SMALL AREA POPULATION ESTIMATES PROJECT SUMMARY REPORT

Qinglin Hu, Xiaojiang Li, Weixing Zhang, and Michael R. Howser

Special thanks to Karyn Backus and Lloyd Mueller for their assistance and guidance with this project and report.

1. Problem statement

Population data is a key component for health statistics, governmental planning, and resource allocations at national, state, and local levels. The US Census Bureau, the federal agency tasked with collecting and disseminating population data for the Nation, provides county-level population estimates annually that include demographic identifiers, such as age, sex, race, and Hispanic origin. While the majority of the states in the U.S. use county as the principal geographic level for local governance, Connecticut and a few others states rely on towns or cities. Currently, the only reliable source for town-level population data with demographic identifiers for Connecticut is the decennial census that occurs every 10 years. For the years between decennial censuses, only town population totals are published by the Census Bureau. As the demographic distributions within each town evolve over time, the decennial counts become outdated and may insufficiently represent the true town populations. Connecticut must wait 10 years for an updated population distribution from the next decennial census. An alternative option is to develop an in-house process to estimate the demographic distribution of each town on an annual basis. With no comprehensive resource for population counts in Connecticut, this task requires identifying and accessing reliable resources and developing a model that accurately estimates population distributions using available resources.

2. Highlights

Identified viable population data sources for various age groups, connected with data providers to receive datasets, and pre-processed inputs for modeling.

Developed a workflow to generate population estimates data yearly by combining multi-sources datasets.

Developed and evaluated three possible models for conversion of ZIP code data to town data

Tested the predictive value of data sources against decennial population data

Developed regression models for 2010 predictors and 2010 population town-age-sex-race/ethnicity subgroups based on select combinations of data sources and age groups.

- Birth and infant death data were used to estimate the 0-4 age group.
- School enrollment counts were used to estimate the 5-9 and 10-14 age groups.
- For the 15-19 age group, a model utilizing both school enrollment and DMV licenses/non-driver IDs was developed.
- DMV licenses/non-driver IDs and residential utility customer data were used to develop a model for the 5-year age groups between 20 and 64 years.
- Medicare data was used to estimate the 5-year age groups for 65 and older.
- These regression model equations were used with updated 2011-2014 predictor data to estimate the population counts in 2011-2014 by town-age-sex-race/ethnicity.

3. Introduction

The founding task of the U.S. Census Bureau is to conduct a count of the population of the United States every 10 years. Known as the decennial census, the purpose of this census is to provide a true count of the population, rather than an estimate, that will be used to allocate representation in the Congress among the states (www.census.gov). Over time, the decennial census has expanded beyond a headcount by collecting a variety of demographic indicators, such as age, sex, and ancestry, and social and economic indicators, such as living arrangements and income strata. Due to its comprehensive assessment, the decennial census is the single most important demographic data collection effort the federal government implements (Hare, 2013).

In 2010, the American Community Survey (ACS) replaced the decennial census as the sole national source of select demographic and economic data for small areas (Spielman and Singleton, 2015). Up through 2000, the decennial census collected extensive demographic and economic data through the census long-form, essentially providing a snapshot of the U.S. population once every 10 years. The ACS asks essentially the same demographic, social, and economic questions as the decennial census long form but is administered throughout the decade to provide information on a continual basis. Instead of a snapshot which becomes outdated, the ACS works as a "rolling survey" and has been described as a "moving video image, continually updated to provide much needed data about our nation in today's fast-moving world" (Cooper, 2005). Compared with the decennial data, the ACS is based on relatively smaller sample sizes and has larger of margins of error (U.S. Census Bureau, 2009; Citro & Kalton, 2007; Spielman, Folch, & Nagle, 2014; Spielman & Folch, 2015).

The utilization of population statistics published from the decennial census and ACS extend well beyond the founding tenant of Congressional representation. Hundreds of billions of dollars in federal funding are allocated using population estimates (Blumerman & Vidal, 2009). At the state or smaller scale, population estimates are used for a wide variety of commercial and public purpose including planning, budgeting, analytical purposes, management, and business decision-making (Swanson & Pol, 2010). Specific examples include selection of sites for public facilities, shopping malls, and housing; disaster prevention and management (Hauer, Evans, & Alexander, 2015); planning transportation routes; analyzing demographic trends; setting geographic boundaries for political districts; public health surveillance; and determining eligibility for government programs (Cai, 2007; Smith & Cody, 2013). Given their importance, it is not surprising that demand of population estimates are increasing to federal, state, and local government; small businesses, corporate groups, research institutes, and so forth.

The expansive reliance of governmental, commercial, and public agendas on small area population estimates creates a need for accurate annual small-area population estimates that include demographics. Numerous efforts have been undertaken to develop methods for accurate, annual small-area demographic population estimation: Housing Unit (HU) method (Hoque, 2010), Regression method (Goldberg & Balakrishnan, 1960; Hoque, 2010;), Censal Ratio method, Component method (Swanson, Schlottmann, & Schmidt, 2010), Sample based method, Inter-censal method, Integrated method (Cai, 2007; Deng & Wu, 2013) , and so on (Swanson & Tayman, 2012). HU method is the most commonly used method for making population estimates in the U.S. because it can produce reliable population estimates at multiple levels of geography by incorporating a wide variety of data sources (Baker et al., 2013; Smith & Cody, 2013; Rayer et al., 2015). HU method is the methodology used for the annual subcounty population estimates published by U.S. Census Bureau (U.S. Census Bureau, 2014). Lacking the granularity of person counts, the HU method cannot estimate demographic characteristics (e.g. age, sex, race/ ethnicity) directly (Bryan, 2004). Another common method is the standard cohort-component approach; however, this method is extremely difficult to implement at subcounty geographic levels since corresponding direct data (e.g. Internal Revenue Service (IRS) Data, college age population) are not available for cities/towns (Tang, 1999). As a result, it is not surprising that no single method is able to produce quality small-area demographic estimates.

Connecticut is the 3rd smallest state by area but the 4th most densely populated of the 50 states in U.S. (U.S. Census Bureau, 2012). While the majority of the states in the U.S. use county as the principal geographic level for local governance, Connecticut relies on 169 towns to govern its 3,590,886 residents (U.S. Census Bureau, 2015). Town populations range in size from over 100,000 to less than 1,000. The objective of this study is to develop and test a multi-level model for producing reliable population estimates for the years between decennial censuses for each of the 169 towns in Connecticut. Town estimates must also provide detailed data by age, sex, race, and Hispanic origin.

4. Data inputs

The initial task for this project was to identify data sources that could provide population information for each town in Connecticut. Standard components of population change include births and deaths (natural increase) and migration. Birth and death events are readily available from the Connecticut vital records system. Estimating migration would require inputs for all ages between birth and death. Based on a literature review, potential sources included housing permits, income tax data, and utility data. Through existing collaborations with other state agencies, the authors identified school enrollment, Department of Motor Vehicle licenses, voter registration, and Medicare enrollment as resources with high potential. Population data from the decennial census would be necessary to evaluate model inputs, provide control totals for towns and demographic subgroups, and validate the model prediction. Lastly, through the process of identifying data sources, it became apparent that some datasets would only be available by ZIP code and thus a conversion table from ZIP code to town would also be required.

The target demographic identifiers for this project are town, five-year age groups, sex, and five mutually exclusive race and Hispanic origin categories. Based on previous work by the Connecticut Department of Public Health (DPH) authors, these aggregations are the lowest level of demographic granularity feasible for modeling that still meet the minimum groupings necessary for most of the local, state, and federal reporting of health statistics. Due to Office of Management and Budget (OMB) federal standards, it is common for government agencies to use only five race categories despite the expanded race and mixed race reporting commonly used today. Furthermore, the proportion of the Hispanic population in some towns is so small that providing race by Hispanic origin would be unreliable.

Table 1. Data sources and applications in small area population projection

Dataset	Population Component	Town	5-year Age Groups * Sex	Race/Ethnicity
Bridged 7/1/2010 POP	All	Y	Y	Y
FSCPE Annual TOWN Estimates	All	Y	N	N
NCHS Annual ASRH Estimates	All	N	Y	Y
Births	Youth	Y	Y	Y
Deaths	Elderly	Y	Y	Y
Household Population	All	Y	Y	N
Group Quarters	All	Y	Y	N
School Enrollment	Youth	Y	Y	Y
DMV	Adults	Y	Y	N
Medicare	Elderly	N	Y	Y
Utility	All	N	N	N
IRS	Employed	N	N	N
Voter Registration	Adult	Y	Y	N

Note: FSCPE represents The Federal-State Cooperative for Population Estimates;
NCHS represents National Center for Health Statistics

4.1 Census data

Decennial census data provides information about a community's entire population by age, sex, households, families, housing units, race and ethnicity origin groups etc. (U.S. Census Bureau, 2013). The decennial census provides a base upon which annual estimates of change can be applied for each post-censal year. It also serves as a benchmark to which model inputs can be compared in order to assess how well they may represent the full population of Connecticut.

> Strengths: actual headcount, provides Minor Civil Division (MCD) level, is the national standard for population data, provides housing units
>
> Limitations: 31 race categories require bridging, bridged counts are only available for total population (although household (HH) and Group Quarters (GQ) may be available in the future)

Extensive decennial census data is available online through the American FactFinder giving access to all variables collected through the census at the MCD level. Decennial census data were used as benchmarks when evaluating model inputs. Household-related factors, such as housing units and occupancy rates, were collected as potential inputs in the model. Tables were downloaded by town for age and sex to create a local data archive of decennial census counts for this project. Due to social factors surrounding self-identification of race, a portion of respondents will report “Some Other Race” and write-in their response, often reporting an ethnicity instead of a race. A special request was submitted to the US Census Bureau to receive town-level population counts tabulated by 5-year age group, sex, and Hispanic origin but with modified race categories. In the modified race dataset, the “Some Other Race” responses reported by respondents have been allocated into the 31 single and mixed race combinations based on written responses or imputation. Using the same bridging constants applied by NCHS (Ingram, 2003), all multiple race responses were allocated into 5 race categories: White, Black, Native American, Asian, and Native Hawaiian and Other Pacific Islander. After bridging, ethnicity and race were collapsed into five race/ethnicity categories: White non-Hispanic, Black non-Hispanic, Native American non-Hispanic, Asian non-Hispanic, or Hispanic (any race).

4.2 Birth cohort

Population growth occurs through natural increase when births outpace deaths causing the population count to increase. Consequently, births and deaths are two significant components of population change. Birth and death events are collected through vital event registries in each of the 50 states and Washington D.C. and include all demographic and geographic components necessary for this project. As the stewards of birth and death statistics for Connecticut, the DPH authors tabulated the number of births and deaths occurring in each of the 169 towns by age (as of July 1), sex, and race/ethnicity for each year from 2005 through 2014.

A major strength of vital event data is that virtually all birth and death events are registered, thereby providing a model input that represents all persons who were born or died in the State of Connecticut. For birth data, the child’s date of birth and sex are certified and town of residence, race and Hispanic origin are self-reported by the mother. For death data, the decedent’s date of death, date of birth, sex, race and Hispanic origin and town of residence are reported by family or gathered from personal identification (Driver’s license) or health records. Limitations of both datasets are the race and Hispanic origin identifiers – neither of which are self-reported by the newborn or deceased. The newborn’s race/ethnicity is assumed to be the same as the mother’s race and ethnicity, which may not be consistent with how the child’s race and ethnicity is reported for the decennial census or ACS. The decedent’s information is subject to reporting errors and incomplete information. Overall, the birth and death records are excellent inputs for natural increase.

After reviewing the Medicare enrollment data, the authors concluded that deaths are already subtracted from the annual enrollment figures. To avoid double counting deaths, a natural increase input was not used. Instead, birth cohort natural increase estimates were derived to represent residents aged 0-4 years. Using births from 2005-2010, each child born who was 0-4 years as of July 1, 2010 was aggregated into a birth cohort and counts were tabulated by town, sex, and race/ethnicity. Births with unknown race or ethnicity information were proportionally allocated into the 5 race/ethnicity categories using the race/ethnicity distribution ratios for each town.

4.3 State Department of Education (SDE)

A common factor statistical measure available for children aged 5 through 18 is school enrollment. Particularly for children aged 5 through 14, school enrollment data may be a reliable input for counting children by town, age, sex, and race/ethnicity as the State Department of Education (SDE) collects this information annually from each of the state's public school districts. SDE also receives enrollment statistics from private schools; however, the demographic components are limited to grade and town of residence.

