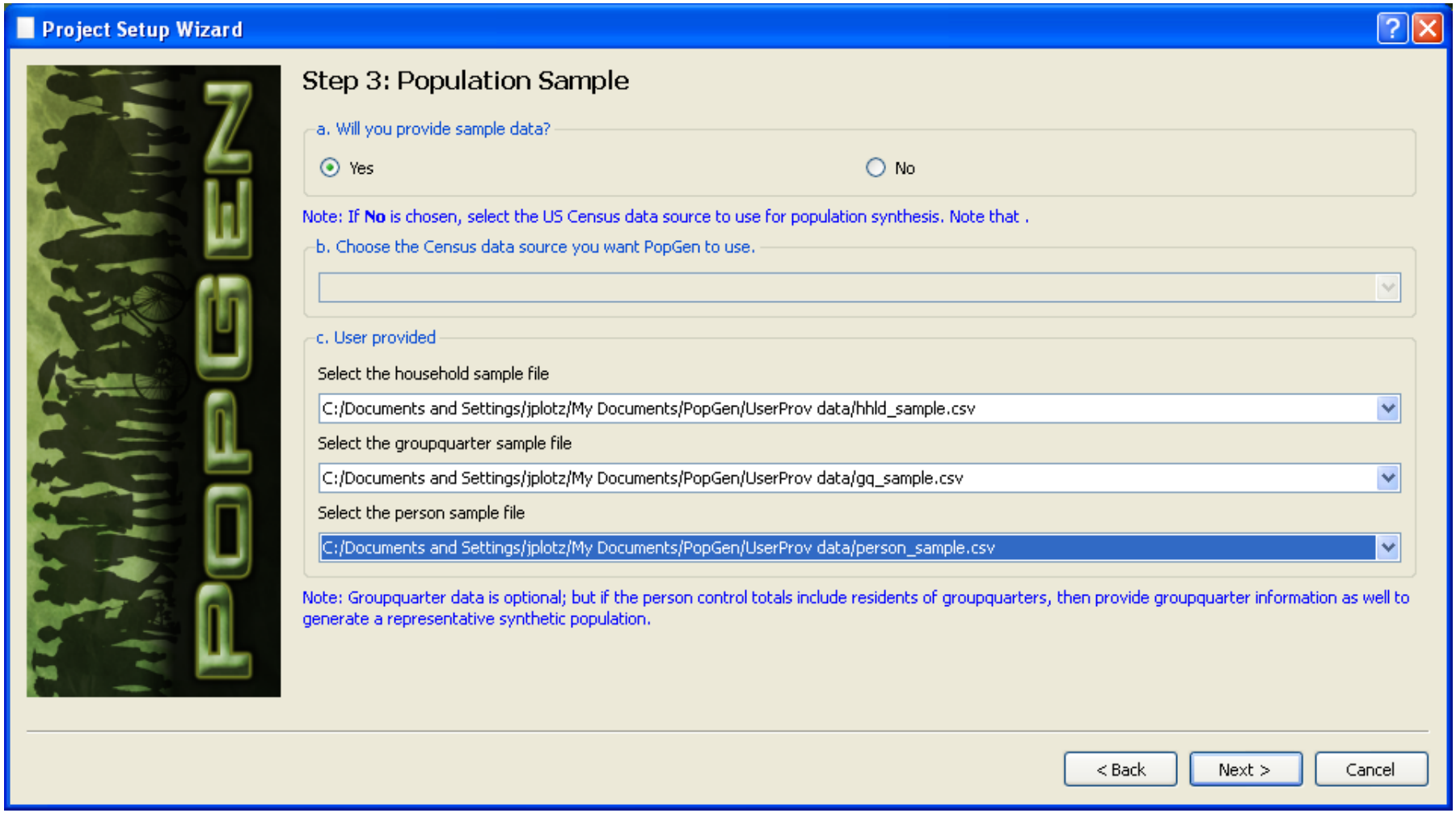
# Step 2 – Geographic Resolution



# Step 3 – Population Sample

- Choose **Yes** for step **3a** since we will be using our own data
- Step **3b** is grayed out since we are not using the default Census data
- Select the appropriate files from your computer for the **Household** and **Person** Sample Files
  - The Groupquarter Sample file is not required but if your person control totals include residents of groupquarters, it is necessary to provide the Groupquarter Sample File in order to generate a representative synthetic population
- Click **Next** to continue

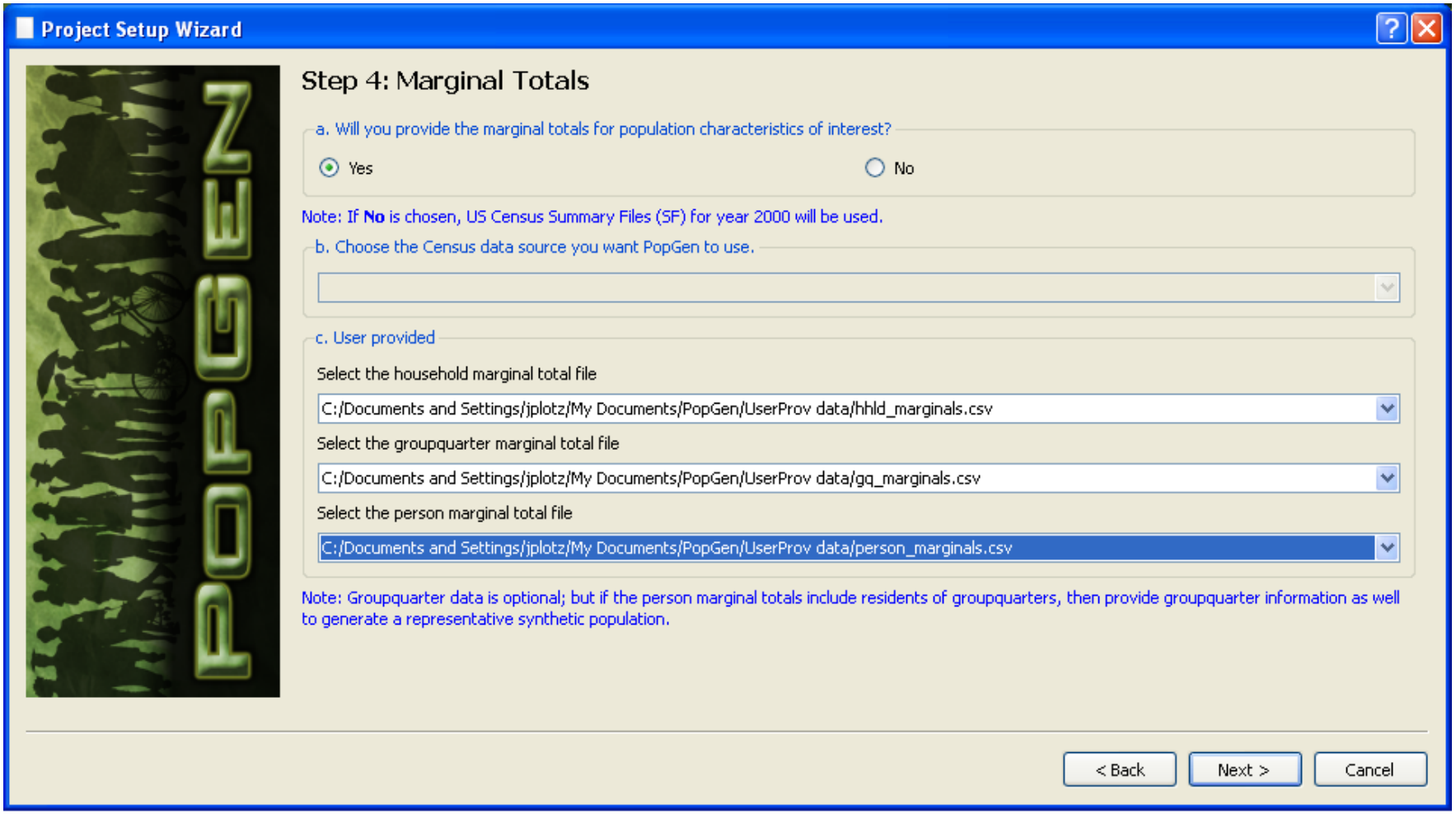
# Step 3 – Population Sample



# Step 4 – Marginal Totals

- Choose **Yes** for step **4a**, since we will be using our own data
- Step **3b** is grayed out since we are not using the default Census data
- Select the appropriate files from your computer for the **Household** and **Person** marginal total files
  - The **Groupquarter** marginal total file is not required, but if you know that your **Person** marginal totals include residents of groupquarters, it is necessary to provide the Groupquarter marginal totals in order to generate a representative synthetic population
- Click **Next** to continue

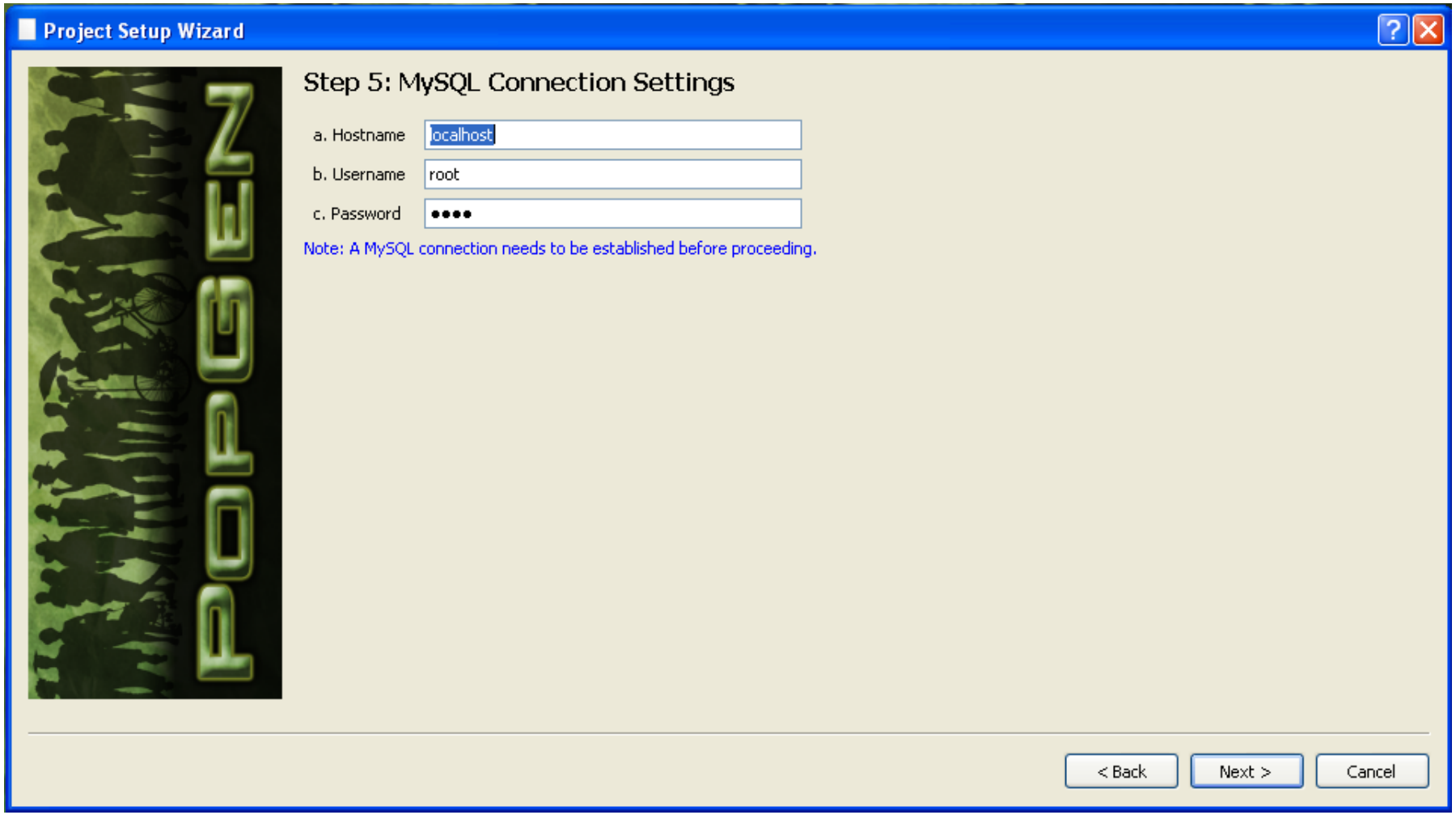
# Step 4 – Marginal Totals



# Step 5 – MySQL Connection Settings

- Choose a Hostname or leave the default as **localhost** for this exercise
- Choose a username or leave the default as **root**
- Choose a password or leave the default as **1234**
- Click **Next** to continue

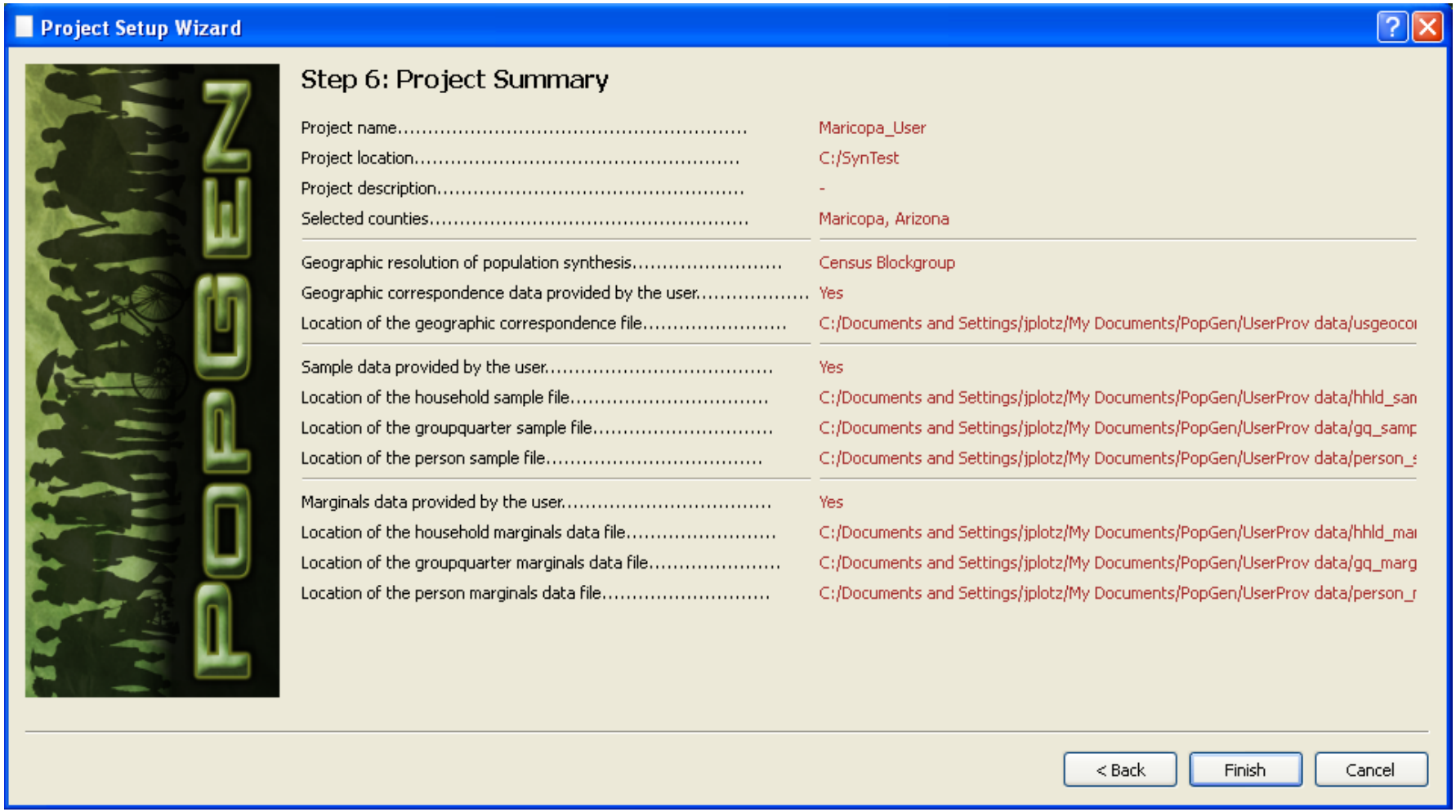
# Step 5 – MySQL Connection Settings



# Step 6 – Project Summary

- A window showing all the previous steps with the chosen options specified is displayed
- Review and finalize that everything is correct
- Click **Finish** to complete the project setup process

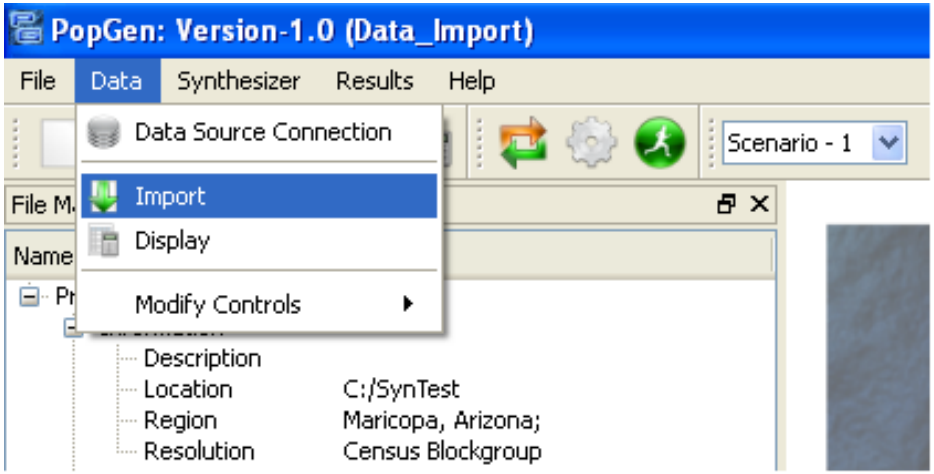
# Step 6 – Project Summary

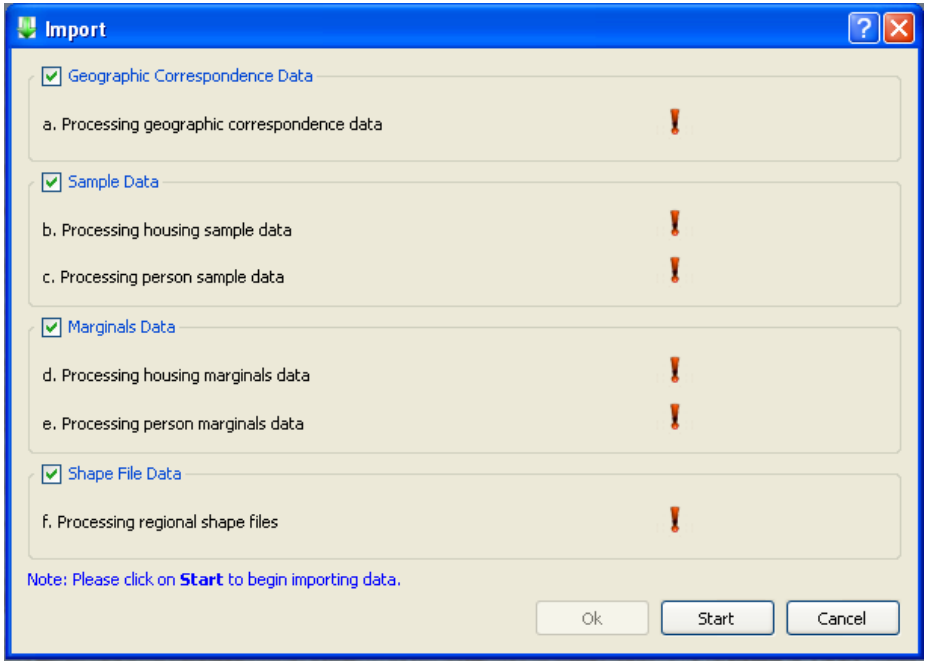


# Data Import

- The data import process has 6 steps:
  1) Processing Geographic Correspondence data
  2) Processing Housing PUMS/ Sample data
  3) Processing Person PUMS/ Sample data
  4) Processing Housing Summary data
  5) Processing Person Summary data
  6) Processing Region's Shape file
- Go to **Data** → **Import** at the top of the screen to start importing data or click the **Data Import icon** in the toolbar

# Data Import

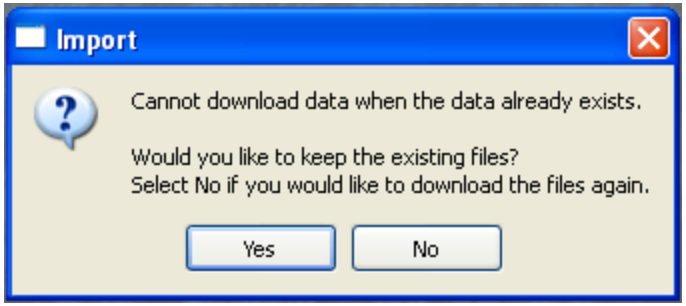

PopGen: Version-1.0 (Data_Import)
File
Data
Synthesizer
Results
Help
Data Source Connection
Import
Display
Modify Controls
Scenario - 1
File M
Name
Description
Location
C:/SynTest
Region
Maricopa, Arizona;
Resolution
Census Blockgroup


Import
Geographic Correspondence Data
a. Processing geographic correspondence data
Sample Data
b. Processing housing sample data
c. Processing person sample data
Marginals Data
d. Processing housing marginals data
e. Processing person marginals data
Shape File Data
f. Processing regional shape files
Note: Please click on Start to begin importing data.
Ok
Start
Cancel

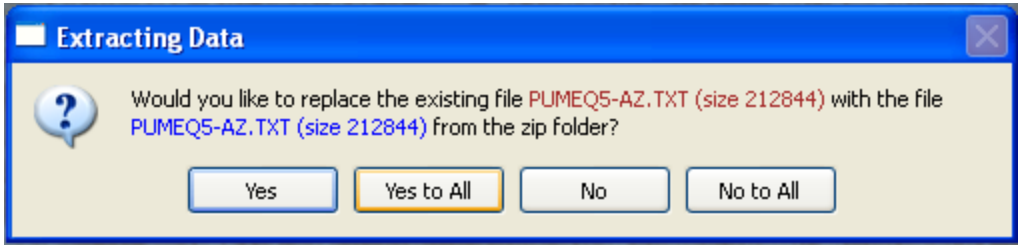
# Data Import

- Click **Start** to begin the import process
- While importing, you will be prompted to overwrite any files if they already exist and also for the selection of the specific variables to be imported into the database
- The final step downloads and extracts the shape files containing the boundaries of the geographies selected

# Data Import




- If you are prompted to keep data files that already exist, you may choose to download the data again (Choose No), or keep the existing data (Choose Yes) that has already been downloaded




- If you are prompted to replace existing files that already exist, you may choose to overwrite them all (No to All) or replace them (Yes to All)

# Data Import

- When the Data Import process is complete, click **OK** to continue
- The data that has been imported will now appear in the **File Manager** on the left hand side of the screen

# Session 6 – Modifying Data

PopGen Training Workshop

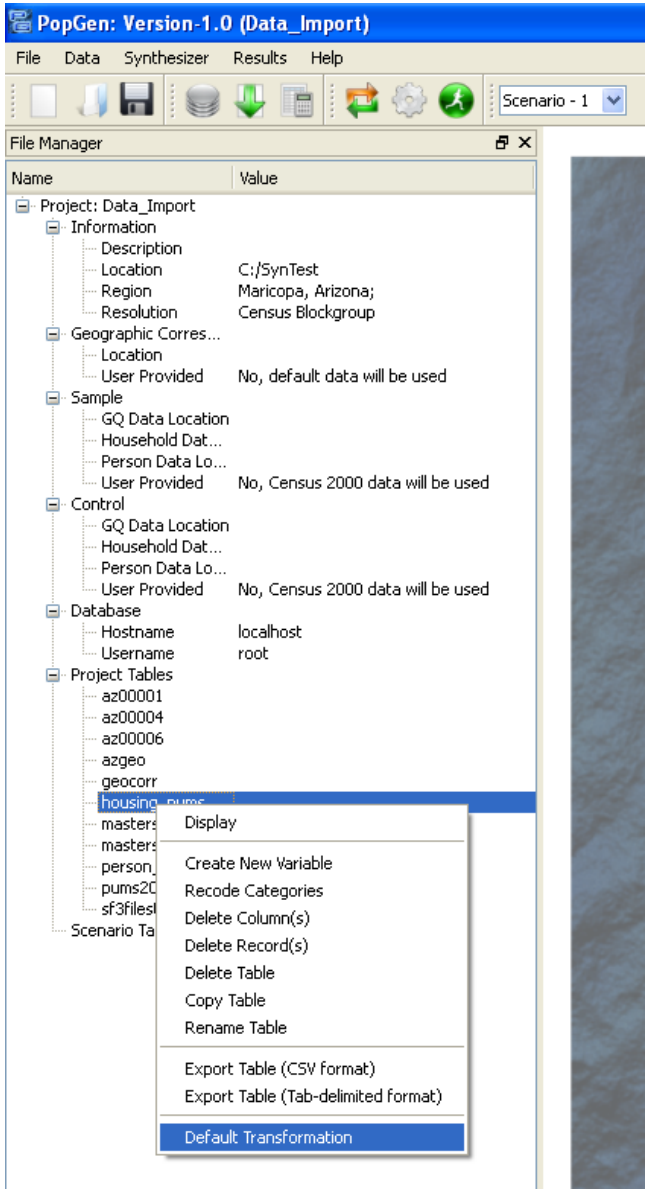
May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Default Transformation

- Now that the data has been imported, we need to run default transformations on the data
- Right-click **housing_pums** under **Project Tables** in the **File Manager**
- Select **Default Transformation** at the bottom
- The default transformation will take a few moments to complete
- Repeat this step for **person_pums** and **mastersftableblockgroup** also located under **Project Tables**

# Default Transformation



# Displaying a Data Table

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Display** to open the data table
- You can view the descriptives or frequencies by clicking on the appropriate button

# Displaying a Data Table

Data Table - hhld_sample

| | state | pumano | hhid | serialno | hhtype | hhldtype | hhldinc | hl |
|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 300 | 1 | 4329 | 1 | 1 | 6 | 3 |
| 2 | 4 | 300 | 3 | 5913 | 1 | 1 | 4 | 4 |
| 3 | 4 | 300 | 4 | 7405 | 1 | 3 | 1 | 3 |
| 4 | 4 | 300 | 6 | 8026 | 1 | 1 | 2 | 5 |
| 5 | 4 | 300 | 9 | 11018 | 1 | 3 | 2 | 6 |
| 6 | 4 | 300 | 10 | 11480 | 1 | 1 | 5 | 4 |
| 7 | 4 | 300 | 12 | 15493 | 1 | 2 | 6 | 2 |
| 8 | 4 | 300 | 13 | 16412 | 1 | 3 | 1 | 5 |
| 9 | 4 | 300 | 14 | 16657 | 1 | 1 | 3 | 7 |
| 10 | 4 | 300 | 18 | 20325 | 1 | 1 | 3 | 3 |
| 11 | 4 | 300 | 21 | 25181 | 1 | 5 | 1 | 1 |
| 12 | 4 | 300 | 22 | 26191 | 1 | 1 | 5 | 2 |
| 13 | 4 | 300 | 23 | 27759 | 1 | 1 | 2 | 4 |
| 14 | 4 | 300 | 26 | 35740 | 1 | 1 | 5 | 2 |
| 15 | 4 | 300 | 27 | 39709 | 1 | 1 | 3 | 7 |

Descriptives

Frequencies

Output

OK

Cancel

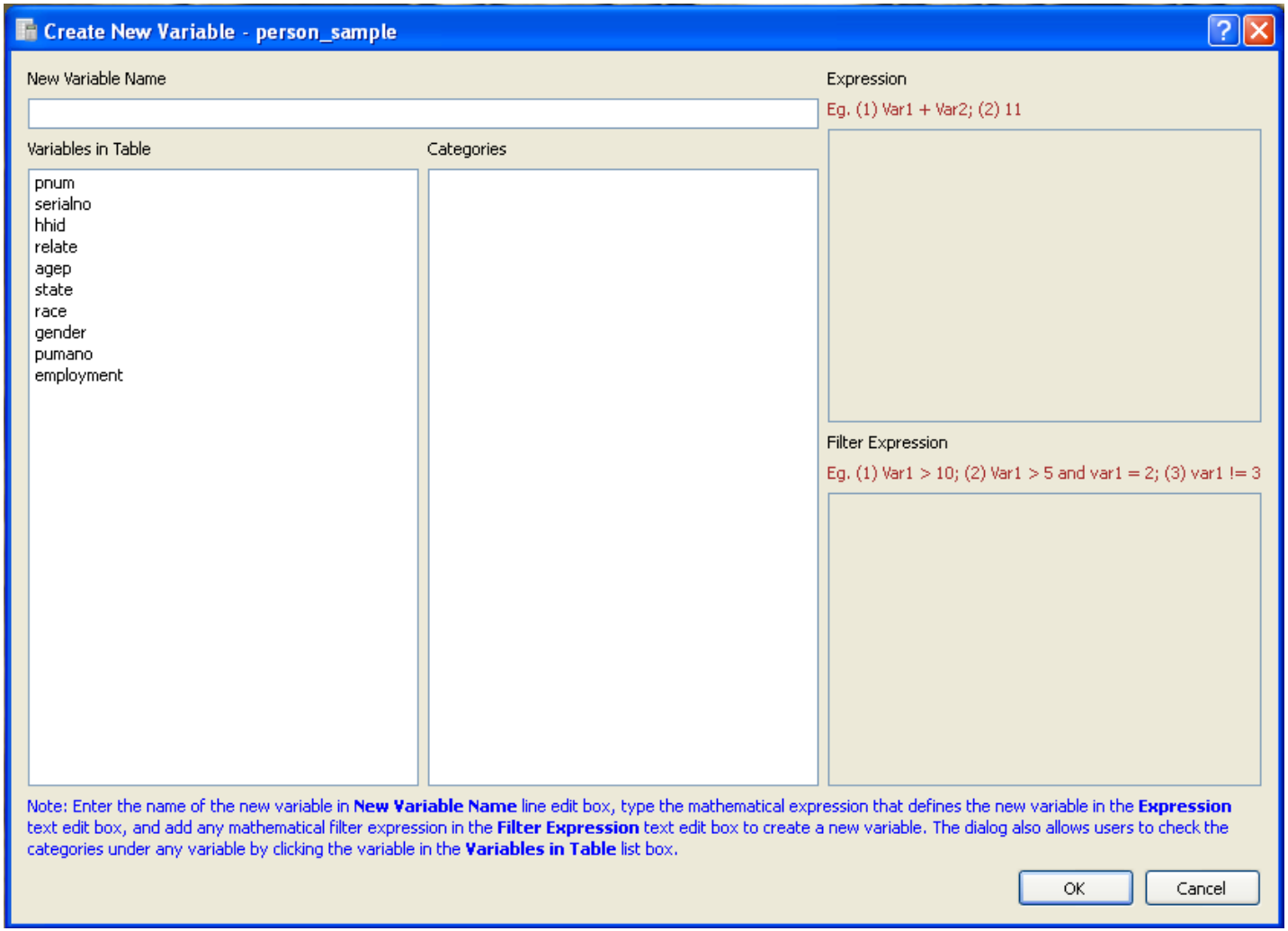
# Create a New Variable

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Create New Variable** from the list
- Give a name to the new variable

# Create a New Variable

- Define the new variable by giving it a mathematical expression in the **Expression** text box
- Add any mathematical filter expressions in the **Filter Expression** text box
- You can also reference the different categories of each variable by clicking the variable

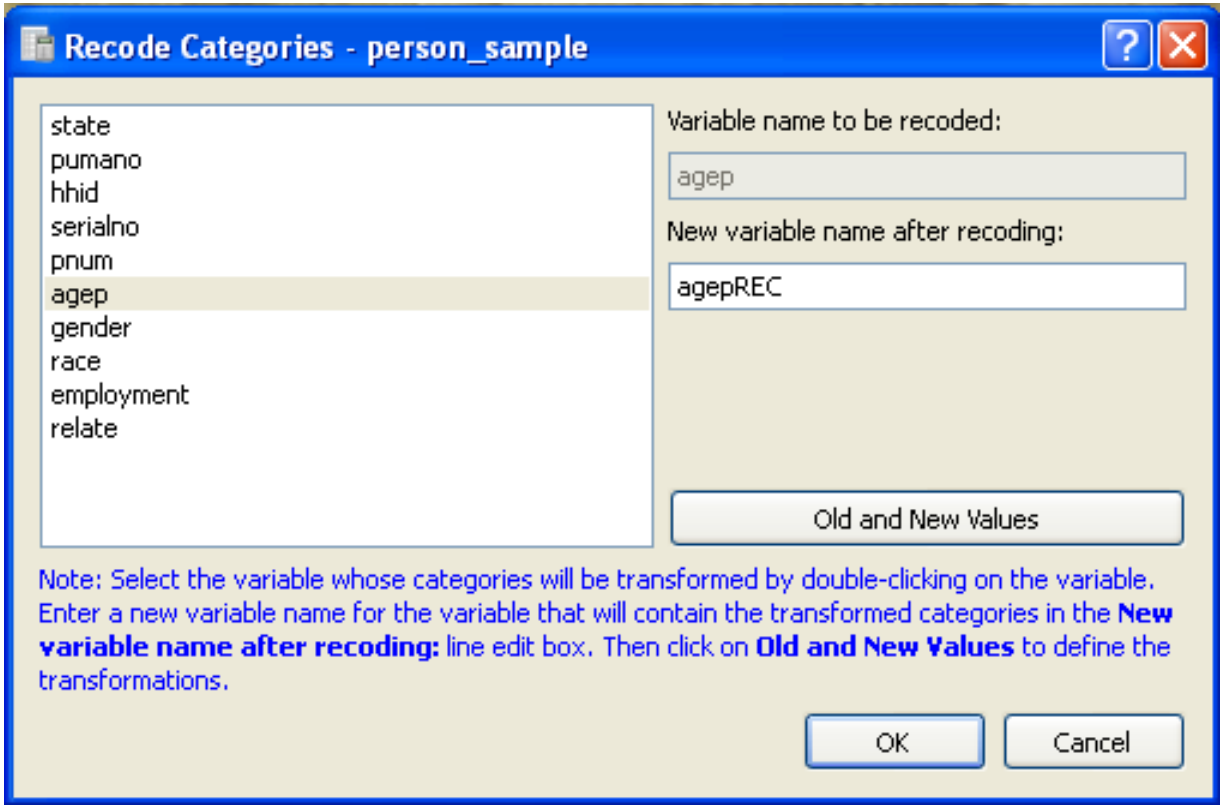
# Create a New Variable



# Recode Categories

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Recode Categories** from the list
- Double-click the variable you wish to recode
- Give a new name to the variable being recoded