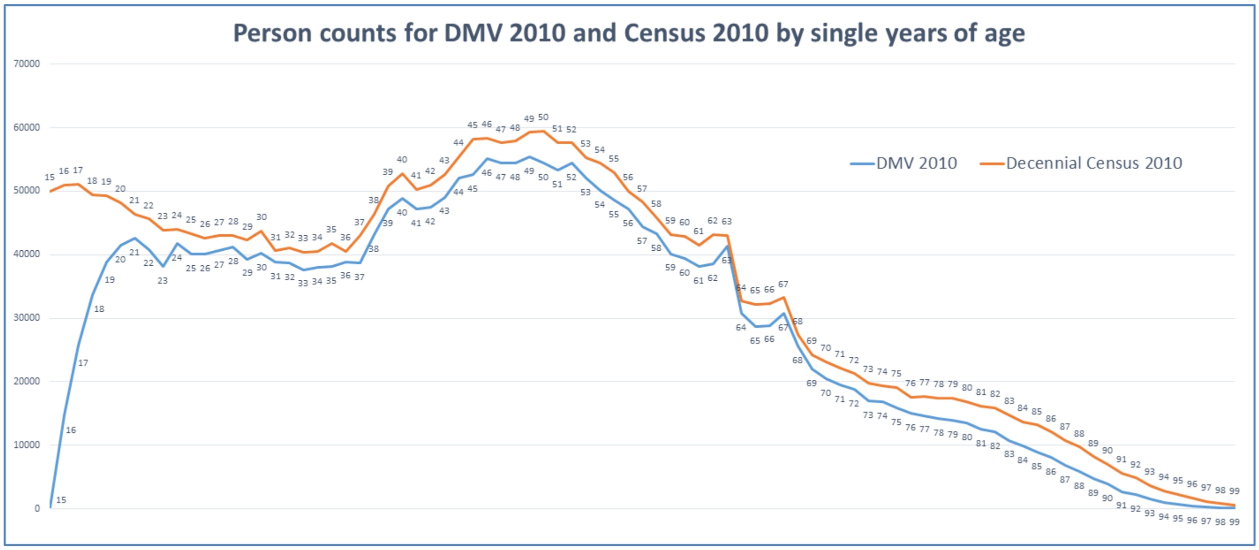
> Strengths: count of students aged 5-14 versus count of persons in census.
>
> Limitations: data sharing limitations associated with confidentiality; doesn't cover 18-19 year olds well, doesn't address dropouts, doesn't include home schooling.

After discussions with SDE staff about data sharing limitations, public school enrollment data was tabulated by town of residence, age as of July 1, 2010, and race and Hispanic origin. Sex was not included since this demographic component was not expected to have meaningful impact on the demographic distributions for ages 5 to 14 and dropping sex would reduce the number of censored race/ethnicity cells for each age group. For 2010, the number of public school students was provided by SDE by town of residence for ages 5-9, 10-14, and 15-19 and for 7 race/ethnicity categories. After reallocating the mixed and other races into the five target race/ethnicity categories, each town's race/ethnicity proportions were applied to each of the three age groups to create town-level public school counts by age group and race/ethnicity. Since private school data provided grade rather than age, grades were collapsed to approximate the target age groups: grades K-4 for 5-9 years, grades 5-9 for 10-14 years, and grades 10-12 for the 15-19 years. These counts by town and age groups were then proportionally distributed into the five race/ethnicity categories using the public school proportions. Public and private school counts were added together to create the final school enrollment inputs.

4.4 Department of Motor Vehicles (DMV) data

Through existing agreements with the Department of Motor Vehicles (DMV), DPH staff had access to motor vehicle license data. For licensure, the DMV collects date of birth, sex, and residence address making DMV data a potential resource. DPH staff tabulated the total number of licenses for 2010 by single year of age and compared the counts to the 2010 census counts to assess the coverage of the state population (Figure 1).

Figure 1. Person Counts for DMV 2010 and Census 2010 data by single years of age.



At the state level, DMV licenses were well correlated with census counts for most ages. Based on these results, DMV data appear to be a viable input for estimating town populations by age and sex for persons 20 years of age and older.

> Strengths: People 16 years of age and older are eligible to apply driver license; CT is primarily a suburban and rural state; the percentage of people who have a driver license is extremely high for adults, making DMV data quite suitable for estimation of people 16 years of age and older; DMV licenses require renewal every 6 years and require updating of residence with 72 hours of a change of address; includes non-drivers IDs; requires residency in CT.
>
> Limitations: May underrepresent urban areas for select ages and income levels as driver's licensure is relating to driving a car; not available to non-legal residents that would be counted by the census; absence of race/ethnicity information; self-reported town of residence name is not always consistent with the 169 official Connecticut towns.

Demographic components for the DMV input were limited to age, sex, and town of residence. Age was calculated as of July 1, 2010 using date of birth and sex as provided. Tabulation of town of residence required substantially more processing. Self-reported addresses often utilize a local area name or a postal area name as the "town." For licenses that reported a name other than one of the 169 official towns, two additional steps were required. Using an in-house crosswalk, the local area and postal area names known to correspond to a single town were reassigned to the official town name. When the reported town name could not be reassigned by name alone, the five-digit ZIP code from the address was used to proportionally allocate licenses into official towns using a conversion method described in Section 5. For example, "Mystic" is a local area that spans two official towns, Groton and Stonington. Using the conversion method, records with a Mystic ZIP code (and not Stonington or Groton as the town name) were split proportionally between Groton and Stonington. DMV data were then tabulated by five-year age group, sex, and town for modeling.

4.5 Medicare

Medicare is a government-funded health care program available to legal citizens aged 65 and over. Qualidigm, a health data organization with access to Medicare enrollment data, was able to provide counts by age as of July 1, 2010, by sex, race and Hispanic origin and ZIP code.

> Strengths: Majority of eligible persons enroll around ages 65-67; updated monthly by purging deaths; contains race/ethnicity information and detailed ZIP codes (ZIP+4).
>
> Limitations: Confidentiality limitations mean that only ZIP codes without town name information are available; non-legal residents not counted; unclear how the +4 portion of the ZIP code is populated; unclear if resident address ZIP code or mailing address ZIP code was provided.

Qualidigm tabulated counts of enrollments by five-year age group as of July 1, 2010, sex, 7 race and Hispanic origin categories, and ZIP+4 for 2010. Mixed and other races were equally distributed across the five target race/ethnicity groups by ZIP code. With no town information, ZIP codes would be coded as one of the 169 official towns first by cross-walk and second by proportional allocation. Using methods described in more detail later, a ZIP+4 to block group to MCD cross-walk was used to assign towns for all reported ZIP+4 values that existed in the cross-walk. For ZIP+4 that were not in the cross-walk and 5-digit ZIP codes known to span town boundaries, counts were proportionally allocated to component towns using a weighted conversion table. The Medicare data was then re-tabulated to create total counts by five-year age groups, sex, race/ethnicity and town for modeling.

4.6 Internal Revenue Service (IRS)

The IRS Statistics of Income Division (SOI) and the U.S. Census Bureau have been working on releasing the United States Migration data for several decades. As an important source of recording the movement of individuals from one place to another, these data are mainly collected based on the year-to-year address changes which are reported on individual income tax returns filed with the IRS during two consecutive calendar years. SOI's migration data present migration patterns by State or by county for the entire United States and are available for inflows—the number of new residents who moved to a State or county and where they migrated from, and outflows—the number of residents leaving a State or county and where they went. The IRS data include the number of returns filed, which approximates the number of households that migrated and number of personal exemptions claimed, which approximates the number of individuals. The individual income tax data are available by State, ZIP code, and size of adjusted gross income.

> Strengths: An input used by other population estimation models, including census data; data used to produce migration data products come from individual income tax returns filed prior to late September of each calendar year and represent between 95 and 98 percent of total annual filings.
>
> Limitations: In this project the IRS data may not suitable for population estimation because the household level data does not provide information about age, gender, or race/ethnicity. Income tax data migration statistics may be too coarse to provide reliable inputs for smaller towns. People who are not required to file United States Federal income tax returns are not included in this file, so the data under-represent low income individuals and the elderly. Returns filed after

September are not included and may be related to complex returns that report relatively high income, and so the migration data set may under-represent the wealthy. The matching process also results in some returns to be excluded from the counts. When the current-year tax return is compared to the prior-year tax return, only the Social Security Number of the primary taxpayer is considered. If a secondary filer exists (as in the case of a married couple filing jointly), that Social Security Number is not recorded or compared in creating the migration dataset. Besides the above limitations, there also exist the data filing delay and filing mainly based on the individuals' willingness circumstances which potentially impacts the data's consistency, competence and reliability.

After reviewing the IRS data and searching other data sources, in terms of the household base, utility data would be a better option as an alternative of IRS Migration Data (see utility data). Thus, for this project, IRS data was deemed to be less informative than other inputs, such as DMV and utility data, and was not included in the model. IRS data may serve as reference data to validate the population estimation results.

4.7 Utility

In estimation of population, utility data provides exhaustive information for the mobility of households (Courgeau, Nedellec, & Empereur-Bissonnet, 2000). Swanson, Carlson, & Roe (1992) and Swanson, Carlson, Roe, & Williams (1995) tested the Local Expert Procedure method with incorporating residential population from utility data. Hepburn, Mayor, and Stafford (1976) used electrical utility data for estimating population in market area. Comparing with using building-permit data, the use of utility data may reduce the size of errors (Starsinic & Zitter, 1968). Rayer et al. (2015) estimated count of households for Florida and each of its cities and counties using electric customers. Thus, utility data is also considered as one of the potential data inputs for this project.

Strengths: Active residential utility meters are perceived to represent full coverage of active households; utility data includes meter counts as well as move orders tabulated by from and to ZIP codes that can estimate migration; does not overlap with Group Quarters.

Limitations: Data comes from multiple utility providers with different methods for identifying active meters. Counts of electricity meters do not necessarily equal the number of occupied housing units as multiple units may be tied to a single meter. For example, it is very common in urban or suburban areas that, some apartment buildings, there is a single meter that covers all of the units. Consistently, the counts of electricity meters are lower than the number of housing units from the decennial census and ACS. Limitations were highlighted in accuracy of data as there are three types of move orders which the migration data is based on which include, turn off orders (disconnects), turn on orders (new service), and transfer of service to landlord or other individual (transfers). This is problematic when the transfer is back to a landlord that lives outside of the Eversource and/or the United Illuminating data. Time of snapshot is another limitation if using utility data. As shown in Table 2, data from different utility companies vary in the time points of the meter counts and in the years for which counts are available.

Table 2. Description of Utility data

Utility Company	Data type	Premise category	Geographical level	Date
Eversource	Total counts	Single family, Multi-family	Town, ZIP+4	2010-04, 2011-04, 2012-04, 2013-04, 2014-04, 2015-04
United Illuminating	Migration summary	Single family, Multi-family, Other	Town, ZIP	2011-04-03 to 2012-04-01, 2012-04-01 to 2013-04-01, 2013-04-01 to 2014-04-01, 2014-04-01 to 2015-04-01
	Total counts	Single family, Multi-family, Other	Town, ZIP	2011-04-03, 2015-04-01
Bozrah Light and Power Company	Total counts	-	Town	2010-04-01, 2011-04-01, 2012-04-01, 2013-04-01, 2014-04-01, 2015-04-01, 2016-02-29
Groton Department of Public Utilities	Total counts	-	Town, ZIP	2016
Town of Wallingford Department of Public Utilities-Electric Division	Total counts	-	Town	2010, 2011, 2012, 2013, 2014, 2015
Jewett City Department of Public Utilities	Total counts	-	Town	2014-2015
Norwalk City of Third Taxing Dist.- Electrical Department	-	-	-	-
South Norwalk Electric and Water	Total counts	-	Town	2011, 2012, 2013, 2014, 2015
City of Norwich Public Utilities	Total counts	Single family, Multi-family	Town, ZIP	2011-04, 2012-04, 2013-04, 2014-04, 2015-04

The "-" indicates that information was not provided.

Pre-processing of the utility data required the conversion of the ZIP code-based counts into town-based counts. Municipality-based utilities provided counts by town of residence; however, Eversource and United Illuminating data required the conversion of ZIP codes to town. The geo-processing steps discussed later were applied to convert all ZIP-based inputs to town. After conversion to town totals, meter counts with an incomplete series of time points were smoothed to create a constant rate of change for each year between existing time points. When 2010 inputs were not available, the rate of change was negatively applied to the 2011 count to estimate the 2010 count. For meter counts with only one time point, the provided count was used for all time points. After converting ZIP codes and populating missing time points, meter counts were summed by town to create a single meter count for each town.

4.8 Voter Registration

Statewide voter registration files were recently published online for public access. The files contain the information required to register to vote, including date of birth, sex, and resident address.

> Strengths: Voter registration files have the potential to provide high geographic precision since physical location of residence is requirement for determining voting district.
>
> Limitations: Voter registration files are maintained by local municipalities and it is unclear if differences exist between towns in the management and updating of registration data, particularly the purging of death records; voter registration is an individual's choice and will not equitably represent all demographic subgroups.

The migration of persons between towns creates ongoing need to purge voters from the previous town's database which may not be feasible if the migrating person does not re-register to vote or does not provide information about his/her prior registration. Given these known limitations, voter registration was not used to provide person counts, but was utilized to validate geo-processing tables. By comparing self-reported ZIP code with the voting district town, a proportional distribution of ZIP code to town was tallied and used to validate the ZIP-to-town conversion ratios provided by GeoLytics, which are discussed in detail in the next section.