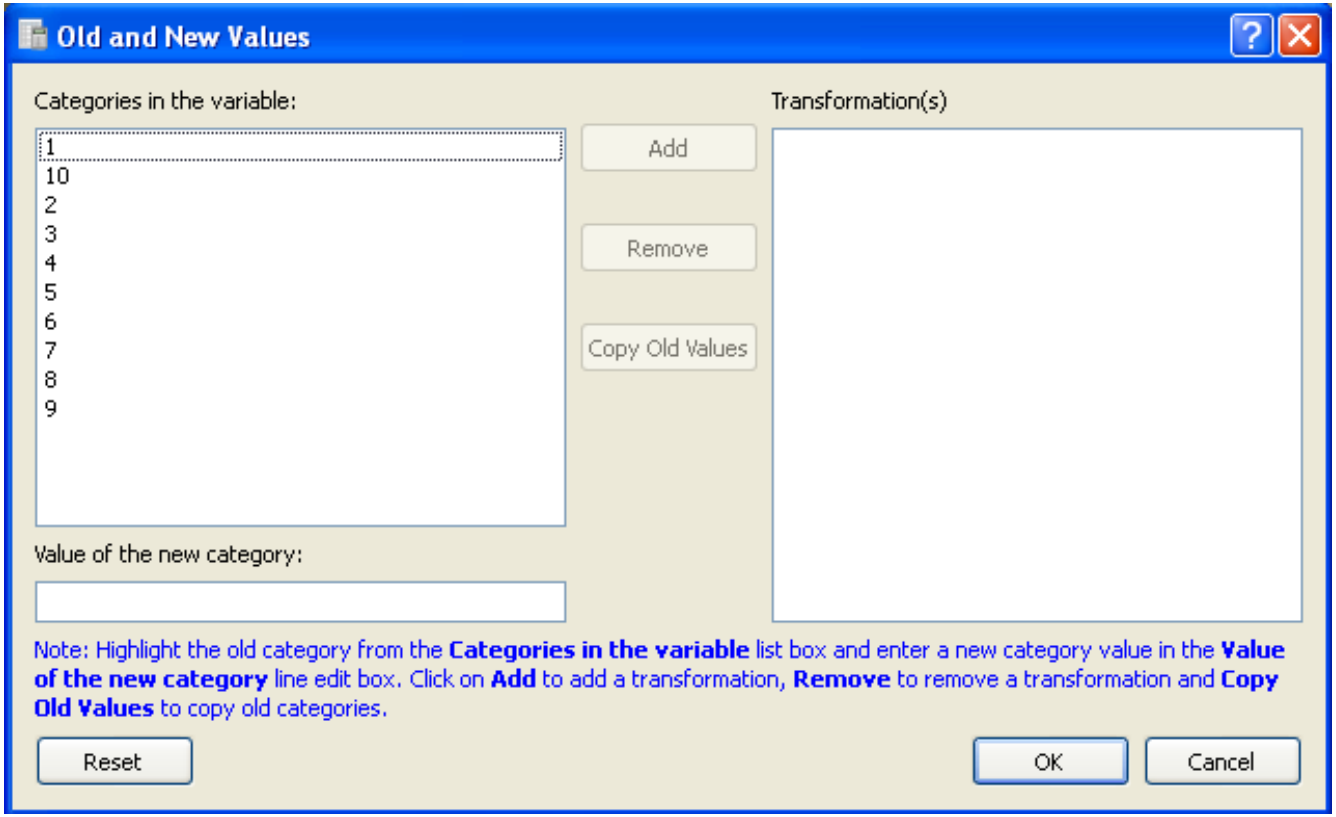
# Recode Categories



# Recode Categories

- Click **Old and New Values** to define the transformations
- Highlight the old category in the **Categories in the Variable** text box and assign it a new value in the **Value of the New Category** text box
- Click **Add** to add the transformation and repeat, if necessary, for other categories
- When all your transformations are shown in the **Transformation** text box, click **OK**

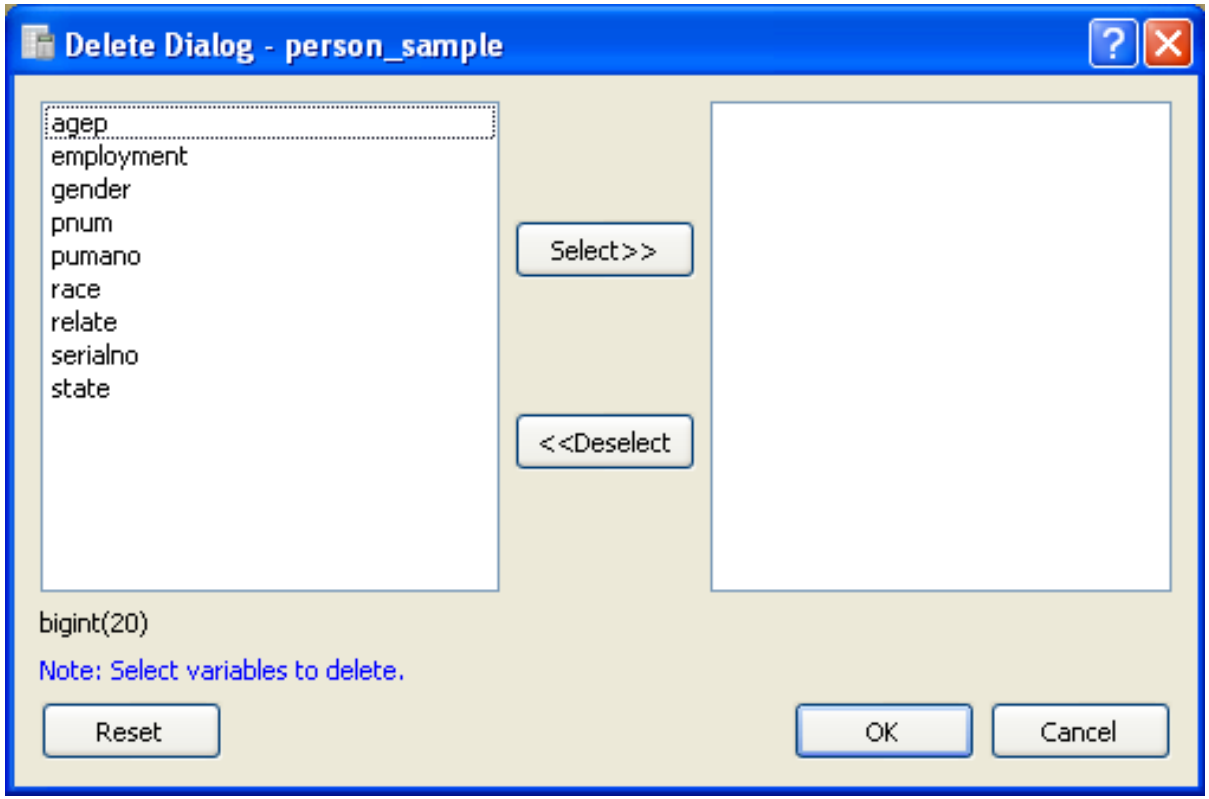
# Recode Categories



# Delete a Column

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Delete Column(s)** from the list
- Choose the variable(s) from the list that you wish to delete and click **Select**
- Click **OK** to continue

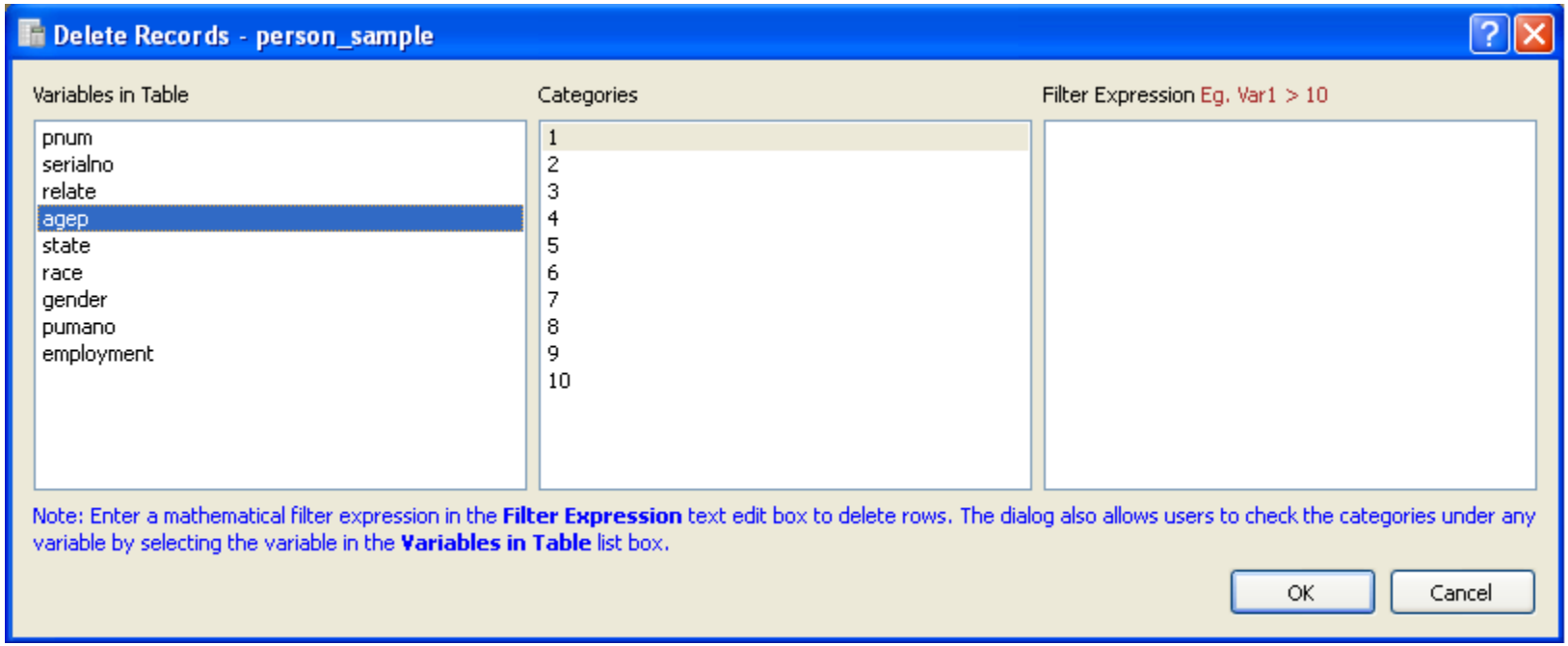
# Delete a Column



# Delete a Record

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Delete Record(s)** from the list
- Enter a mathematical filter expression in the **Filter Expression** text box to delete rows
- Click **OK** when you are finished

# Delete a Record



# Delete a Data Table

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Delete Table** from the list
- A prompt window will appear asking if you wish to continue with the deletion
- Click **Yes** if you are sure that you want to delete the data table

# Copy a Data Table

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Copy Table** from the list
- Enter a name for the table and click **OK**

# Rename a Data Table

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select **Rename Table** from the list
- Enter a new name for the table and click **OK**

# Exporting Data Tables

- Right-click the table of your choice under **Project Tables** or **Scenario Tables** in the **File Manager**
- Select either Export Table **(CSV format)** or Export Table **(Tab de-limited format)** from the list
- Browse for a folder where you would like to save the data table and click **OK**

# Session 7 – Configuring a Project

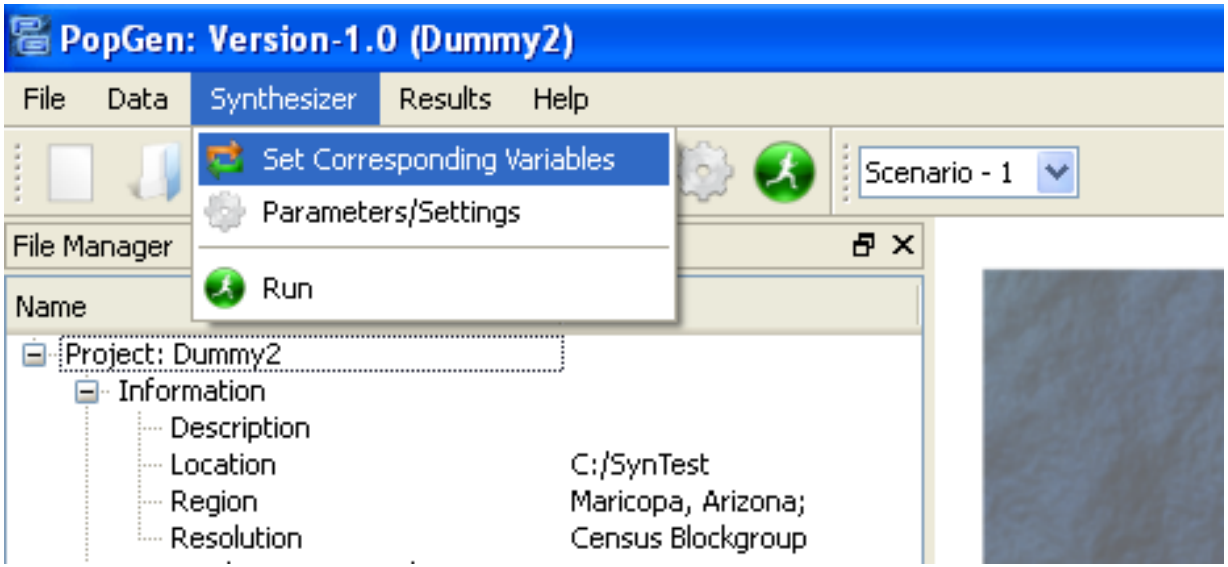
PopGen Training Workshop

May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Set Corresponding Variables

- Next, we need to set the corresponding variables before running the synthesizer
- Go to **Synthesizer** ➡ **Set Corresponding Variables** in the menu bar at the top

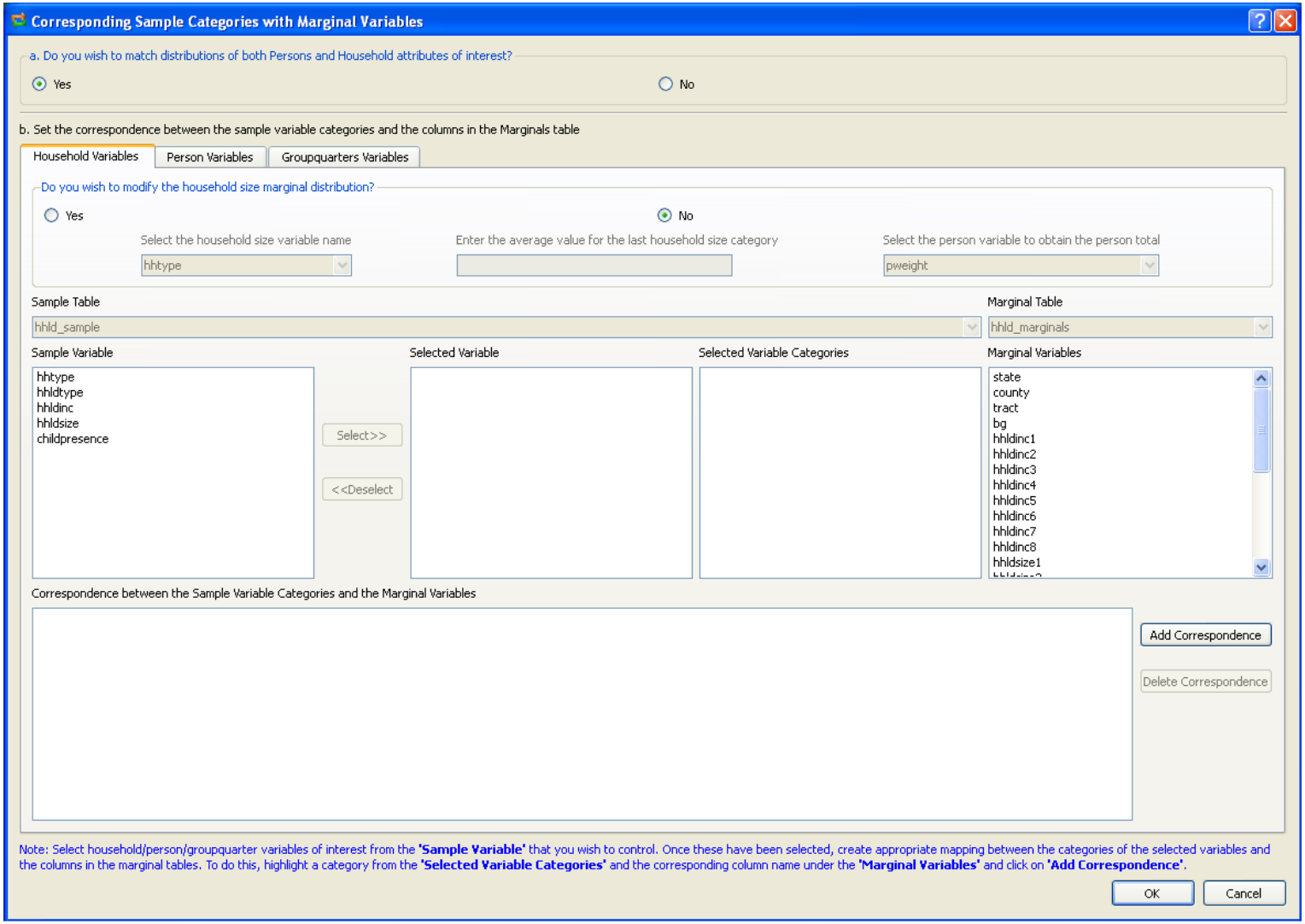


# Set Corresponding Variables

- Select **Yes** for "Do you wish to match distributions of both Persons and Household attributes of interest?"

  *Note*: If you select **No**, the synthesizer will not control for person attributes of interest
- Under the **Household Variables** tab, **No** will be selected by default
  - If you select **Yes**, the person totals will be matched at the expense of the household totals
  - For this exercise, leave the selection as **No**

# Set Corresponding Variables



# Set Corresponding Variables

- To create correspondences between the variables, select the appropriate variables from the **Sample Variable** column and click **Select**
- The variables now appear in the **Selected Variable** column and are ready to be assigned correspondences

# Set Corresponding Variables

- Next, select a variable in the **Selected Variable Categories** column along with its matching variable in the **Marginal Variables** column and click **Add Correspondence**
- Repeat this step for all the variables until they are all given correspondences
- The same procedure needs to be done for the **Person Variables** and **Groupquarters Variables** before continuing
- Select **OK** when all correspondences have been made

# Set Corresponding Variables

**Corresponding Sample Categories with Marginal Variables**

a. Do you wish to match distributions of both Persons and Household attributes of interest?

(•) Yes ( ) No

b. Set the correspondence between the sample variable categories and the columns in the Marginals table

Household Variables | Person Variables | Groupquarters Variables

Sample Table: person_sample

Marginal Table: person_marginals

| Sample Variable | Selected Variable | Selected Variable Categories | Marginal Variables |
|---|---|---|---|
| pweight | agep | agep, Category 10 | agep10 |
| relate | gender | gender, Category 1 | gender1 |
| | race | gender, Category 2 | gender2 |
| | employment | race, Category 1 | race1 |
| | | race, Category 2 | race2 |
| | | race, Category 3 | race3 |
| | | race, Category 4 | race4 |
| | | race, Category 5 | race5 |
| | | race, Category 6 | race6 |
| | | race, Category 7 | race7 |
| | | employment, Category 1 | employment1 |
| | | employment, Category 2 | employment2 |
| | | employment, Category 3 | employment3 |
| | | employment, Category 4 | employment4 |

Select>> <<Deselect

Correspondence between the Sample Variable Categories and the Marginal Variables

agep, Category 1 - agep1
agep, Category 2 - agep2
agep, Category 3 - agep3
agep, Category 4 - agep4
agep, Category 5 - agep5
agep, Category 6 - agep6
agep, Category 7 - agep7
agep, Category 8 - agep8
agep, Category 9 - agep9
agep, Category 10 - agep10
gender, Category 1 - gender1
gender, Category 2 - gender2
race, Category 1 - race1
race, Category 2 - race2
race, Category 3 - race3
race, Category 4 - race4
race, Category 5 - race5
race, Category 6 - race6
race, Category 7 - race7

Add Correspondence

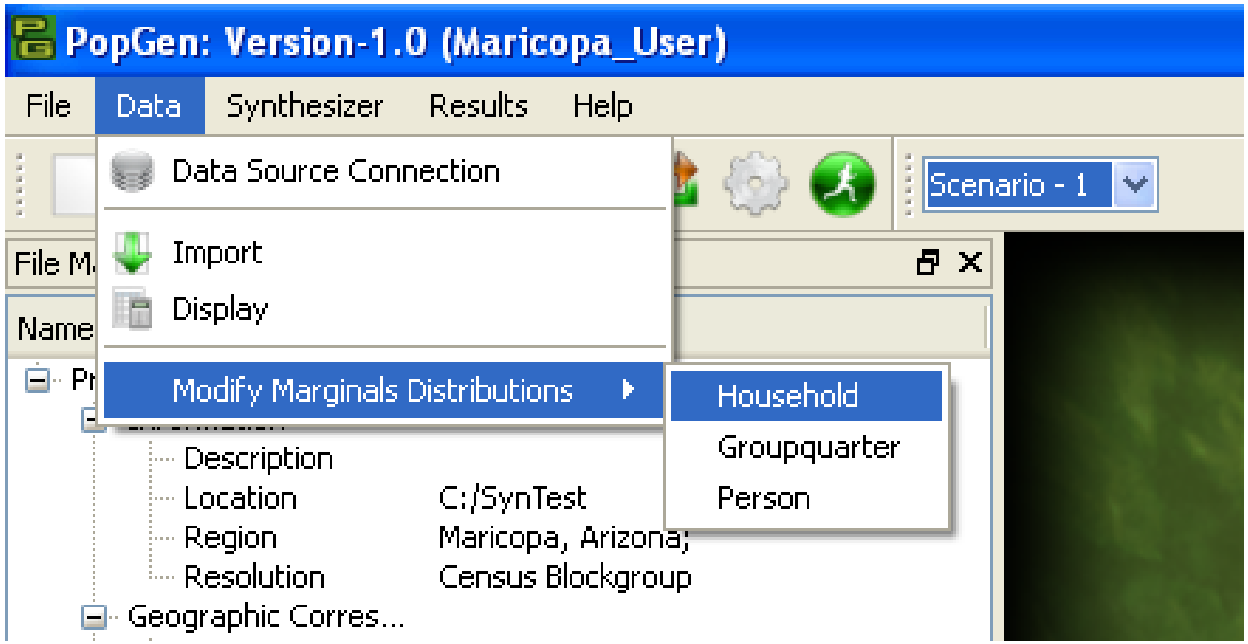
Delete Correspondence

Note: Select household/person/groupquarter variables of interest from the **'Sample Variable'** that you wish to control. Once these have been selected, create appropriate mapping between the categories of the selected variables and the columns in the marginal tables. To do this, highlight a category from the **'Selected Variable Categories'** and the corresponding column name under the **'Marginal Variables'** and click on **'Add Correspondence'**.

OK Cancel

# Modify Marginal Distributions

- Go to **Data** ➡ **Modify Marginals Distributions** ➡ **Household** in the menu bar at the top

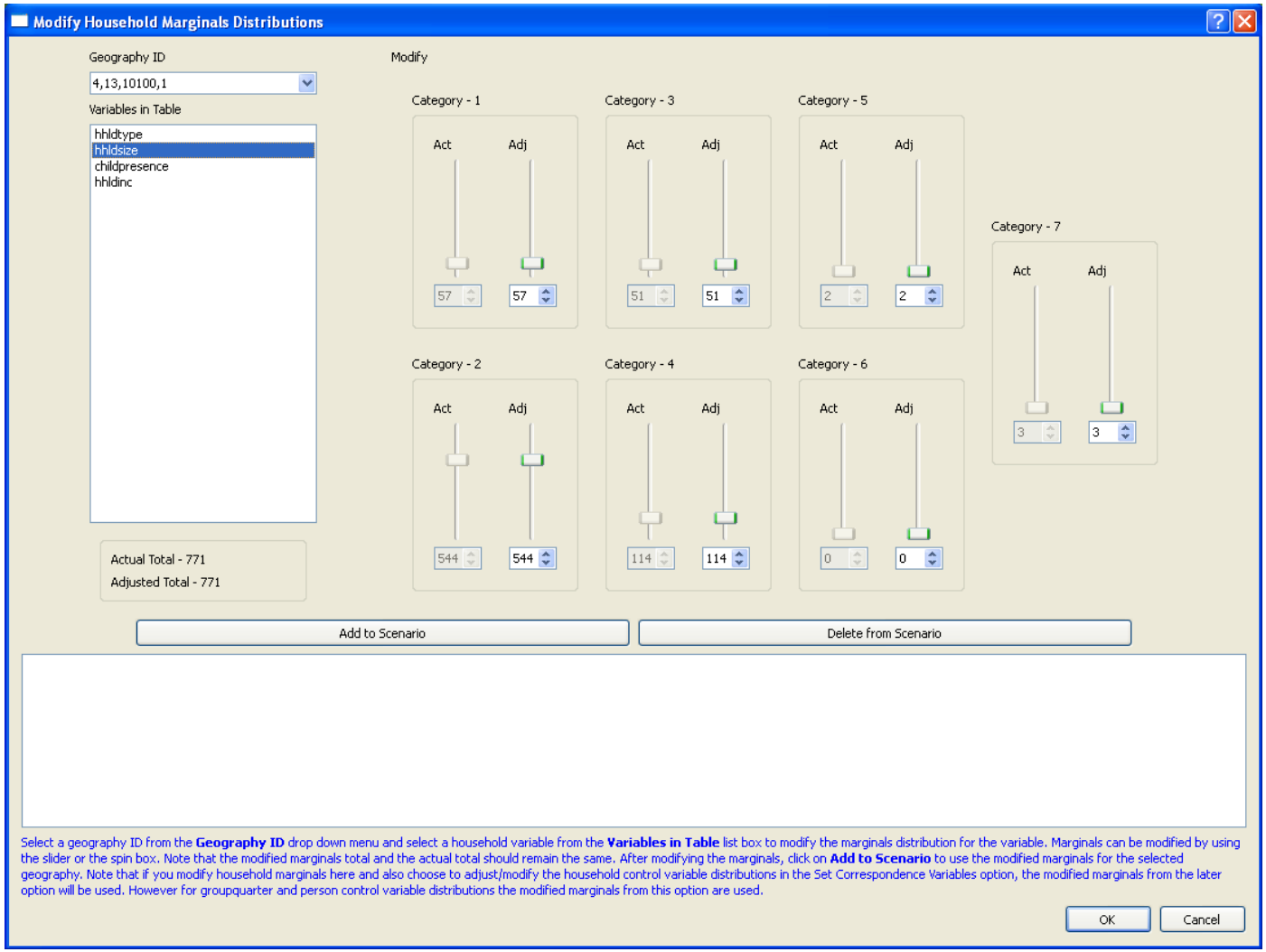


# Modify Marginal Distributions

- Select a geography ID from the drop down menu and then select a household variable from the **Variables in Table** list box to modify the marginals distribution for the variable
- Modify the marginals by moving the slider or adjusting the spin box

*Note*: The modified marginals total and the actual total need to remain the same

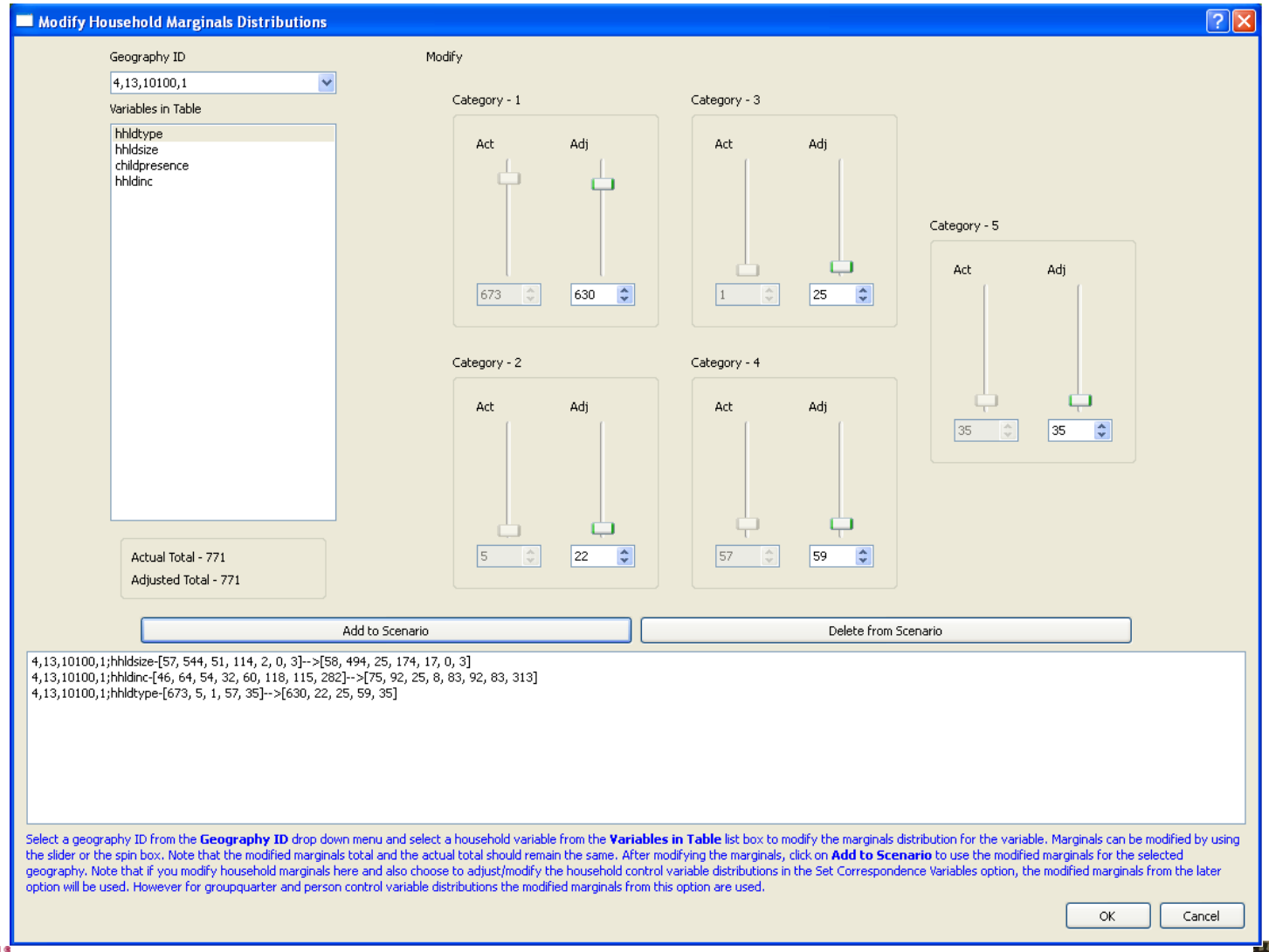
# Modify Marginal Distributions



# Modify Marginal Distributions

- After modifying the variables, click **Add Scenario** to use the modified marginals for the selected geography
- If you modify household marginals here and also choose to adjust/modify the household control variable distributions in the Set Correspondence Variables option, the modified marginals from the later option will be used
- However, for groupquarter and person control variable distributions, the modified marginals from this option are used

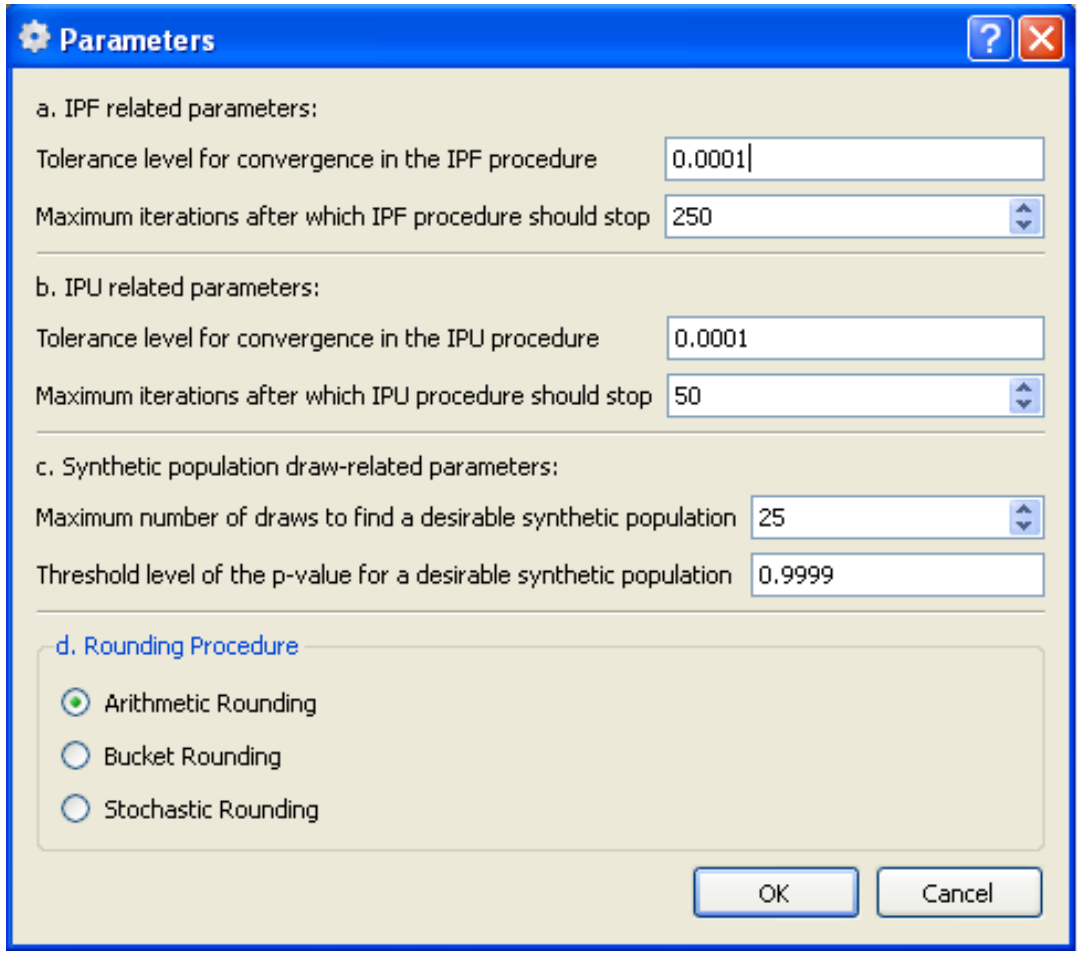
# Modify Marginal Distributions



# Parameters

- Go to **Synthesizer** ➡ **Parameters/Settings** on the menu bar at the top or click the **Parameters** icon in the toolbar at the top of the page
- There are 4 options that you can adjust in the Parameters window:
  1. IPF related parameters
  2. IPU related parameters
  3. Synthetic population draw-related parameters
  4. Rounding procedure

# Parameters


Parameters
a. IPF related parameters:
Tolerance level for convergence in the IPF procedure
0.0001
Maximum iterations after which IPF procedure should stop
250
b. IPU related parameters:
Tolerance level for convergence in the IPU procedure
0.0001
Maximum iterations after which IPU procedure should stop
50
c. Synthetic population draw-related parameters:
Maximum number of draws to find a desirable synthetic population
25
Threshold level of the p-value for a desirable synthetic population
0.9999
d. Rounding Procedure
Arithmetic Rounding
Bucket Rounding
Stochastic Rounding
OK
Cancel