5. Geo-processing Methodology

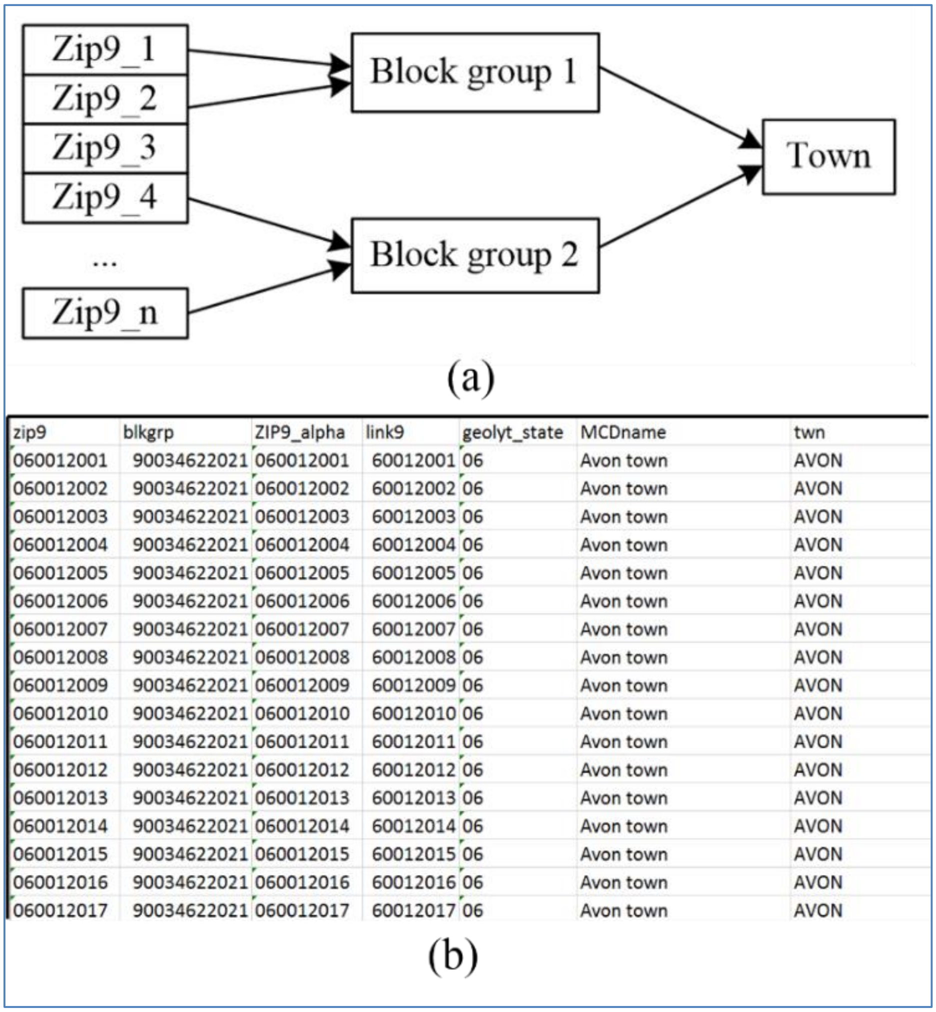
Conversion of ZIP codes to town proved to be a formidable task requiring a multi-level approach to maximize validity. ZIP codes are a construct of the United States Postal Service and support the delivery of mail. While the U.S. Census Bureau provides ZIP Code Tabulated Areas (ZCTAs) and several commercial companies sell ZIP code polygons, ZIP codes as utilized by the U.S. Postal Service are based on points (mail delivery locations) and mail routes (lines) and not polygons. With polygon data being ideal for geoprocessing, this research into ZIP codes highlighted the need to validate in which town a ZIP code(s) fall utilizing a variety of datasets such as utility and voter registration to validate which towns include which ZIP and ZIP+4. This is not an unknown issue, and with postal routes updated as need it was decided the challenge of working with ZIP code data for this project, particularly ZIP+4 data could include different vintages of ZIP code areas and as a result polygon files for ZIP codes were used in limited applications. While no ideal process exists for converting ZIP code data to town, a multi-level approach was taken in this project to utilize the most detailed geographic information first (ZIP+4) and then to partition 5-digit ZIP codes proportionally into towns.

5.1 GeoLytics ZIP+4 to Block Group

ZIP+4 codes use the standard 5-digit code to identify the postal area and includes an additional 4 digits for carrier routes and delivery types. The added detail in the additional 4 digits is valuable for associating the address that the ZIP+4 represents with a smaller geographic area. Smaller geographic associations should increase the accuracy of the town of residence when assigning ZIP codes that span town boundaries into the appropriate town.

To assist with identifying which zip code is within which town, a ZIP+4-to-Block Group table was purchased from GeoLytics. Block Groups are census area designations that exist within individual towns. Since Block Groups are contained within towns, the authors were able assign the census MCD (town) for each Block Group which in turn creates an effective ZIP+4-to-MCD table (Figure 2). Per discussion with GeoLytics staff, the assignment of ZIP+4 to Block Group was based on the centroid of all of the segments of the ZIP+4 carrier route. This centroid method of assignment incurs some error as portions of ZIP+4 carrier routes will span Block Group boundaries; however, the authors believe the assignment error is effectively minimized by the smaller geographic areas represented by the ZIP+4 routes and the nesting of the Block Groups within the towns. Incorrect assignment along the town boundaries is assumed to be evenly distributed on both sides of the boundary and ultimately represents fairly small ZIP+4 areas.

Figure 2. Conversion model based on GeoLytics product,
(a) workflow for aggregating ZIP+4 (denoted ZIP9) to town level,
(b) table structure of GeoLytics product.



zip9	blkgrp	ZIP9_alpha	link9	geolyt_state	MCDname	twn
060012001	90034622021	060012001	60012001	06	Avon town	AVON
060012002	90034622021	060012002	60012002	06	Avon town	AVON
060012003	90034622021	060012003	60012003	06	Avon town	AVON
060012004	90034622021	060012004	60012004	06	Avon town	AVON
060012005	90034622021	060012005	60012005	06	Avon town	AVON
060012006	90034622021	060012006	60012006	06	Avon town	AVON
060012007	90034622021	060012007	60012007	06	Avon town	AVON
060012008	90034622021	060012008	60012008	06	Avon town	AVON
060012009	90034622021	060012009	60012009	06	Avon town	AVON
060012010	90034622021	060012010	60012010	06	Avon town	AVON
060012011	90034622021	060012011	60012011	06	Avon town	AVON
060012012	90034622021	060012012	60012012	06	Avon town	AVON
060012013	90034622021	060012013	60012013	06	Avon town	AVON
060012014	90034622021	060012014	60012014	06	Avon town	AVON
060012015	90034622021	060012015	60012015	06	Avon town	AVON
060012016	90034622021	060012016	60012016	06	Avon town	AVON
060012017	90034622021	060012017	60012017	06	Avon town	AVON

5.2 GIS-based conversion models

Geospatial overlays are often used in Geographic Information Systems (GIS) software to compare or align geographic areas. By overlaying the ZIP code boundaries with the town boundaries, the overlapping areas could be converted into a crosswalk between ZIP code and town where the proportion of the ZIP code area in each town would determine the ratio for splitting a ZIP code. This method requires two boundary files. The State of Connecticut publishes the official town boundary file for public use. A ZIP code boundary file is not published by the United States Postal Service (USPS) which is consistent with the notion that ZIP codes do not represent areas but are instead routes. With no official boundary file for ZIP codes, alternative sources were sought. Through existing licenses, the project staff had access to ZIP code shapefiles from ESRI, Tele Atlas, and Navteq. Project staff elected to use the ESRI shapefile available with the ArcGIS software. This shapefile contains 5-digit ZIP code (ZIP5) polygons.

Three conversion methods were evaluated for this project. The first conversion model is an area-based weighting method, the second method is a built-up area-based weighting method, and the last model is a GeoLytics population density conversion model.

5.2.1 Area-based weighting method

Area-based weighting relies on the assumption that the people are distributed evenly in each ZIP5 polygon. Using GIS software, the town and ZIP code boundaries are overlaid and the polygon areas are merged to create polygons that represent each unique town/ZIP5 combination (see Figure 3). The area of each town/ZIP5 polygon is calculated and used to create a ratio of the ZIP5 code that will be assigned to that town.

Figure 3. The overlap of the boundaries of town and ZIP5 in Connecticut.

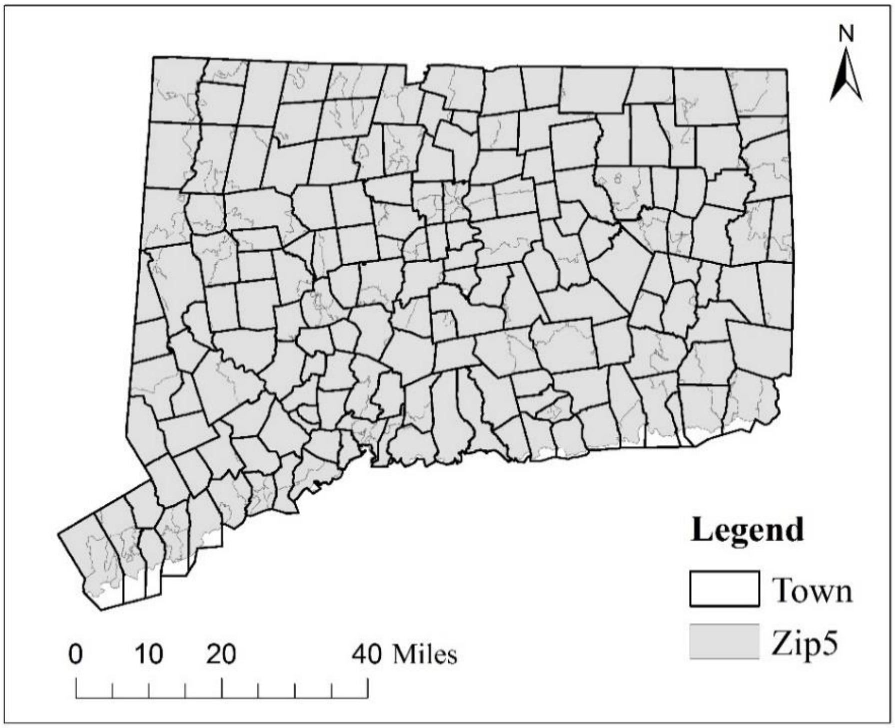
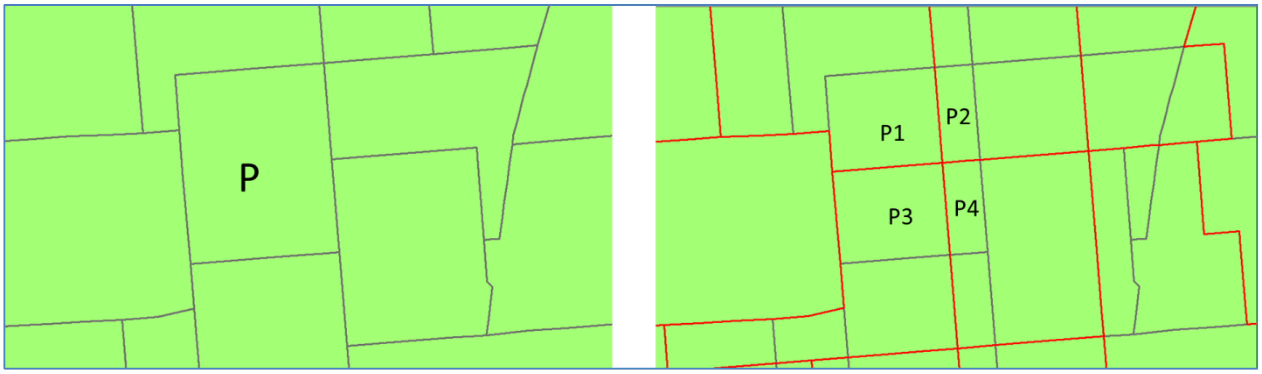


Figure 4 depicts how ZIP5 polygon *P* is overlapped by four towns, where black lines represent ZIP5 boundaries and red lines represent town boundaries. Using DMV data for illustration purposes, the number of DMV drivers in ZIP5 *P* will be assigned to one of the four overlapping towns: P1, P2, P3, and P4. Based on the assumption that the DMV drivers are distributed evenly in ZIP5 *P*, the number of DMV drivers in ZIP5 *P* should be split into the four towns based on the proportion of the ZIP5 P area that overlaps each town.

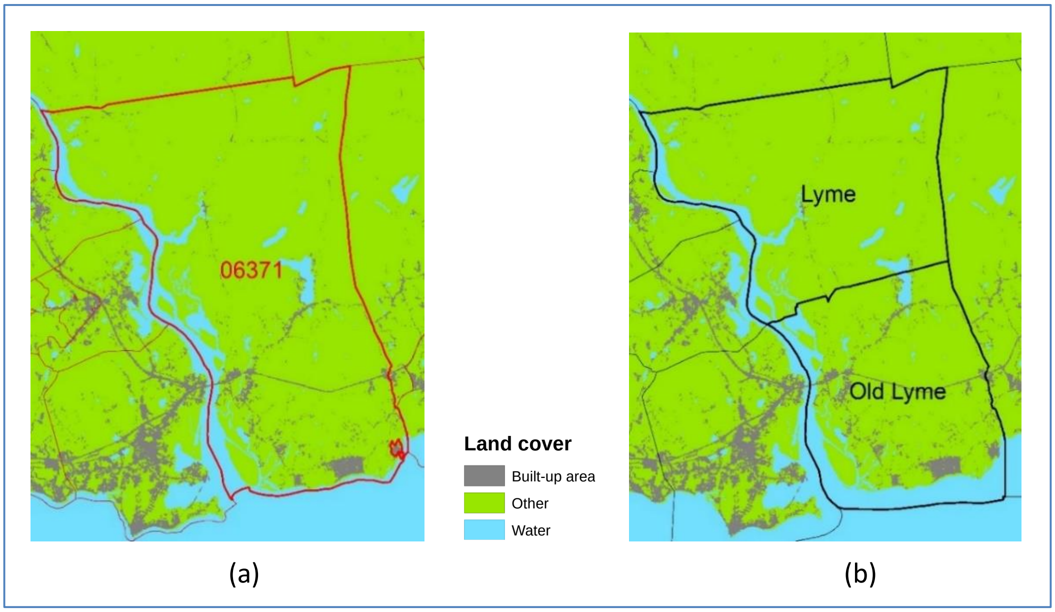
Figure 4. Depiction of the area-based weighting based method.
Black lines represent ZIP5 boundaries. Red lines represent town boundaries.