# Session 8 – Generating a Synthetic Population

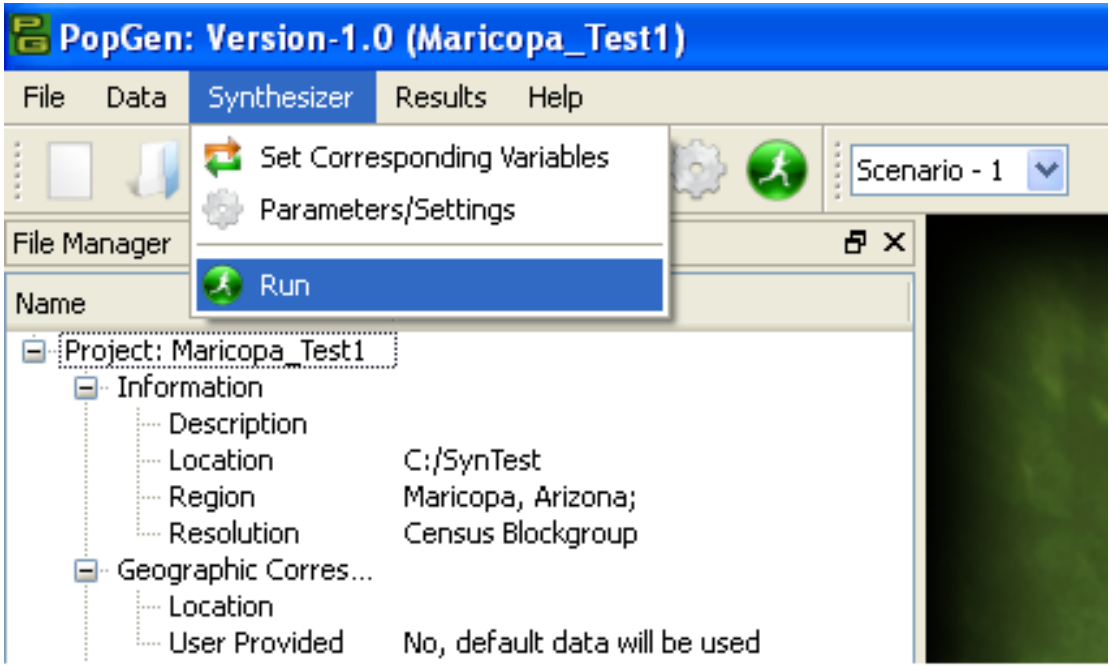
PopGen Training Workshop

May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Running the Synthesizer

- To begin the population synthesis, go to **Synthesizer** ➡ **Run** in the menu bar at the top or click the **Run** icon in the toolbar

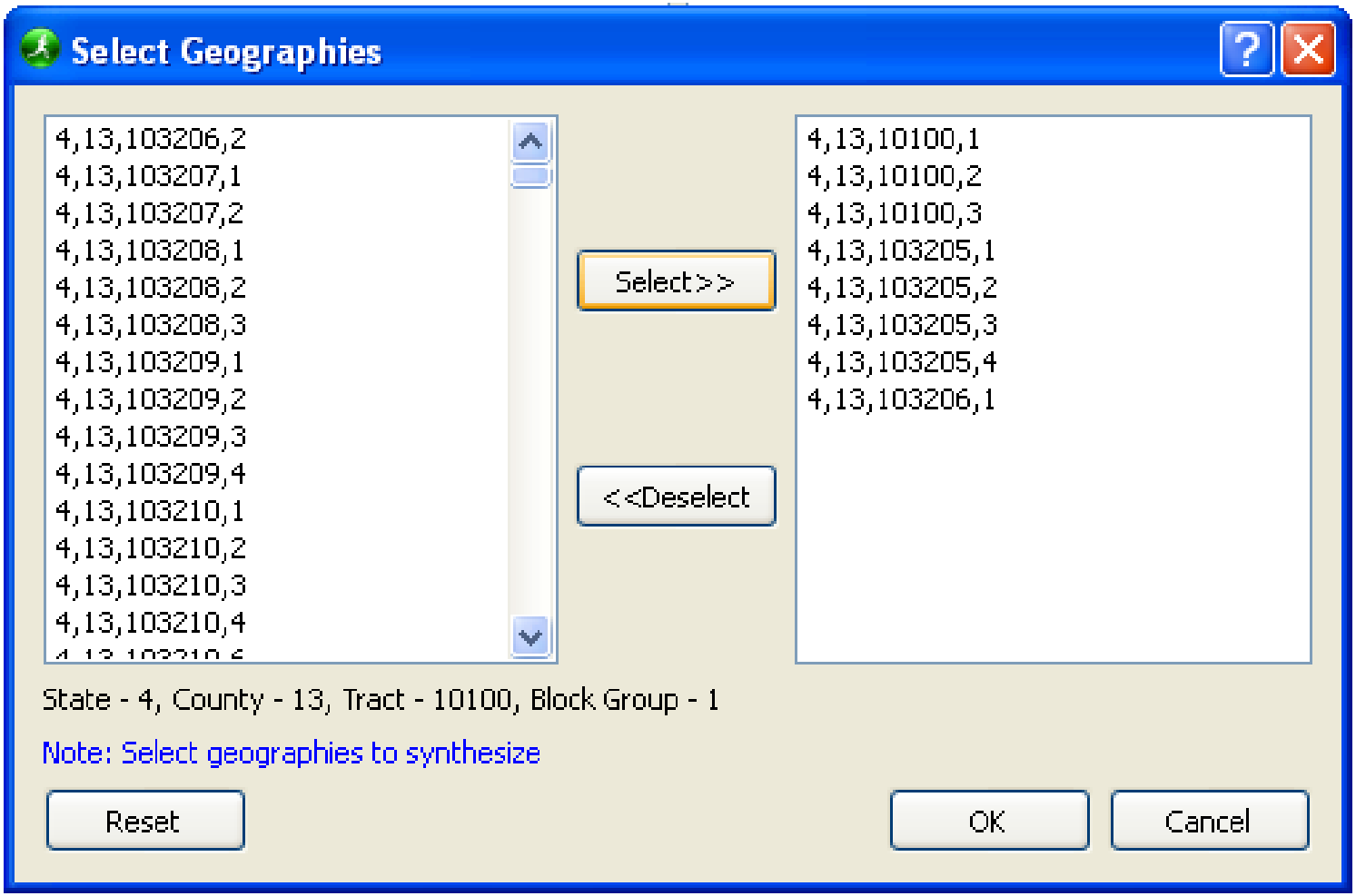


# Running the Synthesizer

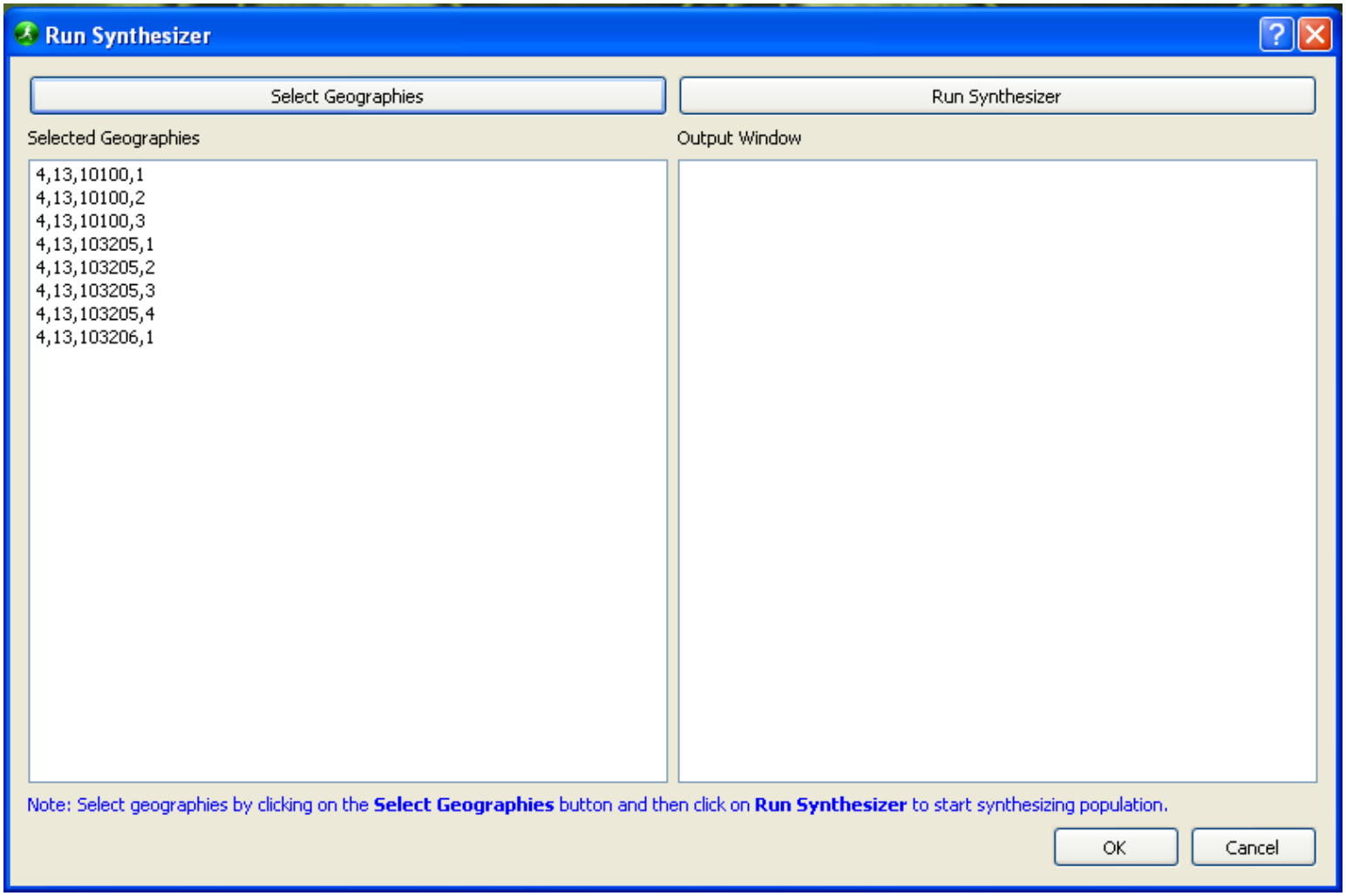
- Once the synthesizer window opens, click **Select Geographies**
- A new window opens with a list of all the census blockgroups
- For this exercise, choose the first 8 blockgroups from the left column and click **Select**
- Click **OK** to continue

*Note*: You can choose as little or as many geographies as you like. The more geographies you choose, the longer the synthesis will take to process.

# Running the Synthesizer



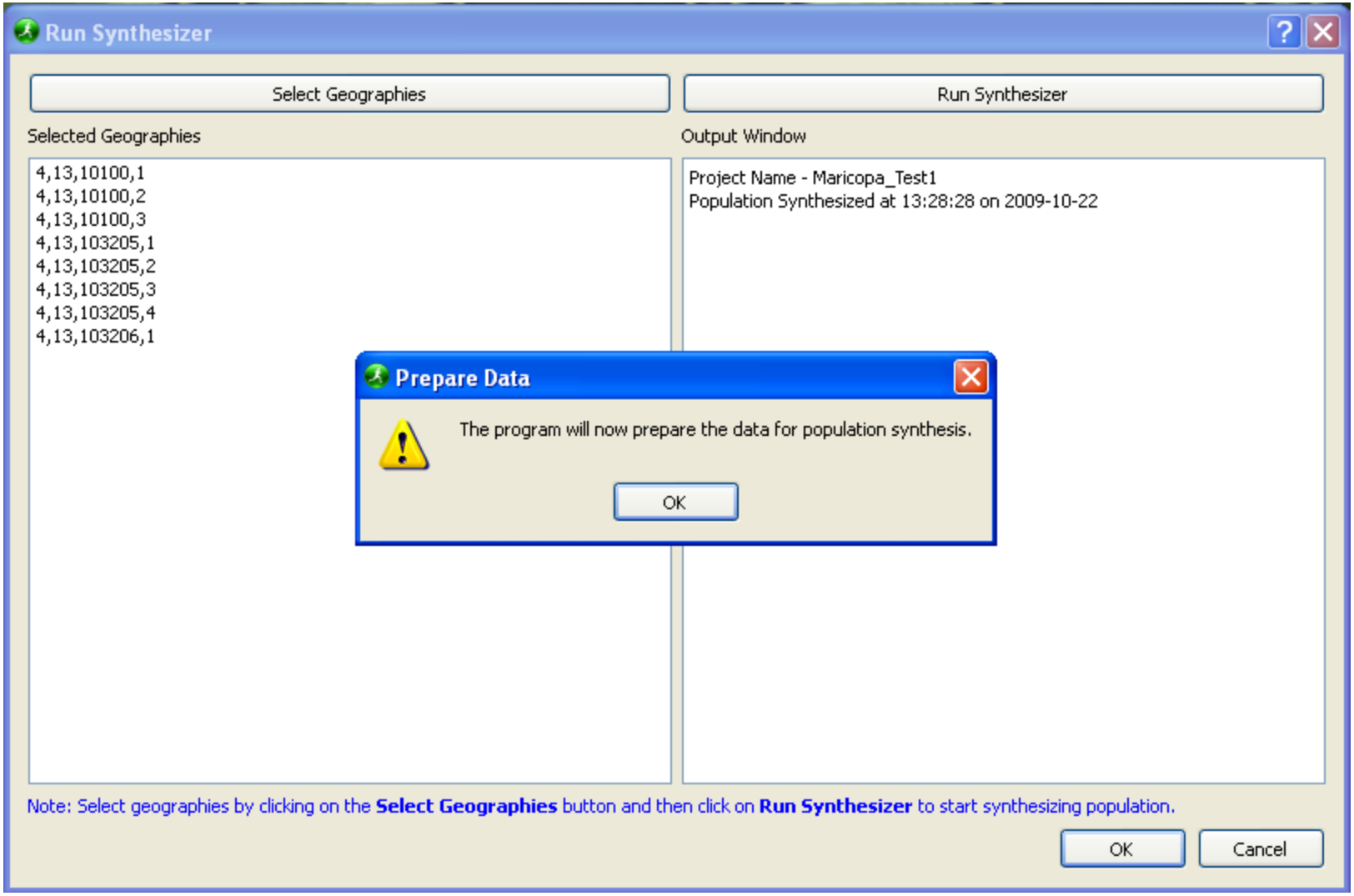
# Running the Synthesizer



# Running the Synthesizer

- With the geographies now selected, we can run the synthesizer
- Click the **Run Synthesizer** button on the top right of the window to start synthesizing
- A pop-up will appear telling you that “The program will now prepare the data for population synthesis”
- Click **OK** to continue

# Running the Synthesizer

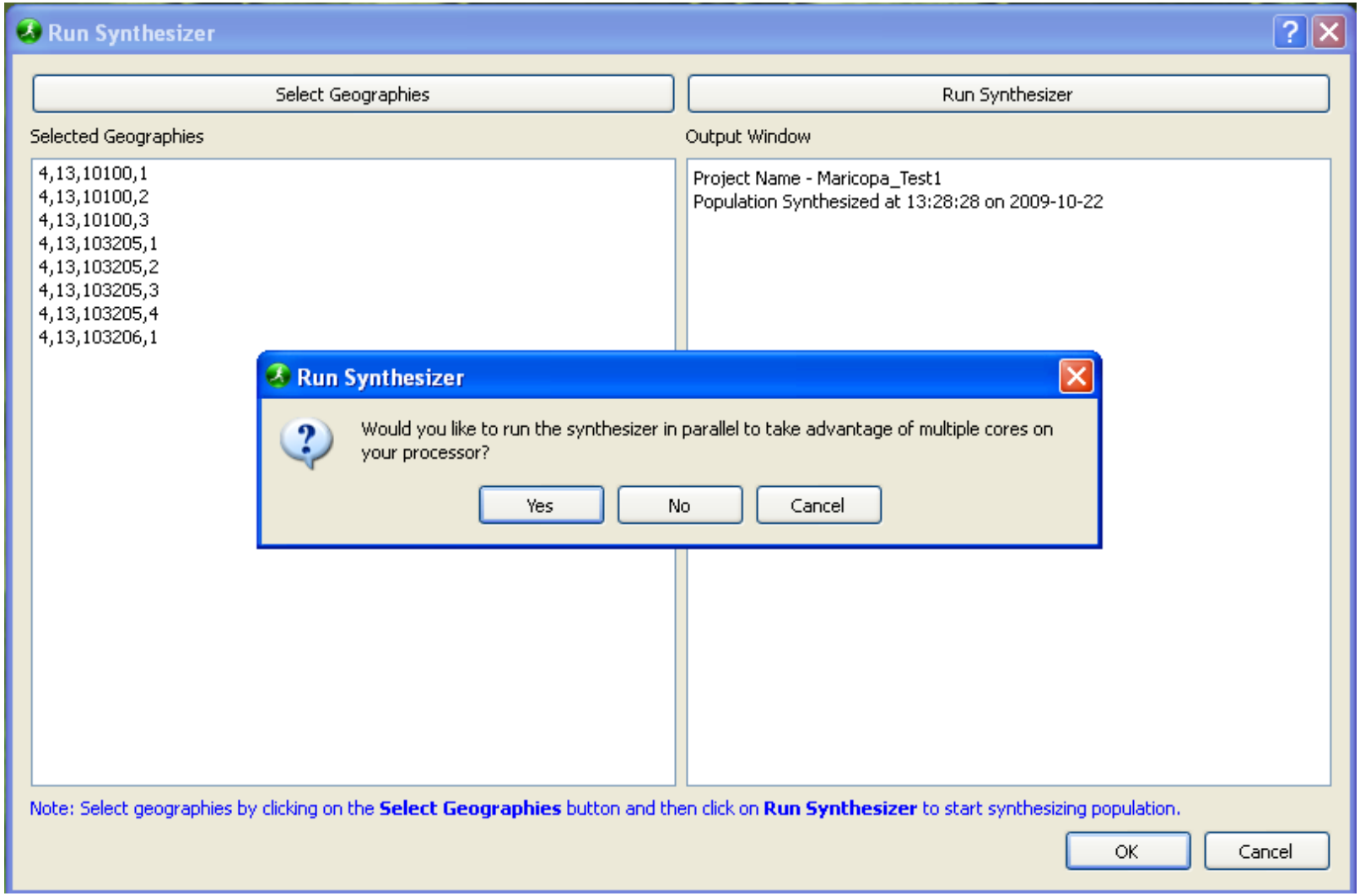


# Running the Synthesizer

- It will take a minute or so to prepare the data
- If you are using a computer with a multiple core processor, a pop-up will appear prompting you if you would like to run the synthesizer in parallel to take advantage of multiple cores on your processor
- Click **Yes** to continue

*Note*: If you were to choose No, the synthesizer would run the process in serial

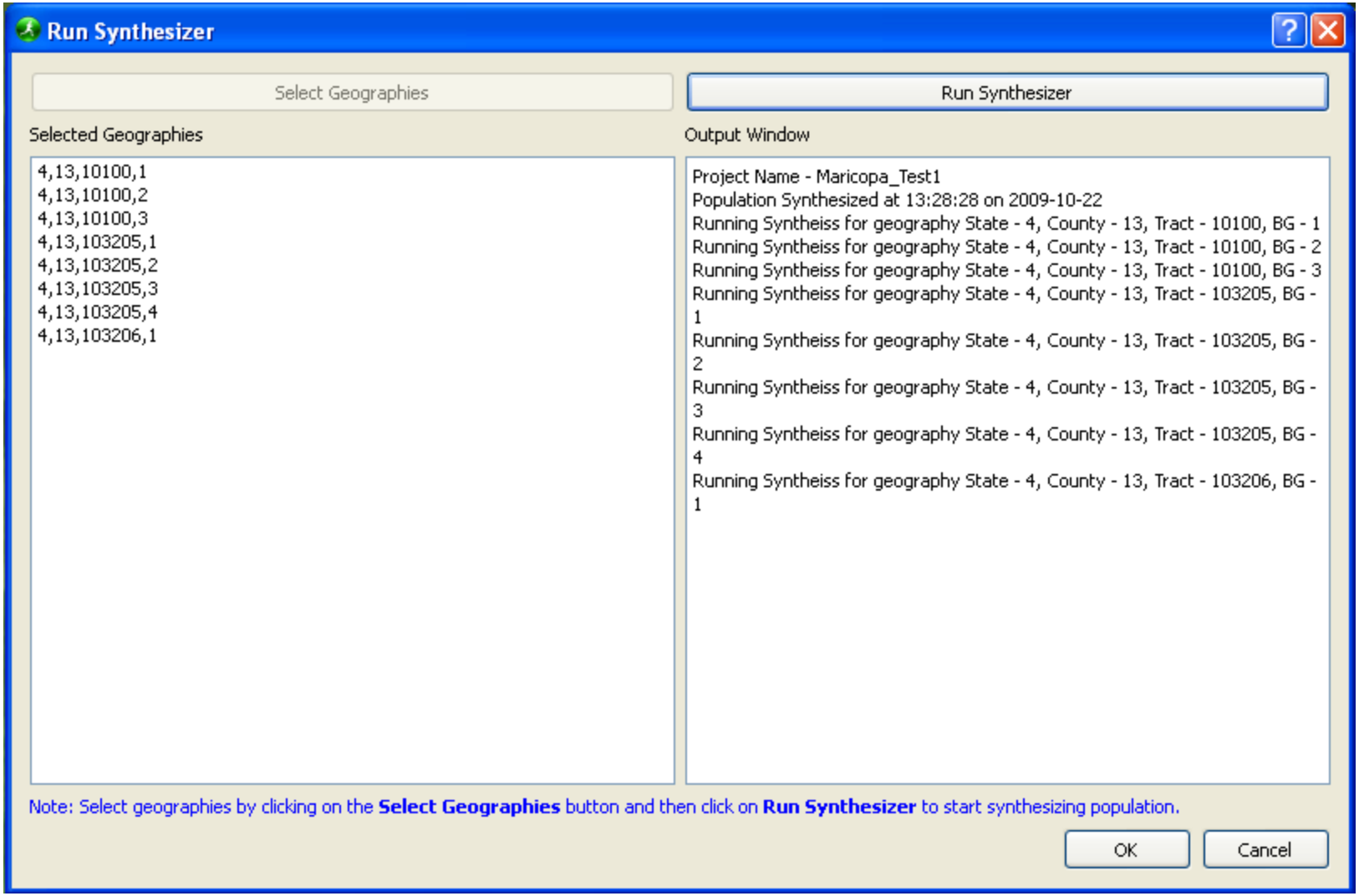
# Running the Synthesizer



# Running the Synthesizer

- The synthesizer will take several minutes to process the results
- Once it has completed, the **Output Window** on the right hand side of the **Run Synthesizer Window** will display each geography that was processed
- Click **OK** to close the window
- Next, we will look at some visualizations of the results

# Running the Synthesizer



# Session 9 – Visualization and Exporting Results

PopGen Training Workshop

May 12 and 13, 2010

ARIZONA STATE UNIVERSITY

# Visualization and Exporting Results

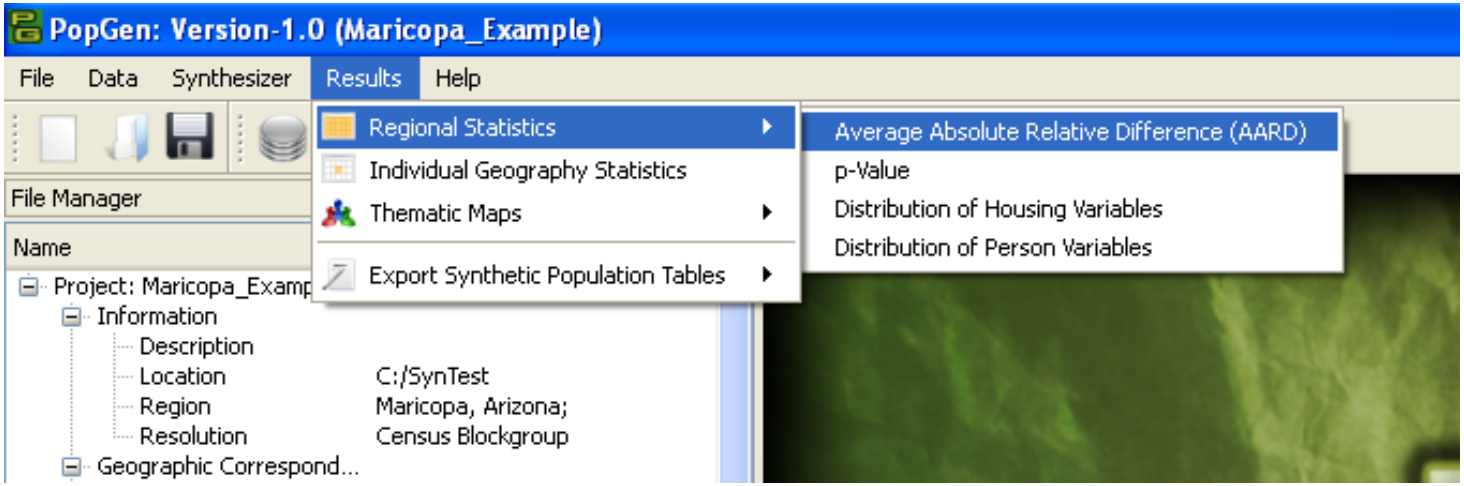
- This session will give an overview of the features available to visualize the synthetic population results
- We will also look at the options available for exporting results onto your system

# Visualization

- Visualizing Results
  - Regional
    - Performance Measures
    - Thematic Maps
  - Individual Geography
    - Performance Measures

# Average Absolute Relative Difference

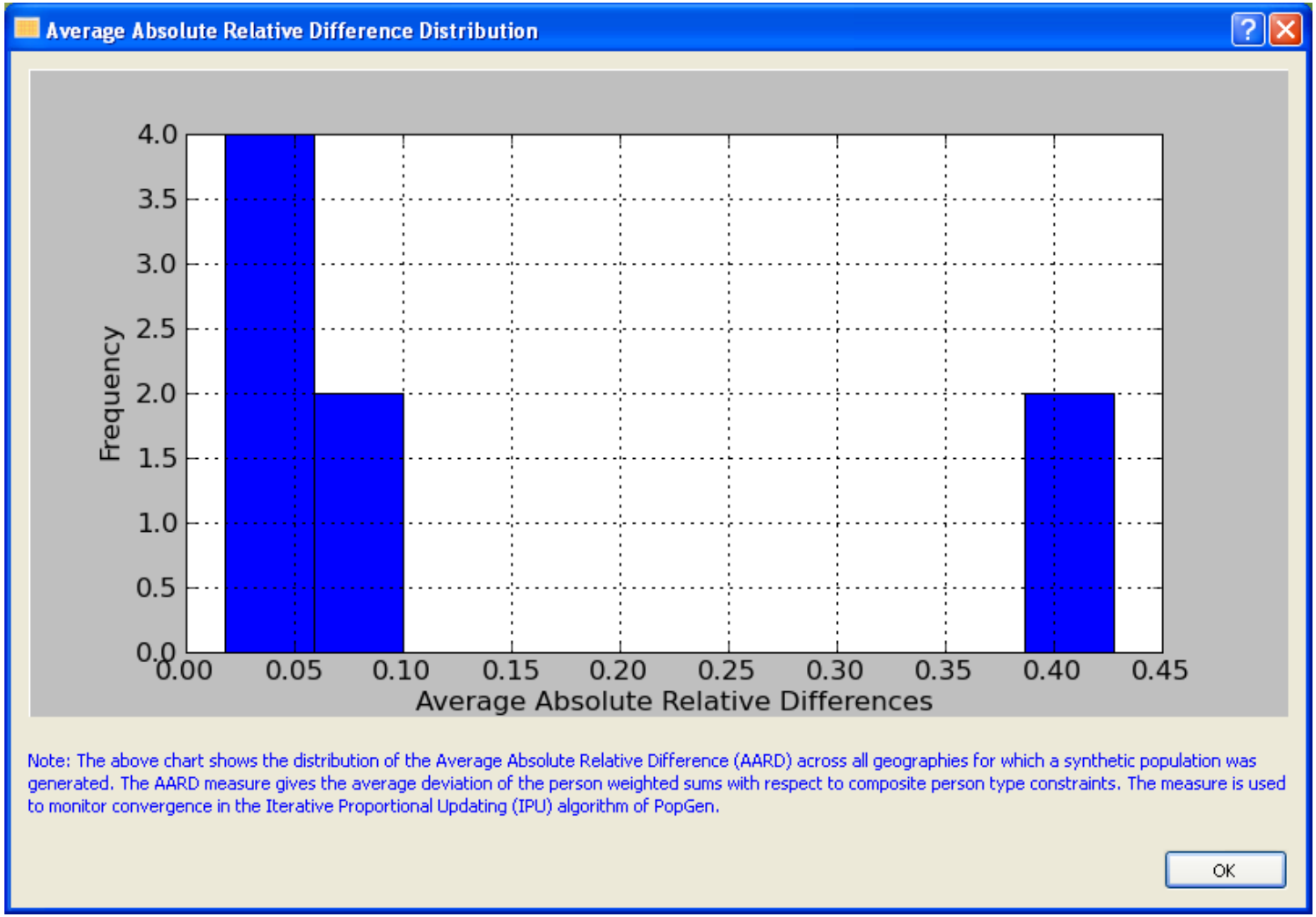
- Go to **Results** ➡ **Regional Statistics** ➡ **Average Absolute Relative Difference (AARD)** on the menu bar at the top



# Average Absolute Relative Difference

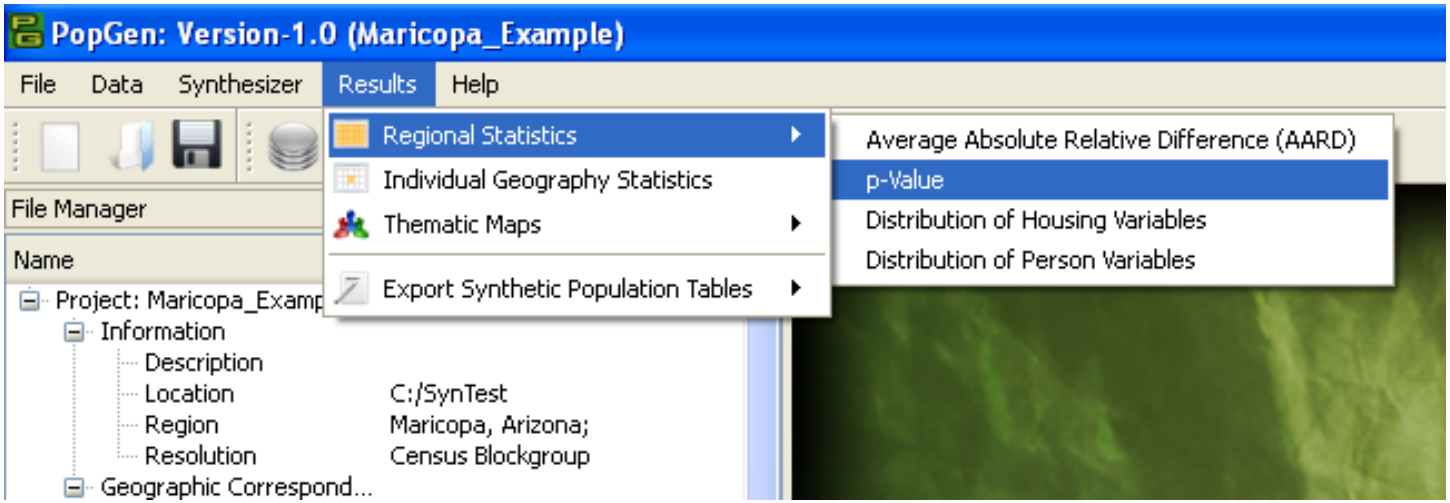
- The Average Absolute Relative Difference (AARD) chart displays the distribution of the AARD across all geographies for which a synthetic population was generated
- The AARD measure gives the average deviation of the person weighted sums with respect to composite person type constraints
- The AARD is also used to monitor convergence in the Iterative Proportional Updating (IPU) algorithm of PopGen

# Average Absolute Relative Difference



# p-Value

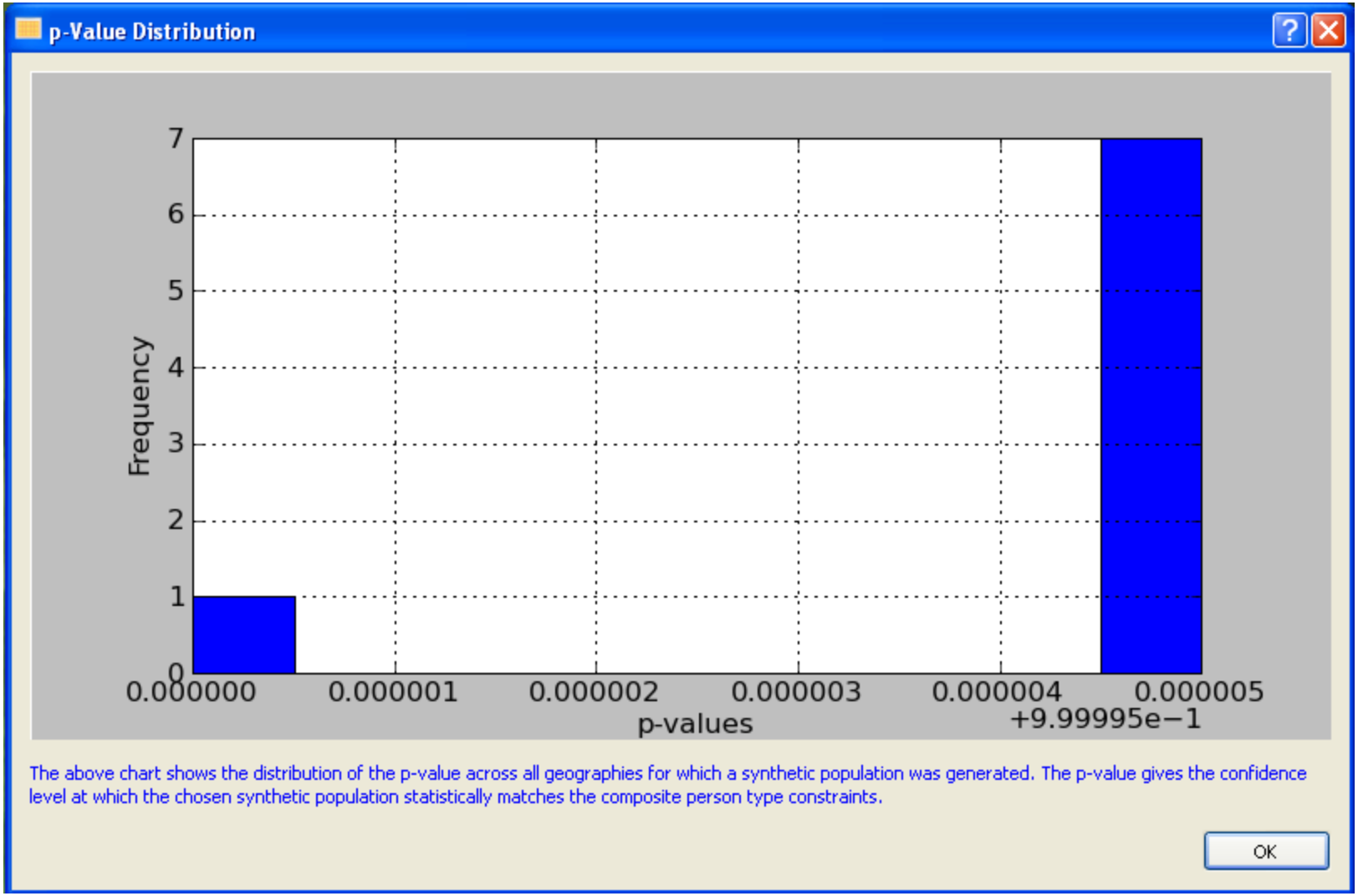
- Go to **Results** ➡ **Regional Statistics** ➡ **p-Value** in the menu bar at the top