5.2.2 Built-up area based weighting method

The area-based weighting method assumes that people are distributed evenly in each ZIP5 area. In reality, people are unevenly distributed geographically, congregating in urban and residential areas. Figure 5 shows the land cover map in ZIP code 06371 (figure 5a) and the town boundaries that overlap 06371 (figure 5b). The southern portion of the ZIP5 area displays more land use for streets, housing, and buildings. When the towns are overlaid, visually it is obvious that Old Lyme has more built up areas than Lyme. This uneven distribution is further validated by comparing the total population in town of Lyme and Old Lyme from decennial census data. To address the uneven distribution of the population, we modified the area-based weighting method to use the built-up area as a weighting factor in the calculation of the town/ZIP5 ratio.

Figure 5. Built-up area weighted based method to convert the ZIP code level IRS data to town level



After performing both weighting methods for converting ZIP5 to town, the results were evaluated. A major limitation of the results was the extensive splintering that occurred. Small sections of overlap were created when town and ZIP5 boundaries were close but did not exactly overlap. These extra splinters increased the ZIP5 to town conversion table by about 30% yet the portion of the population they might represent was extremely small. The other sources for ZIP code boundaries were evaluated only to find that the ZIP code boundaries from each source were highly disparate, undermining the confidence of using any of them to accurately and reliably assign a town. In the end, neither model was viable.

5.2.3 GeoLytics ZIP5-to-MCD Correspondence File

GeoLytics, the vendor that provided the ZIP+4-to-Block Group crosswalk, offers a custom area-to-area correspondence file that can be created using population weights. After verbally discussing our needs, we contracted GeoLytics to develop a ZIP5-to-town correspondence file. The first step required assigning a location to each ZIP+4. The vendor geocoded the addresses in the 2010 USPS ZIP+4 database and calculated the centroid location of the ZIP+4. The ZIP+4 was then assigned the Census 2010 Block Group in which the centroid was located. After assigning all ZIP+4 values to a Block Group, the ZIP+4 values were aggregated to the ZIP5 level by Block Group. Since the Census Block Group exists wholly in each Census MCD, the ZIP5-to-town correspondence could be created while weighting the allocations by Block Group population. This approach effectively weights the correspondence by the population density of the component Block Groups.

Prior to adopting the ZIP5-to-MCD correspondence file, the proportions were compared with two other datasets that provide both town and ZIP code in an effort to validate the results. The voter registration dataset provides addresses for each voter registered in the 169 towns. By cross-tabulating the voting town with the ZIP5, a ratio of ZIP5-to-town was calculated. The same approach was used with the DMV data. While neither the voting registration nor the DMV data could be used for person counts, the proportion of persons reporting a particular ZIP5 and town should be representative of the geographic distribution of the town.

The GeoLytics, Voter Registration, and DMV correspondence ratios were merged into a single file by ZIP5 and town and manually reviewed for consistency in ratios. The vast majority of ratios were consistent for the three inputs. A few ZIP5 values showed variation between all three sources making it difficult to determine which source was most accurate. This was expected as a few ZIP5 areas expand irregularly across multiple towns while representing a small population. None of the GeoLytics ZIP5 ratios were found to be extremely discrepant from the other two sources. Given the interdependence of the ratios between town and ZIP5 to sum to 100%, no edits were made to the proportions provided by GeoLytics. On a few occasions, a ZIP5 was not found in all 3 inputs. In such cases, the ZIP5 was reviewed to determine if it was associated with the same towns in the remaining two inputs. If so, the ZIP5 was added to the GeoLytics file as a new ZIP5 entry. This was also expected; as postal codes change over time but some residents continue to use historical codes and residents sometimes reported a mailing ZIP5 (P.O. Box). The GeoLytics conversion ratios were finalized for the mapping of ZIP5 codes to towns for each input where zip code data is provided.

5.3 Hierarchy for geo-processing

For inputs that contain ZIP+4 data only (Utility, Medicare), the ZIP+4-to-Town crosswalk was applied first. ZIP+4 records that were found in the table were assigned an MCD using the intermediary Block Group. ZIP+4 records that were not found in the crosswalk were retained for additional processing using the ZIP5-to-Town correspondence ratios. A small percentage of records that remained without an MCD assignment were proportionally distributed among all of the towns.

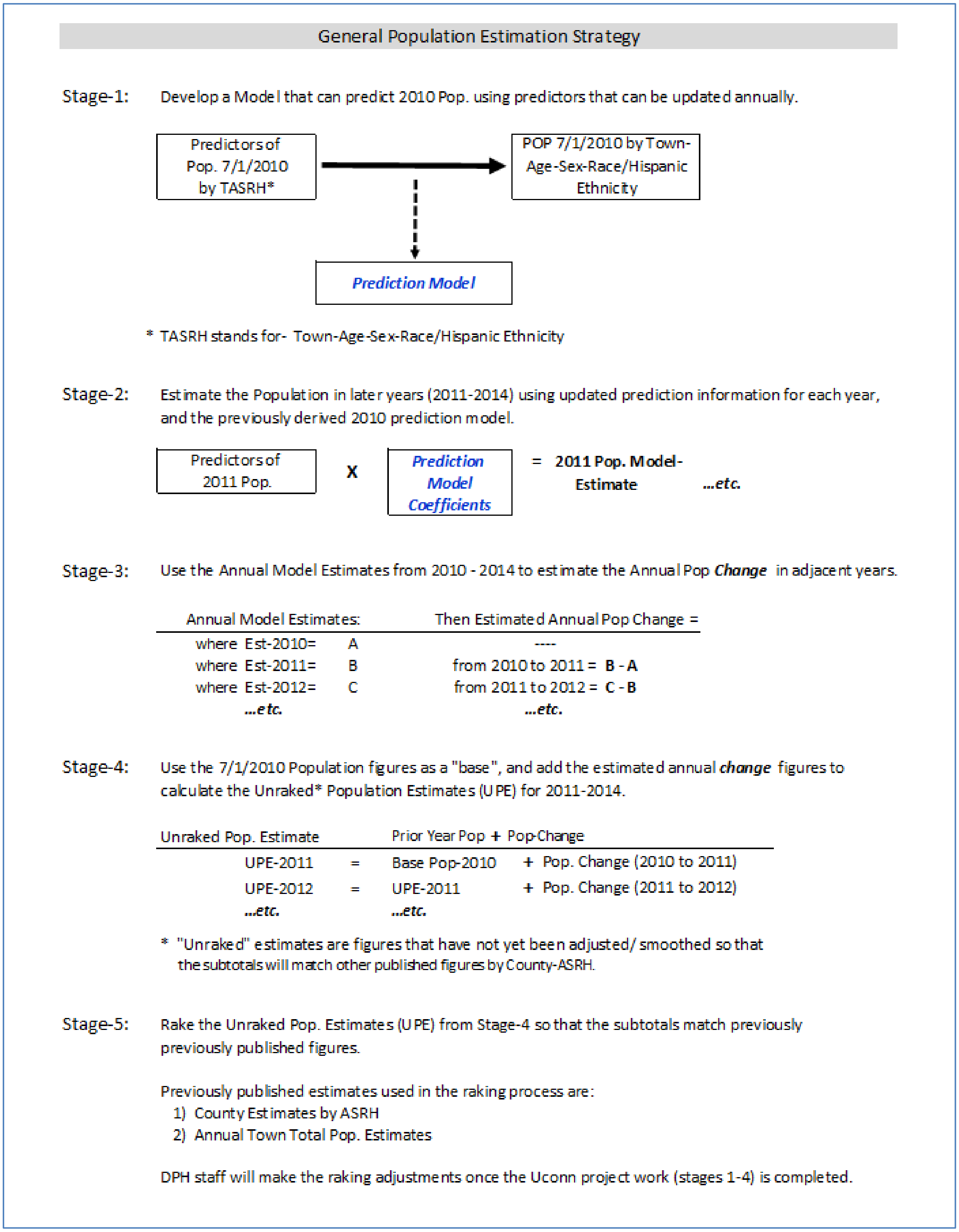
For the DMV data, both town name and Zip code were available but ZIP+4 codes were uncommon. Licenses that reported a town name that was the same as one of the 169 MCD names were assigned the self-reported town. Licenses that reported a local area name or postal area name that is known to correspond to a single MCD were assigned that MCD (e.g., Bantam was assigned to Litchfield). When the MCD could not be assigned by reported town name alone, the five-digit ZIP code from the address was used to proportionally allocate licenses into official towns using the ZIP5-to-Town correspondence ratios. A small percentage of records that remained without an MCD assignment were proportionally distributed among all of the towns.

6. Prediction Model Development Methodology

General method:

The general strategy adopted for developing population estimates is outlined in the “General Population Estimation Strategy” table (see Figure 6).

Figure 6. General Population Estimation Strategy



General Population Estimation Strategy

Stage-1: Develop a Model that can predict 2010 Pop. using predictors that can be updated annually.

Predictors of Pop. 7/1/2010 by TASRH*
POP 7/1/2010 by Town-Age-Sex-Race/Hispanic Ethnicity
Prediction Model

* TASRH stands for- Town-Age-Sex-Race/Hispanic Ethnicity

Stage-2: Estimate the Population in later years (2011-2014) using updated prediction information for each year, and the previously derived 2010 prediction model.

Predictors of 2011 Pop.
X
Prediction Model Coefficients
= **2011 Pop. Model-Estimate**
...etc.

Stage-3: Use the Annual Model Estimates from 2010 - 2014 to estimate the Annual Pop ***Change*** in adjacent years.

Annual Model Estimates:		Then Estimated Annual Pop Change =
where Est-2010=	A	----
where Est-2011=	B	from 2010 to 2011 = **B - A**
where Est-2012=	C	from 2011 to 2012 = **C - B**
...etc.		***...etc.***

Stage-4: Use the 7/1/2010 Population figures as a "base", and add the estimated annual ***change*** figures to calculate the Unraked* Population Estimates (UPE) for 2011-2014.

Unraked Pop. Estimate		Prior Year Pop **+** Pop-Change	
UPE-2011	=	Base Pop-2010	**+** Pop. Change (2010 to 2011)
UPE-2012	=	UPE-2011	**+** Pop. Change (2011 to 2012)
...etc.		***...etc.***	

* "Unraked" estimates are figures that have not yet been adjusted/ smoothed so that the subtotals will match other published figures by County-ASRH.

Stage-5: Rake the Unraked Pop. Estimates (UPE) from Stage-4 so that the subtotals match previously previously published figures.

Previously published estimates used in the raking process are:

1) County Estimates by ASRH
2) Annual Town Total Pop. Estimates

DPH staff will make the raking adjustments once the Uconn project work (stages 1-4) is completed.

Different potential predictor variables are available for different age strata. As a result, the analyses were segmented to correspond to the available predictor candidates. Five separate age strata were modeled with five different models (Figure 7). These age strata are 0 – 4, 5 – 14, 15 – 19, 20 – 64, and 65 – 85 years. At the heart of this process is the analytic approach for identifying subsets of available predictors that are useful in estimating population counts.   A common characteristic of the data available for modeling and all strata was that key independent variables were continuous measures while categorical factors such as age race sex might be used to classify the predictors for estimating population by town age, sex, and race/ethnicity.

Figure 7 shows that for 0-14 and 65+, raw, unadjusted model inputs approximate the state population. For 15-64, the raw DMV inputs underrepresent actual counts; however, the undercount is stable across age groups.