# p-Value

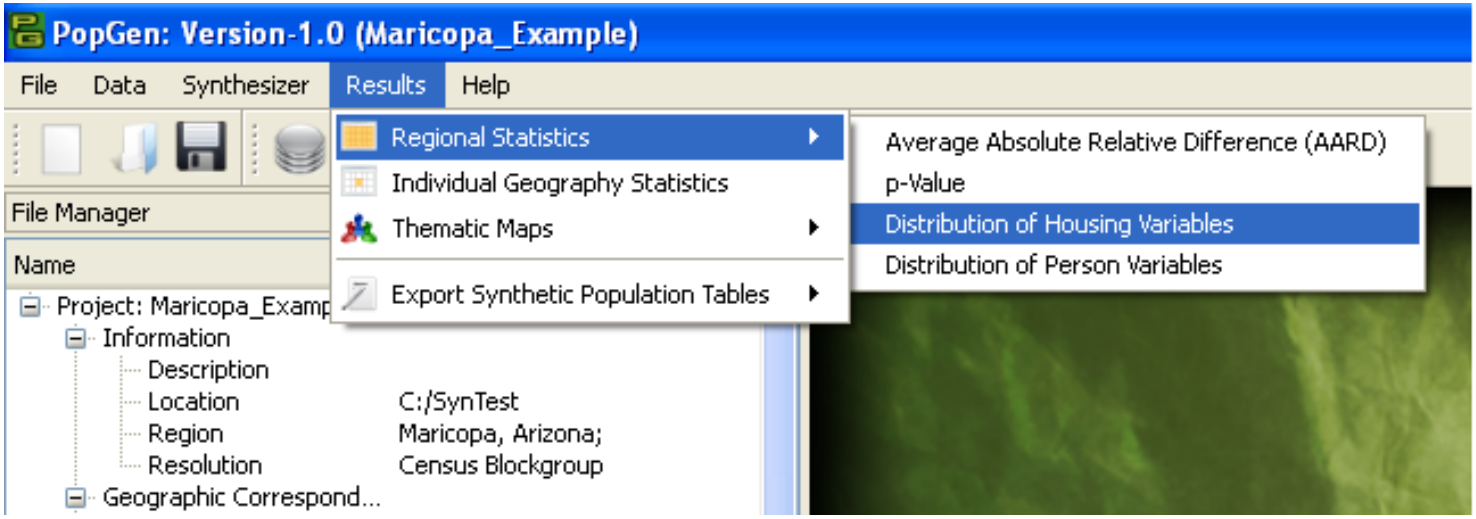
- The chart shows the distribution of the p-value across all geographies for which a synthetic population was generated
- The p-value gives the confidence level at which the chosen synthetic population statistically matches the composite person type constraints

# p-Value


p-Value Distribution
Frequency
0
1
2
3
4
5
6
7
0.000000
0.000001
0.000002
0.000003
0.000004
0.000005
+9.99995e−1
p-values
The above chart shows the distribution of the p-value across all geographies for which a synthetic population was generated. The p-value gives the confidence level at which the chosen synthetic population statistically matches the composite person type constraints.
OK

# Distribution of Housing Variables

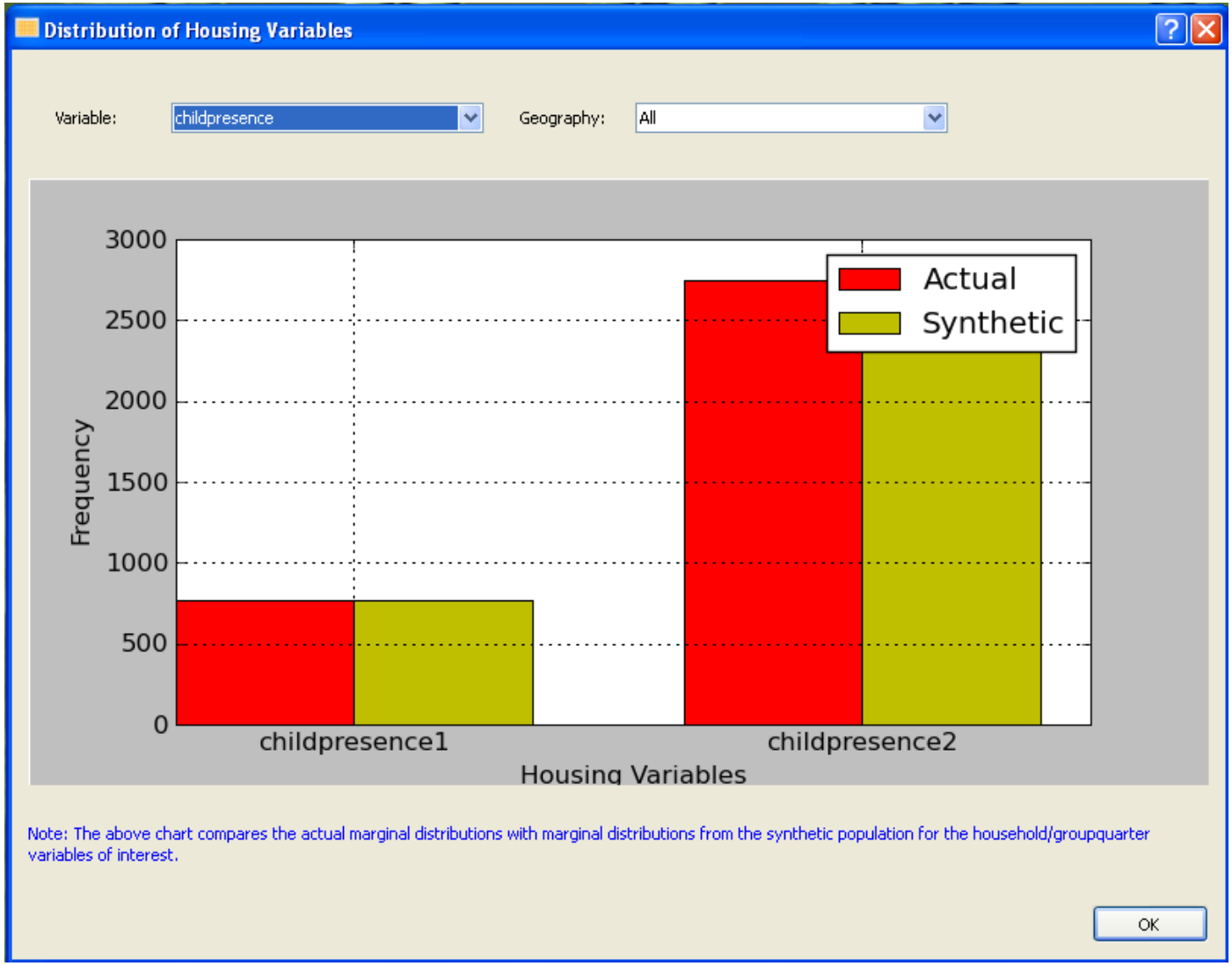
- Go to **Results** ➡ **Regional Statistics** ➡ **Distribution of Housing Variables** in the menu bar at the top



# Distribution of Housing Variables

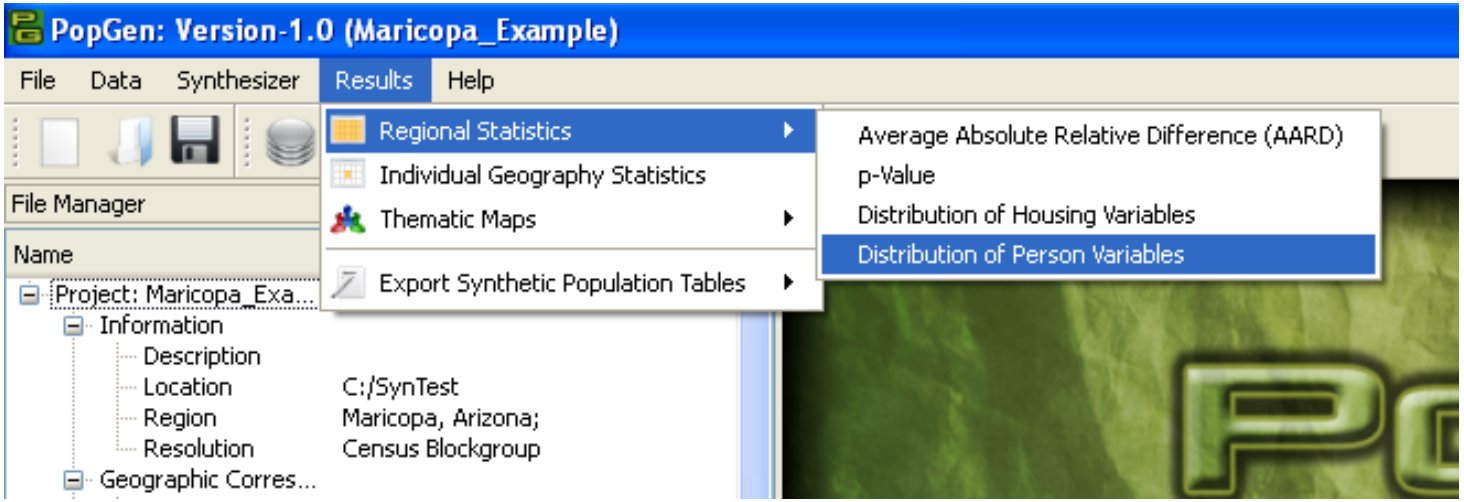
- The distribution of housing variables chart compares the actual marginal distributions with marginal distributions from the synthetic population for the household/groupquarter variables of interest

# Distribution of Housing Variables



# Distribution of Person Variables

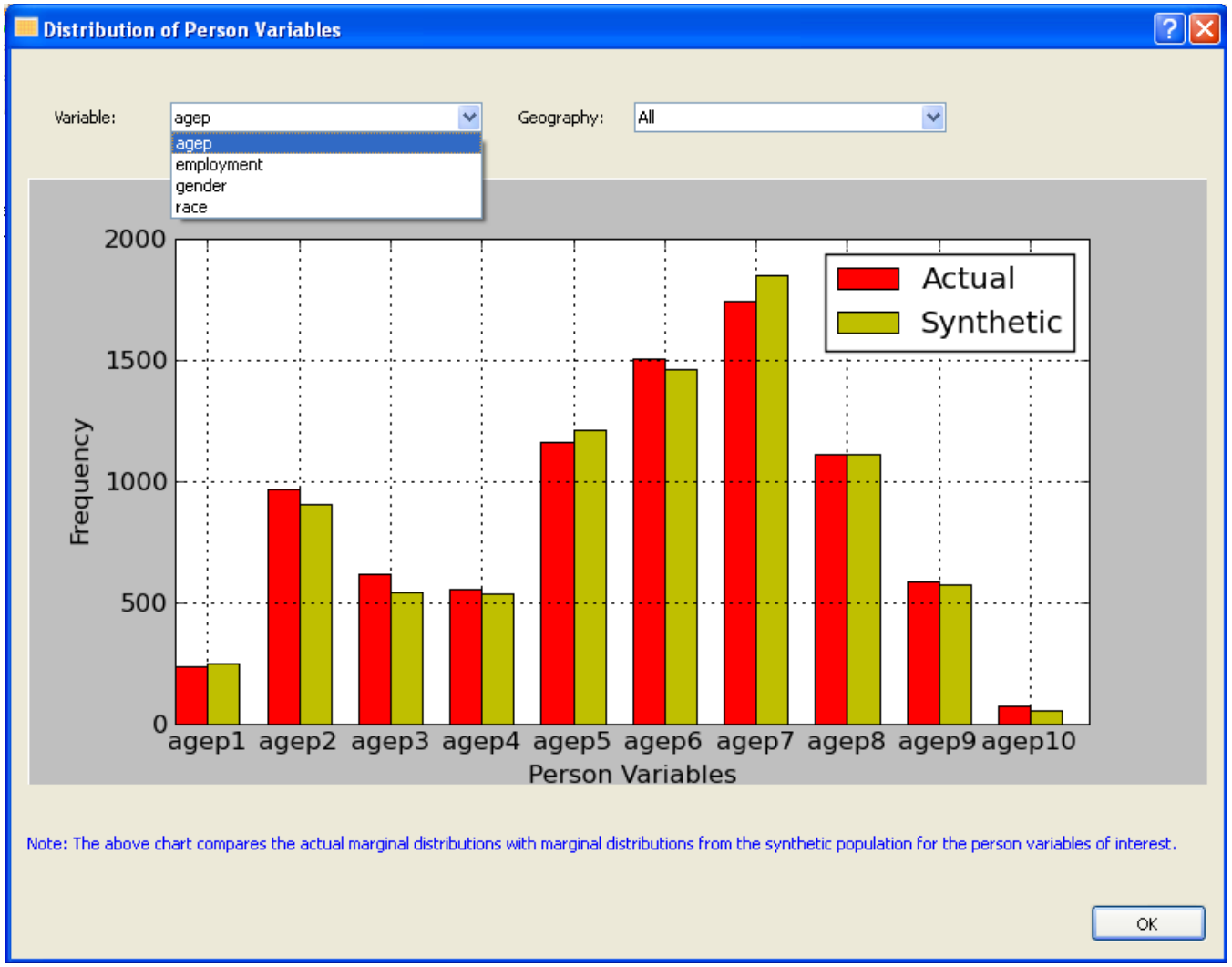
- Go to **Results** ➡ **Regional Statistics** ➡ **Distribution of Person Variables** in the menu bar at the top



# Distribution of Person Variables

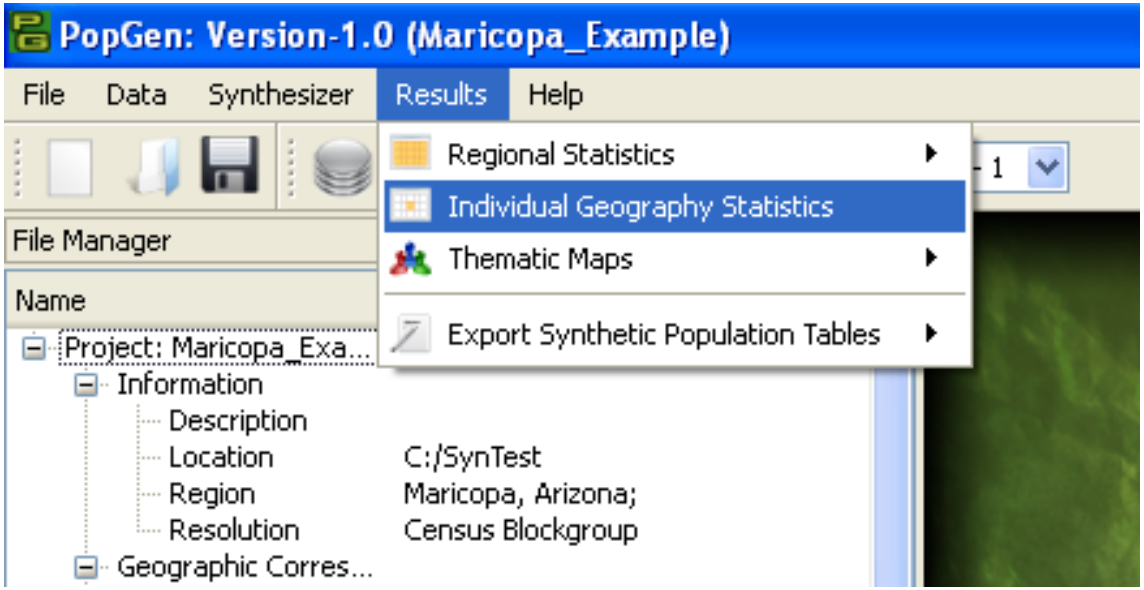
- The distribution of person variables chart compares the actual marginal distributions with marginal distributions from the synthetic population for the person variables of interest

# Distribution of Person Variables



# Individual Geography Statistics

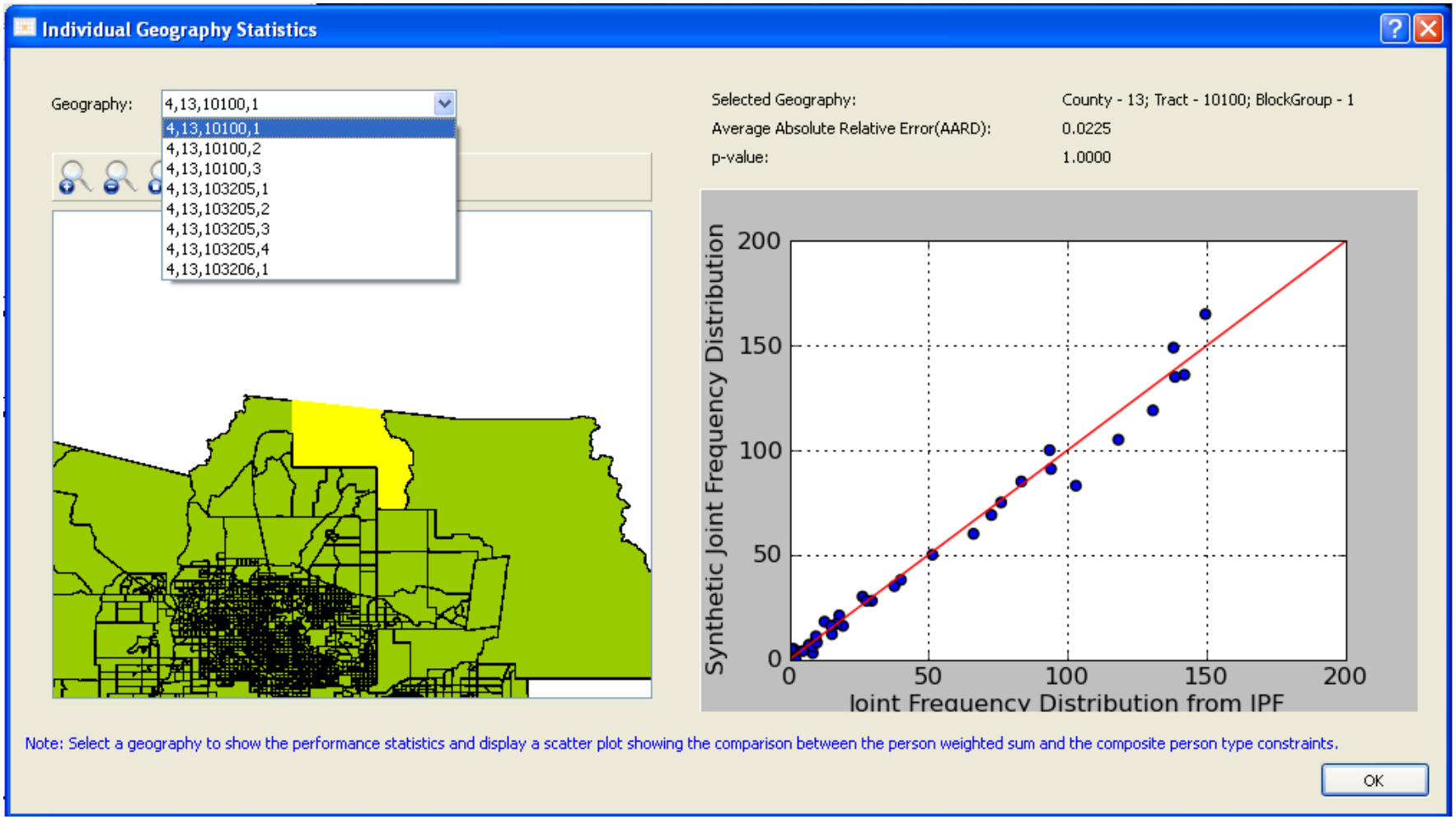
- Go to **Results** ➡ **Individual Geography Statistics** in the menu bar at the top