Figure 7. Comparison of input population counts with April 1, 2010 Census Versus 2010 Raw, Unadjusted Model Inputs

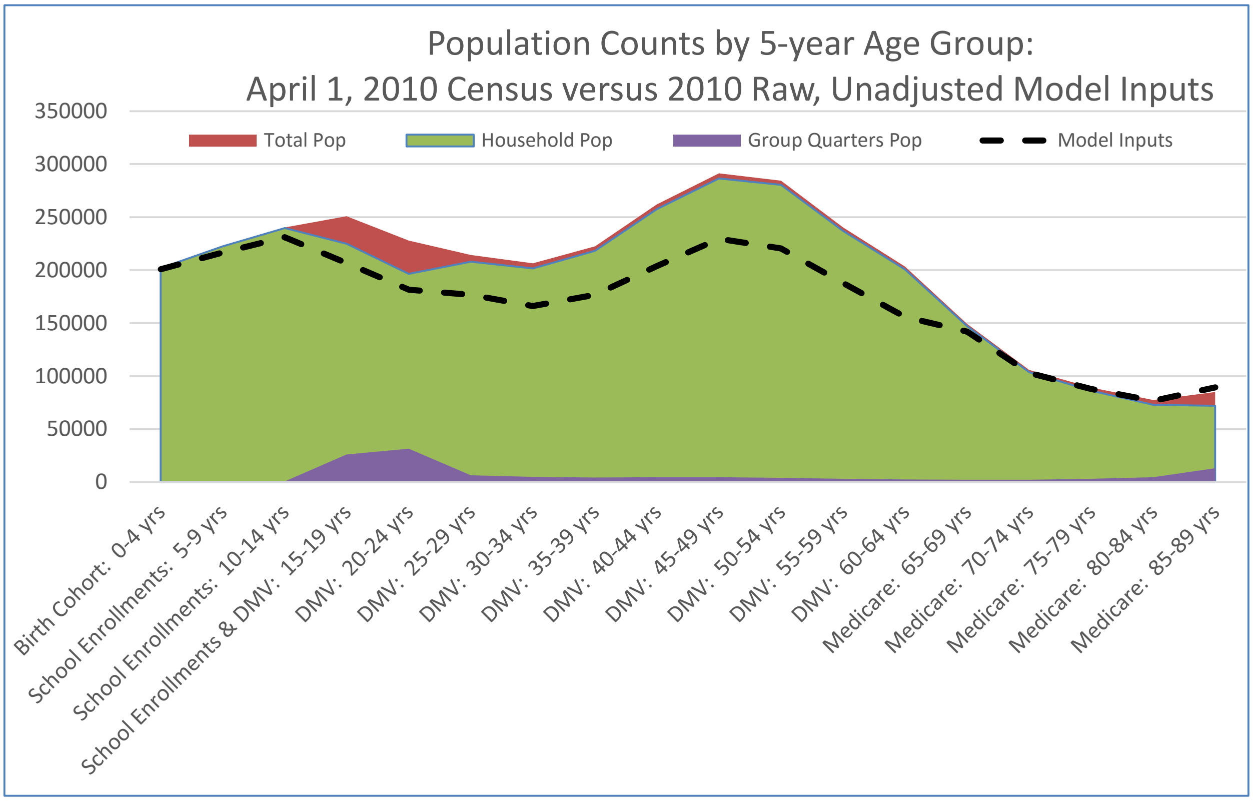
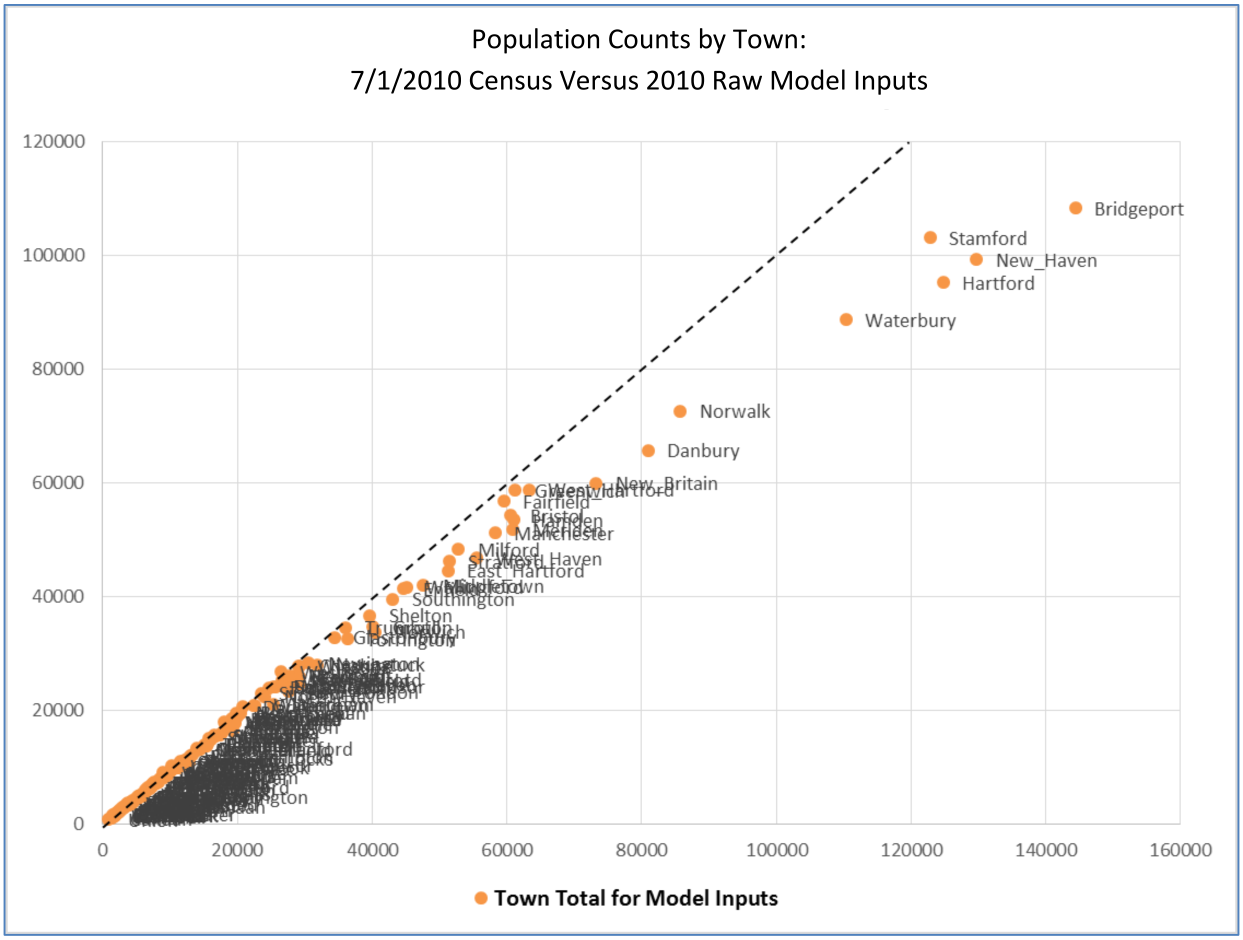


Figure 8. Population Counts by Town: 7/1/2010 Census Versus 2010 Raw Model Inputs



As illustrated in Figures 7 and 8, there is a gap between the simple sum of the raw predictor variable counts and the actual 2010 population counts. This gap was addressed by developing regression models using the 2010 Census-based population where we have detailed town-age-sex-race/Hispanic-ethnicity (TASRH) population figures. Models fit to the 2010 population by TASRH were used with new annual predictor data for 2011-2014 to produce new population estimates for 2011-2014 by TASRH. A multivariate adaptive regression splines (MARS) approach was adopted for these analyses. This approach has the advantage of allowing partially automated model development, and identification of relevant variable interactions, while also including controls that assure the selection of a parsimonious model that is not "over fitted" (see Friedman, 1991 and Zhang, 2010 for a discussion of MARS). The SAS procedure, AdaptiveReg (version 14.1) was selected for this purpose.

An important consideration for us in the model selection process was the potential for selecting "optimal" models that include "too many variables", or "overfitting" the model. Overfitting occurs when models are derived that include parameters that are selected which primarily fit the random error, or "noise" in the data, and therefore give little information about the underlying relationship of the predictors and outcomes. To reduce the chance of selecting too many model predictors we used the "Varpenalty" option in the SAS Adaptivereg procedure. We assigned a "moderate penalty" (Varpenalty=0.05) for incrementally increasing the number of variables in each model. This is consistent with the penalty-level magnitudes described by Friedman (1991). In addition, the overall model selection criteria used are resistant to overfitting. Models were selected using the Generalized Cross Validation score (GCV) criteria. The GCV score provides an efficient approximation of "leave-one-out" cross-validation procedures, where each observation is iteratively compared with model estimate values derived from all other observations (N-1). The model producing the best fit (lowest GCV value) while not adding too many new parameters was selected. Once an initial model was selected in this manner, the residuals were analyzed to determine whether the relative size of any significant deviant residuals was “large”. This residual analysis was aimed at identifying prediction errors that were “large” from the standpoint of each town, even though they might be small in terms of the overall model (see Appendix C). When studentized-residuals were significantly different than zero at $P< 0.05$ and accounted for more than 30% of the town-age-sex-race/ethnicity population count being estimated, the affected observations were re-weighted to obtain a better fit on the next iteration of the model. Weights were increased from a default value of 1.0 to 2.0 for these “extreme” cases. The maximum number of re-weighting iterations required for any model was three. In several cases no extreme residuals were identified. In most cases the overall model fit was very good ($R^2 \sim 0.99$).

As illustrated in Figure 6, the selection of predictive models that fit 2010 population estimates allows us to use the derived coefficients and calculate model-based estimates for later years (2011 – 2014). These model-based estimates are used to calculate annual population *change*. When the annual change figures are added to the base 2010 population, we have a new set of “unraked” population estimates for 2011 – 2014. The unraked estimates are figures that have not been adjusted or smoothed so that the subtotals will match other published figures, e.g. by County – ASRH. The final raking, smoothing stage will control the estimates subtotals so that they agree with 1) County estimates by ASRH, and 2) annual town total population estimates.

Specific Model Details:

1. Age 0-4: Connecticut birth and death data were used to identify persons who would have been under the age of 5 as of July 1, 2010. Births and deaths from 2005-2010 are assigned an “age” value based the time difference between the child’s date of birth (DOB) and July 1, 2010 using the SAS mdy-function. Childhood deaths were subtracted from birth counts to identify the number of net survivors. Mother’s race ethnicity information was used from birth records. Decedents reported race and ethnicity was used for classifying death records. For the purposes of this project, we assume limited migration during the first 5 years of life thereby allowing the birth geography to be maintained for this age group. Cohorts for 2011 – 2014 were identified in a similar manner.

2. Ages 5-9 and 10-14: Public and private school enrollment data were provided by the Connecticut State Department of Education (SDE) by age and town and by race/Ethnicity and town. While public-school data is available in more detail, censoring rules followed by SDE limit access to complete counts by external parties. To reduce the amount of censoring in the 2010 tables requested, we eliminated sex from the cross classifications. The main purpose in using the school enrollment data for comparison with 2010 population figures in order to calculate “adjustment ratios” for each T-A-R/E subgroup in our prediction model. Then, we will need to split these both-sexes estimates by sex. In the future, those adjustment factors can be used to extrapolate from annual enrollment “change counts” to estimate population counts for 5-19 year old groups. This analysis will also motivate our next steps about how to handle the reported enrollment data for 15-19 year olds.

3. Age 15-19: This age group was estimated using a combination of school enrollment counts and driver's license counts for the 15 to 19 age group. School enrollment data, we believe, only captured most children up to the age of 17. Driver's license figures increase gradually with age from 17 up through 24. The DMV driver's license data allowed us to identify the town, age (as of July 1, yyyy) and the sex of each licensed individual. Since enrollment counts and driver's license counts can be affected by economic factors we included a measure of town-specific poverty as a model covariate. The town-specific percent below poverty (from the 2010 Census) was multiplied by town population count in each cell to estimate the approximate number of persons below poverty by town, age, sex. DMV data included both residence ZIP Codes and self-reported town names. Both of these fields were used to determine town of residence.

4. Age 20-64: Driver's license counts by town age and sex were used as the fundamental predictors for the household population in this group. We chose household population of the dependent variable for modeling because we assume in many cases individuals living in group quarters facilities might not obtain driver's licenses at the same rate as the balance of the population. Nevertheless, we knew there would be some fraction in the GQ population that would still maintain driver's licenses, so we included the GQ population estimate as a covariate. As with the 15 – 19-year-old group, the estimated number of people below the poverty level was also used as a covariate. Appendix A includes town level population estimates for the age cohort 20-64 which compare the household population from the 2010 decennial census and compares to the predicted household population minus the 2010 decennial census household population.

5. Age 65+: Medicare enrollment figures were made available to us by Qualidigm, Inc. We were able to produce a complete cross classification of these data by town, age, sex, and race/Ethnicity. Appendix B includes town level population estimates for the age cohort 65 and over which compare the population from the 2010 decennial census and compares to the predicted population minus the 2010 decennial census population.

The SAS models for age cohorts 0-4, 5-9, 10-14, 15-19, 20-64, and 65 years of age and over have been archived to enable population estimates to be updated annually as new input data is available (see Appendix D).

7. Results

A summary of the multivariate adaptive regression splines models is provided in Table 3, "Connecticut 2010 Population Estimation Models: Results Summary Matrix". In most cases the overall model fit was very good (R^2 ~ 0.99). The maximum number of iterations required to fit extreme cases was three. In several models no extreme residuals were identified.

Table 3. Connecticut 2010 Population Estimation Models: Results Summary Matrix

		Final Model Characteristics							
Population component Covered	Notes:	Predictors variables/data			Outcome Variable	Fitted 2010 Model**			
		Continuous Variables	Categorical - Strata*	N-table cells		GCV	R^2	N Basis functions	Model-df
Ages 0-4	1	Birth & infant deaths in 5 prior years	TASRH	1,690	7/1/2010 Total Pop.	175	0.998	42	83
Ages 5-9, 10-14	2	Public School enrollment	TARH	1,690	7/1/2010 Total Pop.	382	0.999	49	97
		Private School enrollment	TA (grade)	338					
Age 15-19	3	Public+Private School data	TAS/TA	338/169	7/1/2010 Total Pop.	596	0.999	34	67
		DMV Driver's License data	TAS	338					
		Estimate Pop Below Poverty level (2010)	T	169					
Ages 20-64	4	DMV Driver's License data	TAS	3,718	Household Population Census 2010	1944	0.997	49	95
		Estimate Pop Below Poverty level (2010)	T	169					
		Electric Meter counts adjusted to 20-64 pop. size.	T	169					
		4/1/2010 GQ Pop	TA	1859					
Ages 65-85+	5	Medicare enrollments	TASRH	8,450	7/1/2010 Total Pop.	50	998	78	155

* Strata abbreviations: T-Town; A-5-yr Age groups, S-Sex, RH-Race/Hispanic Ethnicity

** Model Characteristics:

GCV- Generalized Cross-Validation score. Provides a measure of how well model predictions fit the observed data, adjusted for the number of predictors. Lower numbers indicate better fit. It provides and approximate measure of leave-one-out validation scores. The formula for the GCV is- GCV = RSS / (N * (1 - EffectiveNumberOfParameters / N)^2), where RSS is the residual sum-of-squares measured on the training data and N is the number of observations.

R^2 "R-squared " is equal to the ratio of [Explained variation / Total variation] , and varies between 0.0 and 1.0.

N Basis functions Multivariate adaptive regression splines (MARS) use basis functions as components. Basis functions typically define the relationship between a predictor or a set of predictors and the dependent variable for a specific segment or subset of the predictors range. A single basis function can include multiple predictors. The number of model basis functions is displayed.

Model-df Degrees of freedom associated with the regression model.

The multivariate adaptive regression splines (MARS) model results also provide a measure of the relative importance of the variables selected for each model. The relative important score set the value of the most important predictor to 100, and then score the other predictors proportionally. This information (Table 4) provide the reader with a clear sense of which variables were most important in the final models.

Table 4. Model-1-5 by Age Cohort, 2010

Model-1 for Ages 0-4 years, 2010

By Town, Age(1-group), Sex and Race/Ethnicity

Variable Importance			
Variable	**Number of Bases**	**Importance***	**Variable Description**
B_2010	35	100.00	CT Births- Deaths for persons 0-4 yrs as of 7/1/2010.
MCD	33	9.96	Town of residence (n=169)
RACEETH5name	27	9.86	Race by Hispanic Ethnicity (5 groups)
SEXname	1	0.12	Sex (2 groups)

***** The most important predictor is given an arbitrary score of 100. The values for other predictor are scaled relative to 100.

Model-2 for Ages 5-14 years, 2010

By Town, Age, and Race/Ethnicity

Variable Importance			
Variable	**Number of Bases**	**Importance***	**Variable Description**
SDE_POP_2010	38	100.00	Public + Private School enrollment
MCD	40	7.39	Town of residence (n=169)
RACE	28	6.95	Race by Hispanic Ethnicity (5 groups)
AGE	10	1.08	Two 5-yr age groups: 5-9 and 10-14 yrs, where age is as of 7/1/2010

***** The most important predictor is given an arbitrary score of 100. The values for other predictor are scaled relative to 100.

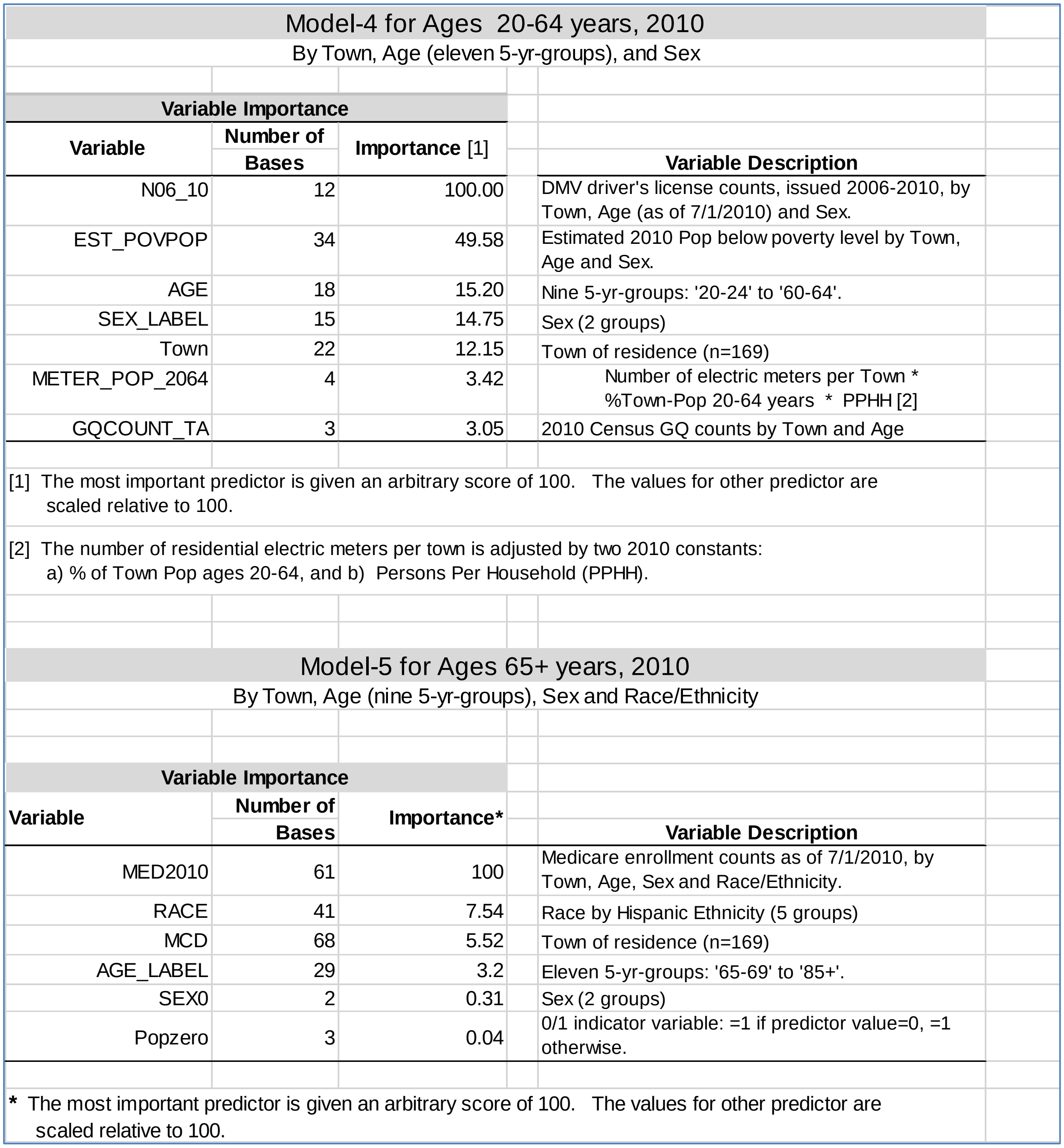
Model-3 for Ages 15-19 years, 2010

By Town, Age (1-group), and Race/Ethnicity

Variable Importance			
Variable	**Number of Bases**	**Importance***	**Variable Description**
Est_POVpop	21	100.00	Estimated 2010 Pop below poverty level by Town, Age and Race/Ethnicity
POP_2010	19	85.10	School Enrollment + DMV License counts for person 15-19 years old as of 7/1/2010.
MCD	27	34.95	Town of residence (n=169)
SEX	4	2.17	Sex (2 groups)

***** The most important predictor is given an arbitrary score of 100. The values for other predictor are scaled relative to 100.

Table 4. (Continued)



Model-4 for Ages 20-64 years, 2010

By Town, Age (eleven 5-yr-groups), and Sex

Variable Importance			
Variable	Number of Bases	Importance [1]	Variable Description
N06_10	12	100.00	DMV driver's license counts, issued 2006-2010, by Town, Age (as of 7/1/2010) and Sex.
EST_POVPOP	34	49.58	Estimated 2010 Pop below poverty level by Town, Age and Sex.
AGE	18	15.20	Nine 5-yr-groups: '20-24' to '60-64'.
SEX_LABEL	15	14.75	Sex (2 groups)
Town	22	12.15	Town of residence (n=169)
METER_POP_2064	4	3.42	Number of electric meters per Town * %Town-Pop 20-64 years * PPHH [2]
GQCOUNT_TA	3	3.05	2010 Census GQ counts by Town and Age

[1] The most important predictor is given an arbitrary score of 100. The values for other predictor are scaled relative to 100.

[2] The number of residential electric meters per town is adjusted by two 2010 constants: a) % of Town Pop ages 20-64, and b) Persons Per Household (PPHH).

Model-5 for Ages 65+ years, 2010

By Town, Age (nine 5-yr-groups), Sex and Race/Ethnicity

Variable Importance			
Variable	Number of Bases	Importance*	Variable Description
MED2010	61	100	Medicare enrollment counts as of 7/1/2010, by Town, Age, Sex and Race/Ethnicity.
RACE	41	7.54	Race by Hispanic Ethnicity (5 groups)
MCD	68	5.52	Town of residence (n=169)
AGE_LABEL	29	3.2	Eleven 5-yr-groups: '65-69' to '85+'.
SEX0	2	0.31	Sex (2 groups)
Popzero	3	0.04	0/1 indicator variable: =1 if predictor value=0, =1 otherwise.

* The most important predictor is given an arbitrary score of 100. The values for other predictor are scaled relative to 100.

Final 2010 model prediction equations were derived from the SAS AdaptiveReg application using the process recommended by Kuhfeld, 2013. The equations for each model are complicated, and have been saved as separate SAS files for later use (i.e. for 2015+ estimates). Selected results from final 2010 models are presented in the Appendices for the 20-64 and the 65+ age models. Graphs are presented that illustrative the town-specific fit of the predicted and actual 2010 population data used to derive these models.

Limitations:

Some of the most significant challenges occur where the data is “thin”, especially for small towns and among the Native American population. It is not clear for example whether individuals self-report the race and ethnicity consistently across various domains, on birth records, for school enrollment, for DMV licensing, and for Medicare. However for most race ethnicity groups, even in the smallest towns the relationship between the predictors and outcomes is fairly stable.

In some models the predictors do not contain all ASRH components. In those cases (e.g. DMV data lacks race/ethnicity) the missing demographic detail was estimated using the complete 2010 population data. The use of the 2010 reference data, e.g. for race/ethnicity, has the limiting consequence that the distribution by race ethnicity will not change over time in certain models. The model for the 20-64 years population used household population as the dependent variable. Since Connecticut, like most other states does not have access to annual GQ data by TASRH, under our current data collection system, annual changes in Connecticut’s GQ population cannot be calculated accurately. Consequently group-quarters population figures are assumed to be constant at 2010 levels for this population group. Connecticut DPH is currently working to improve GQ data collection so that at least the largest facilities will provide more detailed annual data.

School enrollment, Medicare, and DMV datasets do not have full coverage, so, we still need to use a ratio adjustment (derived from regression models) for the estimates to fully cover the total population. The adjustment ratios vary for each model. Nevertheless, the population coverage for each of these datasets is very high, so we expect these data will allow us to make accurate town-level estimates.

References:

Bryan, T. (2004). Population estimates. The methods and materials of demography, 2, 523-560.

Baker, J., Alcántara , A., Ruan, X., Watkins, K., & Vasan, S. (2013). A comparative evaluation of error and bias in census tract-level age/sex-specific population estimates: Component I (Net-migration) vs Component III (Hamilton–Perry). Population Research and Policy Review, 32(6), 919-942.

Baker, J., Alcántara, A., Ruan, X., Watkins, K., & Vasan, S. (2014). Spatial weighting improves accuracy in small-area demographic forecasts of urban census tract populations. Journal of Population Research, 31(4), 345-359.

Cai, Q. (2007). New techniques in small area population estimates by demographic characteristics. Population Research and Policy Review, 26(2), 203-218.

Citro, C. F., & Kalton, G. (Eds.). (2007). Using the American community survey: Benefits and challenges. National Academies Press.

Courgeau, D., Nedellec, V., & Empereur-Bissonnet, P. (2000). Duration of Residence in the Same Dwelling. A Test of Measurement using Electricity Utility Company Records. Population: An English Selection, 12, 335-342.

Deng, C., & Wu, C. (2013). Improving small-area population estimation: An integrated geographic and demographic approach. Annals of the Association of American Geographers, 103(5), 1123-1141.

Friedman, J. H. (1991), “Multivariate Adaptive Regression Splines,” Annals of Statistics, 19, 1–67.

Goldberg, D. and T. R. Balakrishnan. 1960. A Partial Evaluation of Four Estimation Techniques. Paper presented at the Annual Meeting of the Social Statistics Section, American Statistical Association.

Hauer, M. E., Evans, J. M., & Alexander, C. R. (2015). Sea-level rise and sub-county population projections in coastal Georgia. Population and Environment, 37(1), 44-62.

Hepburn, G. C., Mayor, T. H., & Stafford, J. E. (1976). Estimation of market area population from residential electrical utility data. Journal of Marketing Research, 230-236.

Hoque, M. N. (2010). An evaluation of small area population estimates produced by component method II, ratio-correlation, and housing unit methods. The Open Demography Journal, 3, 18-30.

Ingram D.D., Parker J.D., Schenker N., Weed J.A., Hamilton B., Arias E., Madans J.H. United States Census 2000 population with bridged race categories. National Center for Health Statistics. Vital Health Stat 2(135). 2003. Retrieved from http://www.cdc.gov/nchs/data/series/sr_02/sr02_135.pdf

Kuhfeld , Warren F. and Cai , Weijie, 2013. Introducing the New ADAPTIVEREG Procedure for Adaptive Regression. Retrieved from https://support.sas.com/resources/papers/proceedings13/457-2013.pdf

Rayer, S., Smith, S. S., Doty, R., & Roulston-Doty, S. (2015). Households and Average Household Size in Florida: April 1, 2015. Florida Population Studies, 49, 1-6.