# Individual Geography Statistics

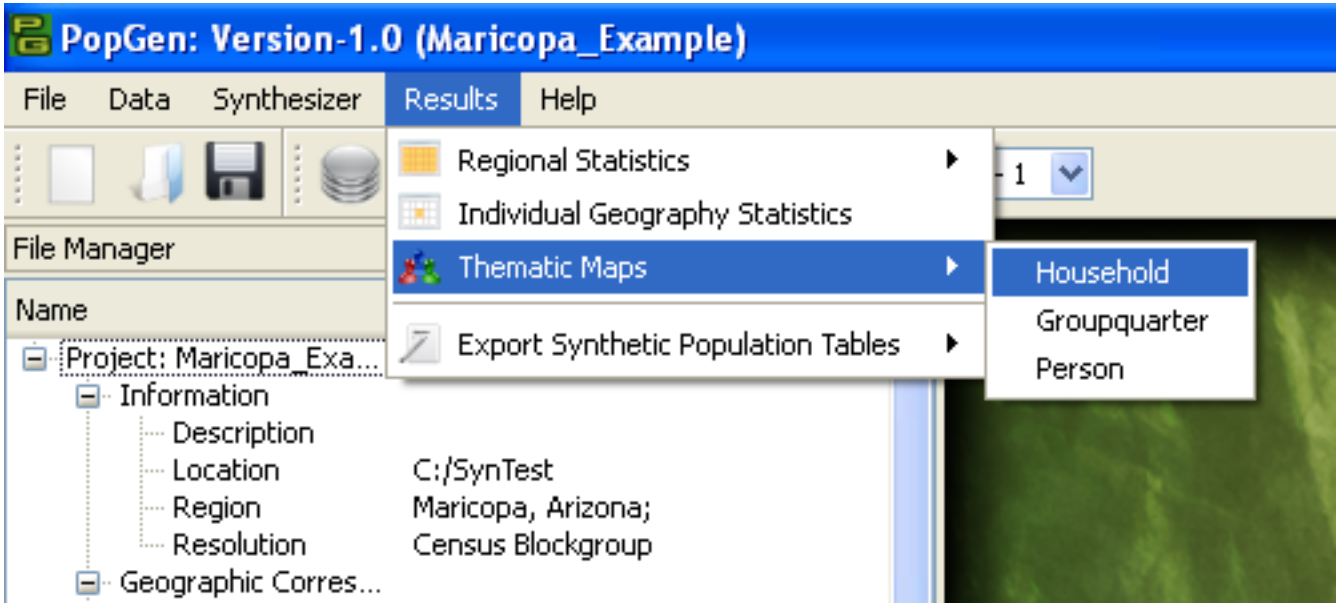
- Select one of the geographies in the drop down menu to show the performance statistics
- A scatter plot is displayed showing the comparison between the person weighted sum and the composite person type constraints

# Individual Geography Statistics



# Thematic Maps – Household

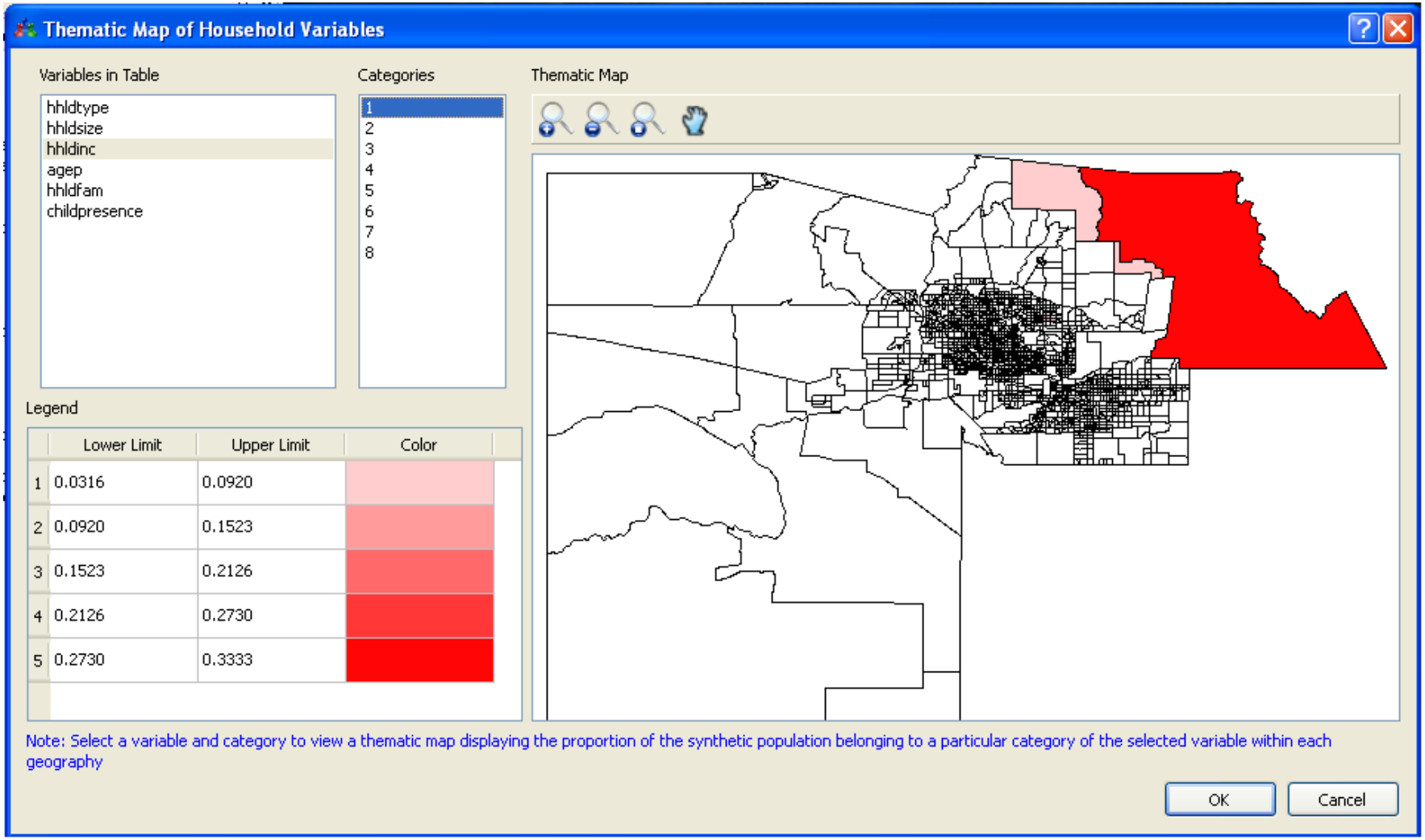
- Go to **Results** ➡ **Thematic Maps** ➡ **Household** in the menu bar at the top



# Thematic Maps – Household

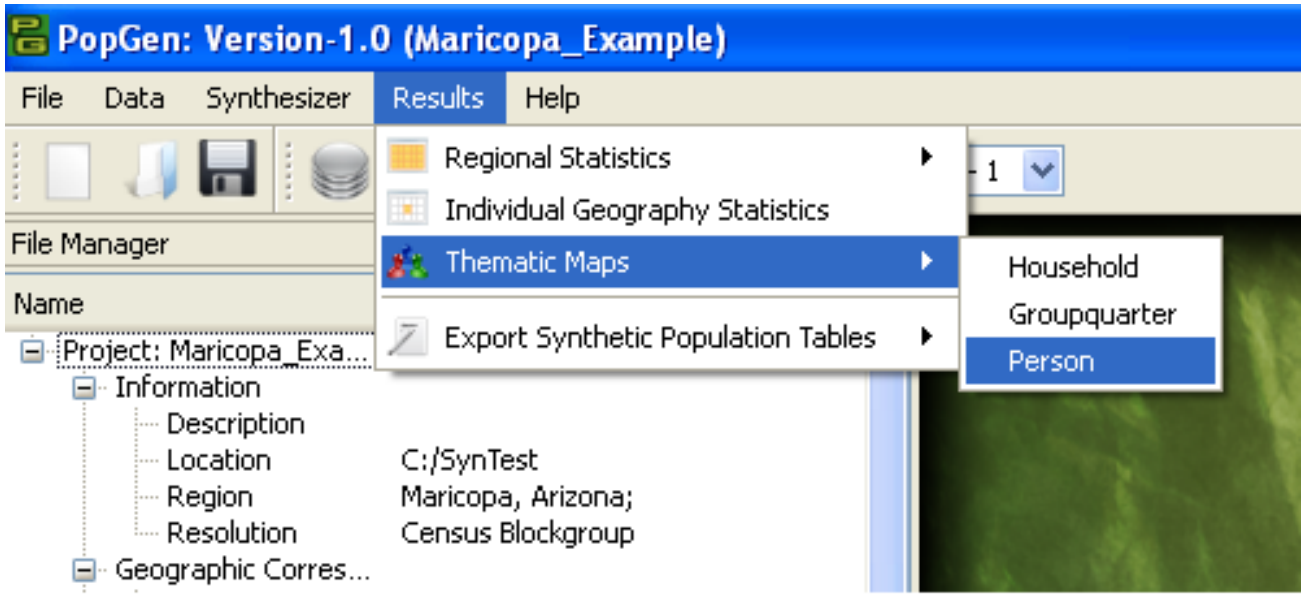
- Select a household variable and then choose a category from that variable
- The thematic map displays the proportion of the synthetic population belonging to the selected category within each geography
- A legend on the lower left shows the upper and lower limits of the selected category

# Thematic Maps – Household



# Thematic Maps – Person

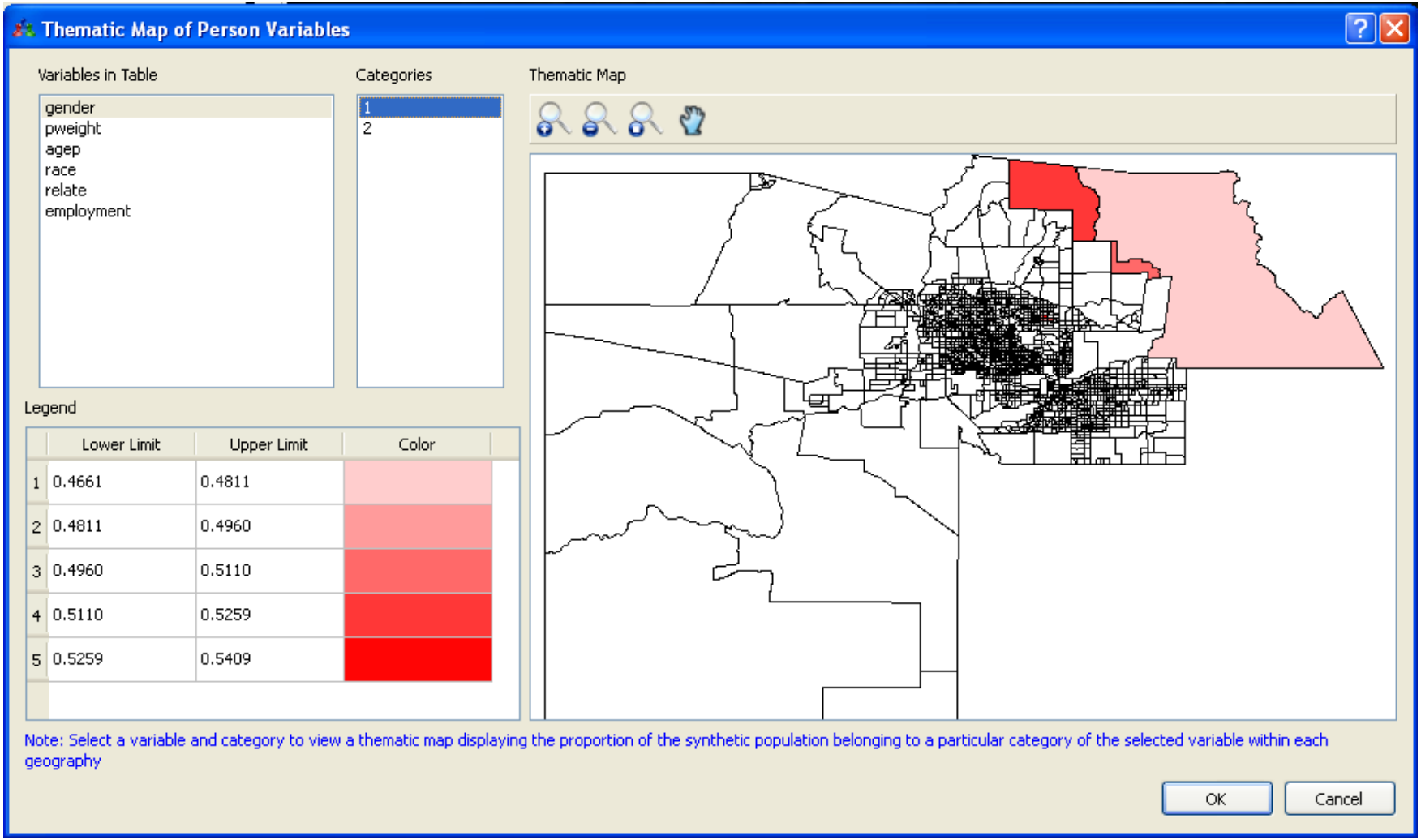
- Go to **Results** ➡ **Thematic Maps** ➡ **Person** in the menu bar at the top



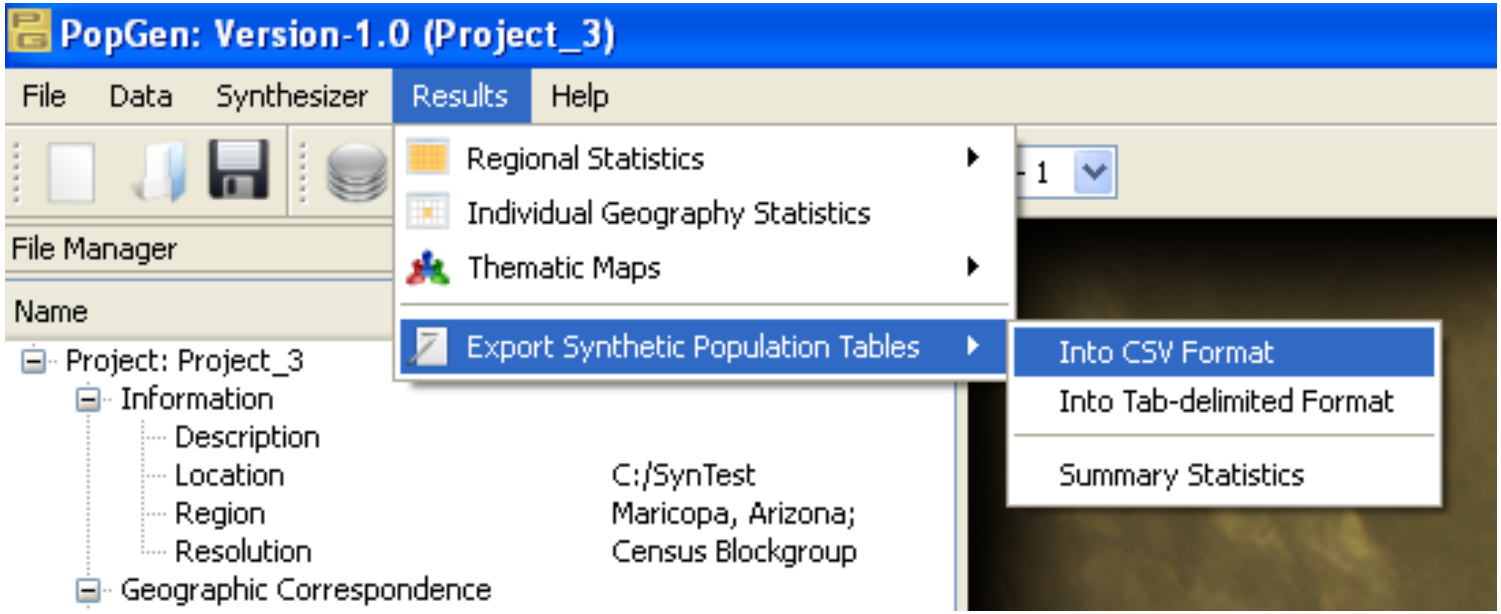
# Thematic Maps – Person

- Select a person variable and then choose a category within that variable
- The thematic map displays the proportion of the synthetic population belonging to the selected category within each geography
- A legend on the lower left shows the upper and lower limits of the selected category

# Thematic Maps – Person



# Exporting Results

- Go to **Results** ➡ **Export Synthetic Population Tables** in the menu bar at the top
- Select **Into CSV Format** or **Into Tab-delimited Format**
- Browse for a folder where you want to store the table on your computer and click **OK**
- You can also choose to export a summary of the result statistics by selecting **Summary Statistics**

# Exporting Results