Smith, S. K., & Cody, S. (2013). Making the housing unit method work: An evaluation of 2010 population estimates in Florida. Population Research and Policy Review, 32(2), 221-242.

Spielman, S. E., Folch, D., & Nagle, N. (2014). Patterns and causes of uncertainty in the American Community Survey. Applied Geography, 46, 147-157.

Spielman, S. E., & Folch, D. C. (2015). Reducing uncertainty in the American Community Survey through data-driven regionalization. PloS one, 10(2), e0115626.

Starsinic, D. E., & Zitter, M. (1968). Accuracy of the housing unit method in preparing population estimates for cities. Demography, 5(1), 475-484.

Swanson, D. A., Carlson, J., & Roe, L. (1992). A variation of the housing unit method for estimating the population of small, rural areas: A case study of the local expert procedure. Survey Methodology, 18(1).

Swanson, D. A., Carlson, J., Roe, L., & Williams, C. (1995). Estimating the Population of Rural Communities by Age and Gender: A Case Study of the Effectiveness of the Local Expert Procedure. Small Town, 25(6).

Swanson, D. A., & Pol, L. G. (2010). Applied Demography: Its Business and Public Sector Components. Demography-Volume I, 321.

Swanson, D. A., Schlottmann, A., & Schmidt, B. (2010). Forecasting the population of census tracts by age and sex: An example of the Hamilton–Perry method in action. Population Research and Policy Review, 29(1), 47-63.

Swanson, D. A., & Tayman, J. (2012). Subnational population estimates (Vol. 31). Springer Science & Business Media.

Tang, Z. (1999). Domestic Migration Estimation in Subcounty Areas: A Case Study in Massachusetts. Population Estimates Methods Conference, Washington, D.C.: U.S, 1999.

U.S. Census Bureau. (2009). A Compass for Understanding and Using American Community Survey Data: What Researchers Need to Know.

U.S. Census Bureau (2012). United States Summary: 2010, Population and Housing Unit Counts (Report). p. 41.

U.S. Census Bureau. Methodology for the Subcounty Total Resident Population Estimates (Vintage 2014): April 1, 2010 to July 1, 2014. Retrieved from http://www.census.gov/popest/methodology/index.html

U.S. Census Bureau (2015). Annual Estimates of the Resident Population: April 1, 2010 to July 1, 2015.

Zhang, Heping and Singer, Burton H. (2010) Recursive Partitioning and Applications, 2nd edition. Springer, ISBN 978-1-4419-6823-4.

Appendices:

Appendix A: Illustrations of predictive model fit – Comparison of selected estimate with the original 2010 population values:

- Model for ages 20-64: Model estimates and 2010 Population counts are plotted by town for each age-sex subgroup in this model. Points are only plotted where the 2010 population was greater than zero.

Appendix B: Illustrations of predictive model fit – Comparison of selected estimate with the original 2010 population values:

- Model for ages 65+: Model estimates and 2010 Population counts are plotted by town for each age-sex-race/ethnicity subgroup in this model. Points are only plotted where the 2010 population was greater than zero.

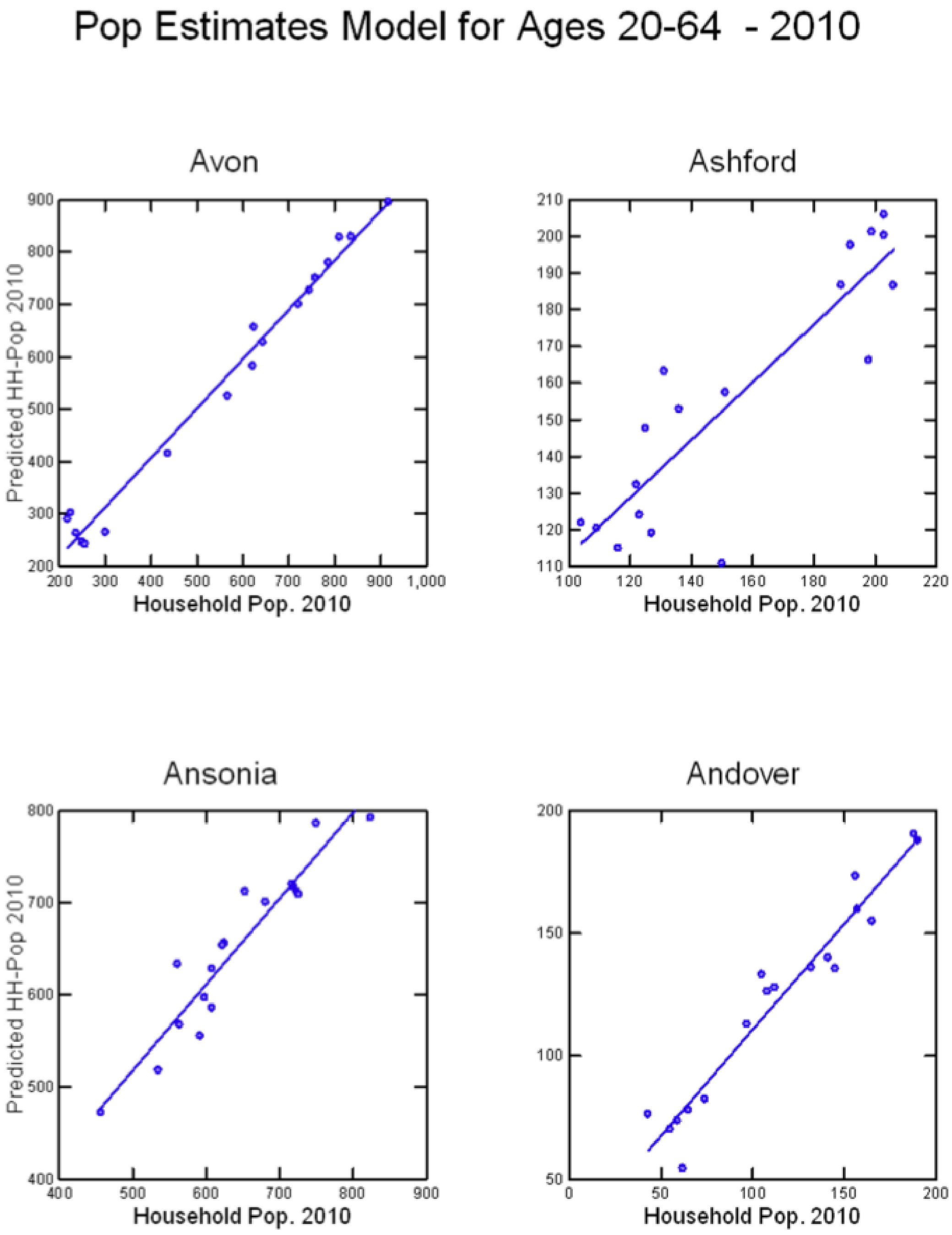
Appendix C: Two Descriptions of the Annual Population Estimation Process:

- Formula -1: The intuitive description of the estimation process used in the body of this report
- Formula-2: An alternate presentation of the formula-1

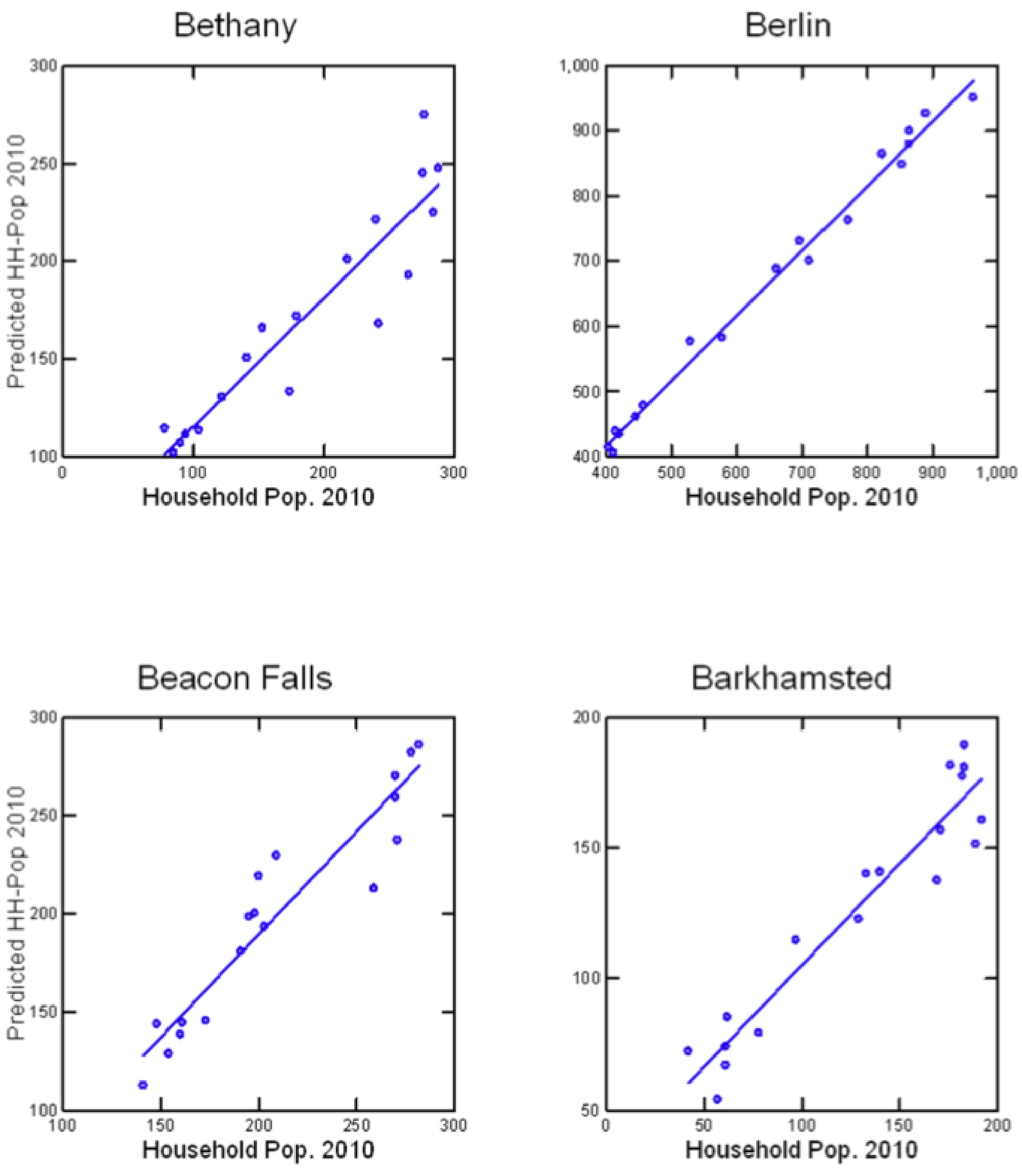
Appendix D: SAS Programs for Annual Prediction by Model Type:

- SAS models for future population estimates using new, annually updated input data include:
    - Model for 0-4 years of age
    - Model for 5-9 years of age
    - Model for 10-14 years of age
    - Model for 15-19 years of age
    - Model for 20-64 years of age
    - Model for 65-85 years of age

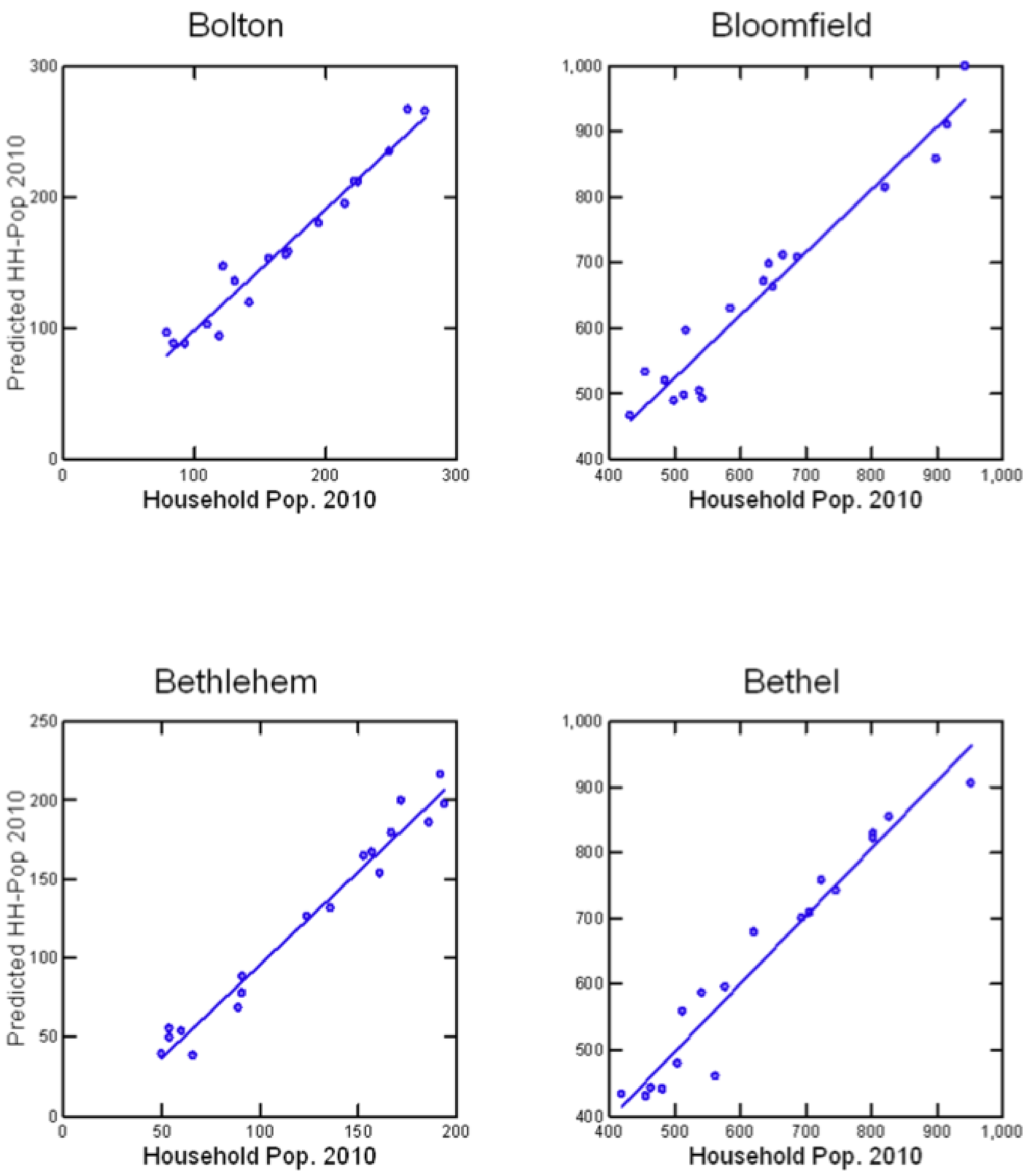
Appendix A: Population Estimates Model for Ages 20 to 64 – 2010 by Town



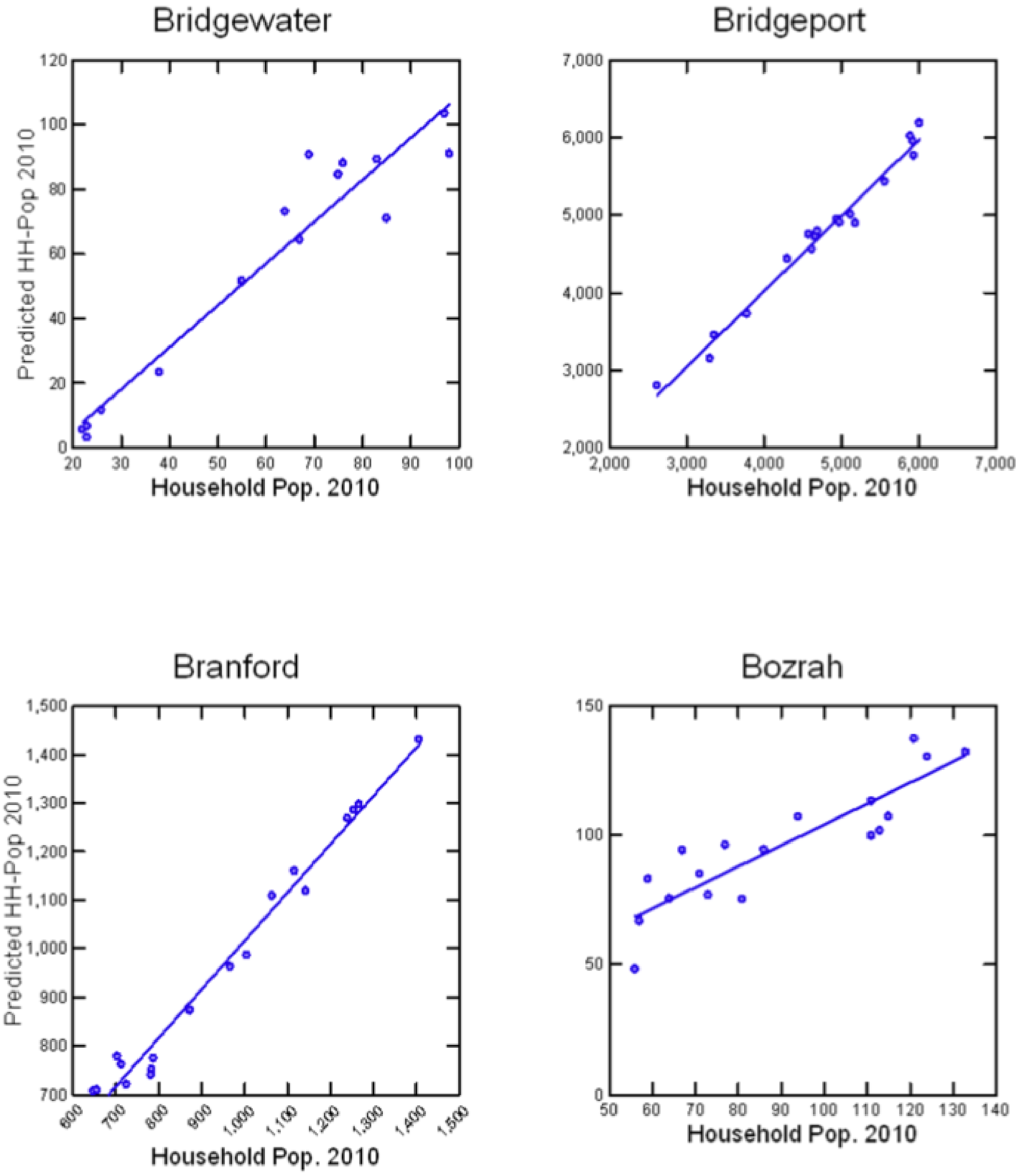
Pop Estimates Model for Ages 20-64 - 2010



Pop Estimates Model for Ages 20-64 - 2010



Pop Estimates Model for Ages 20-64 - 2010



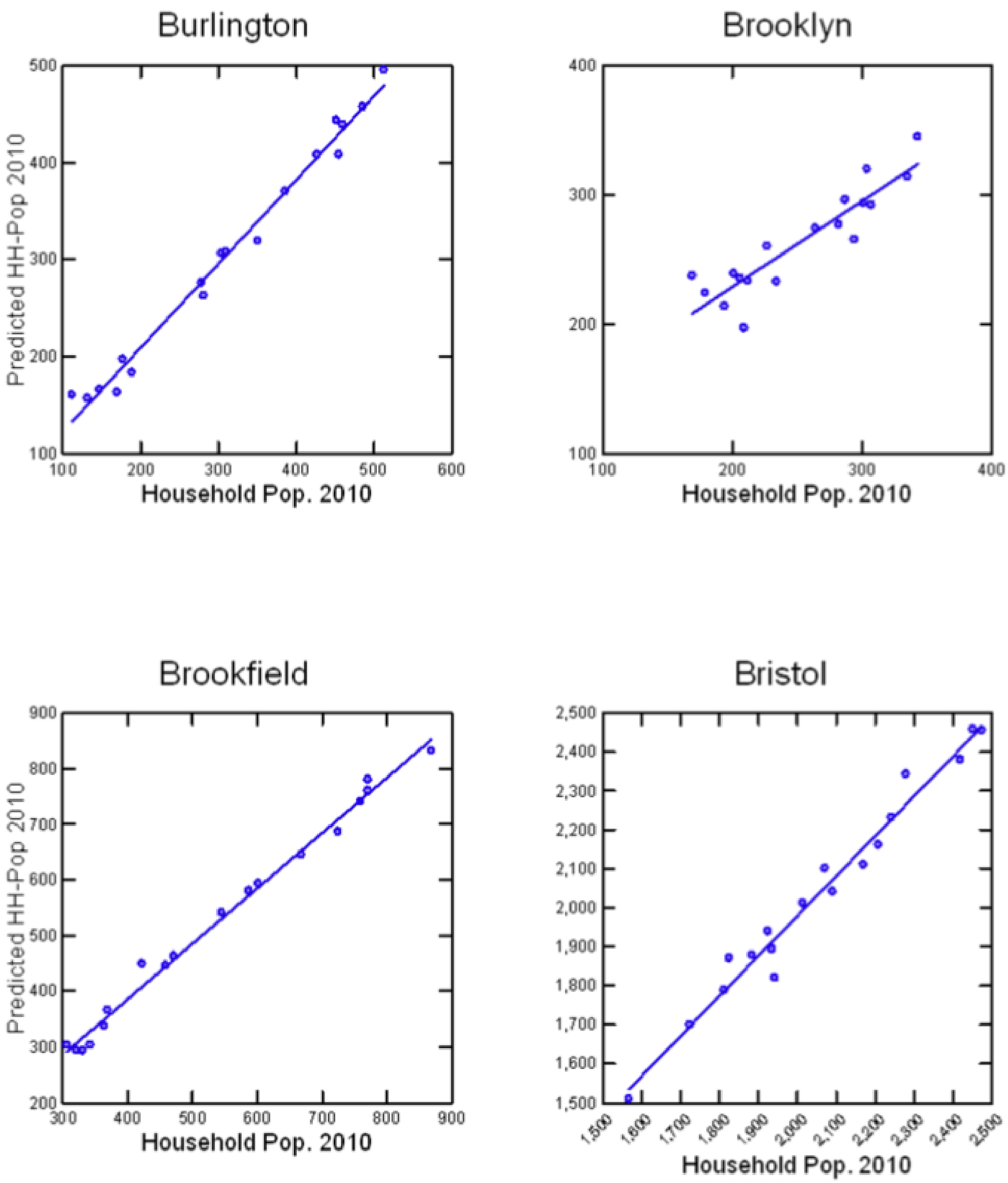
Bridgewater

Bridgeport

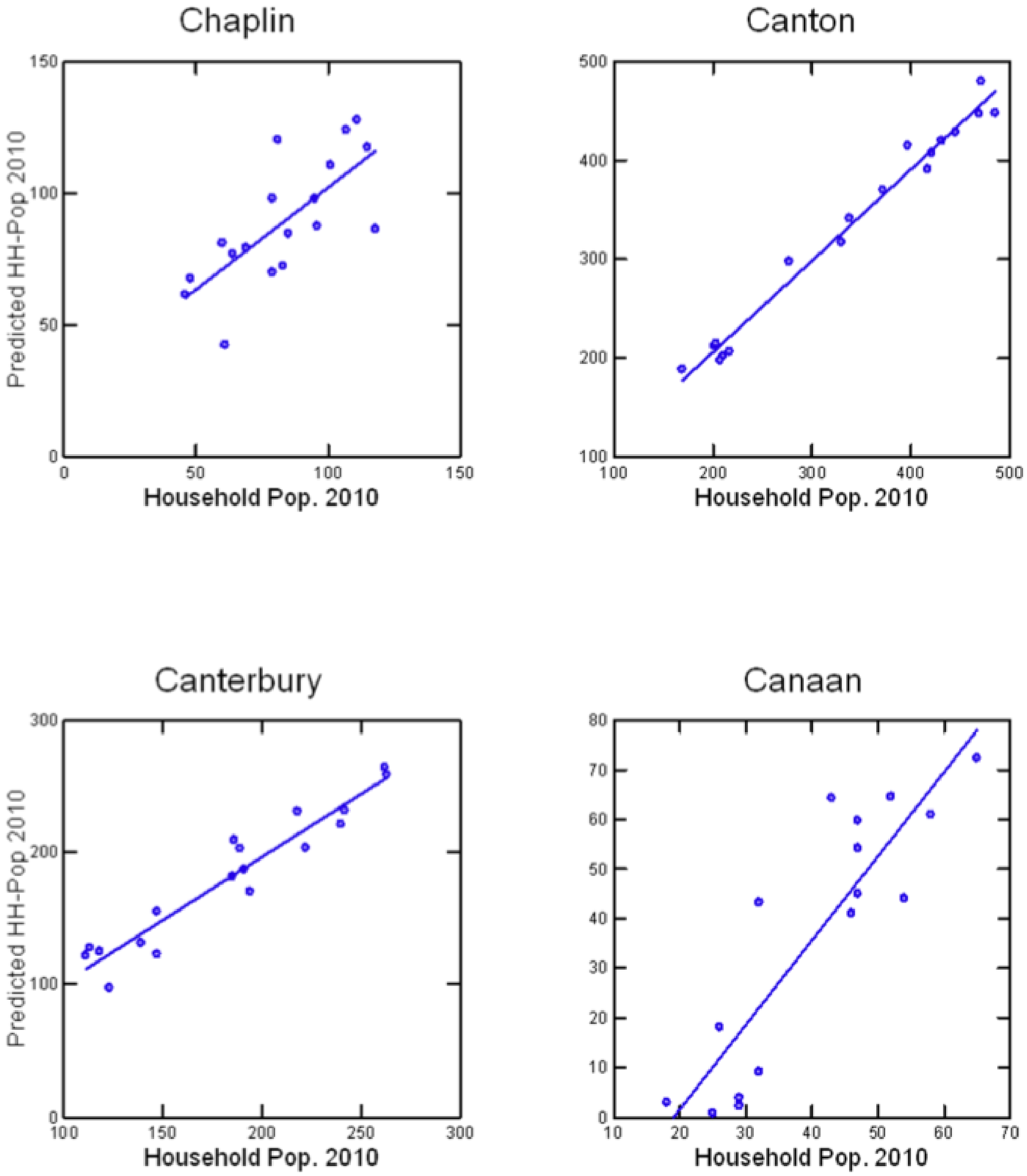
Branford

Bozrah

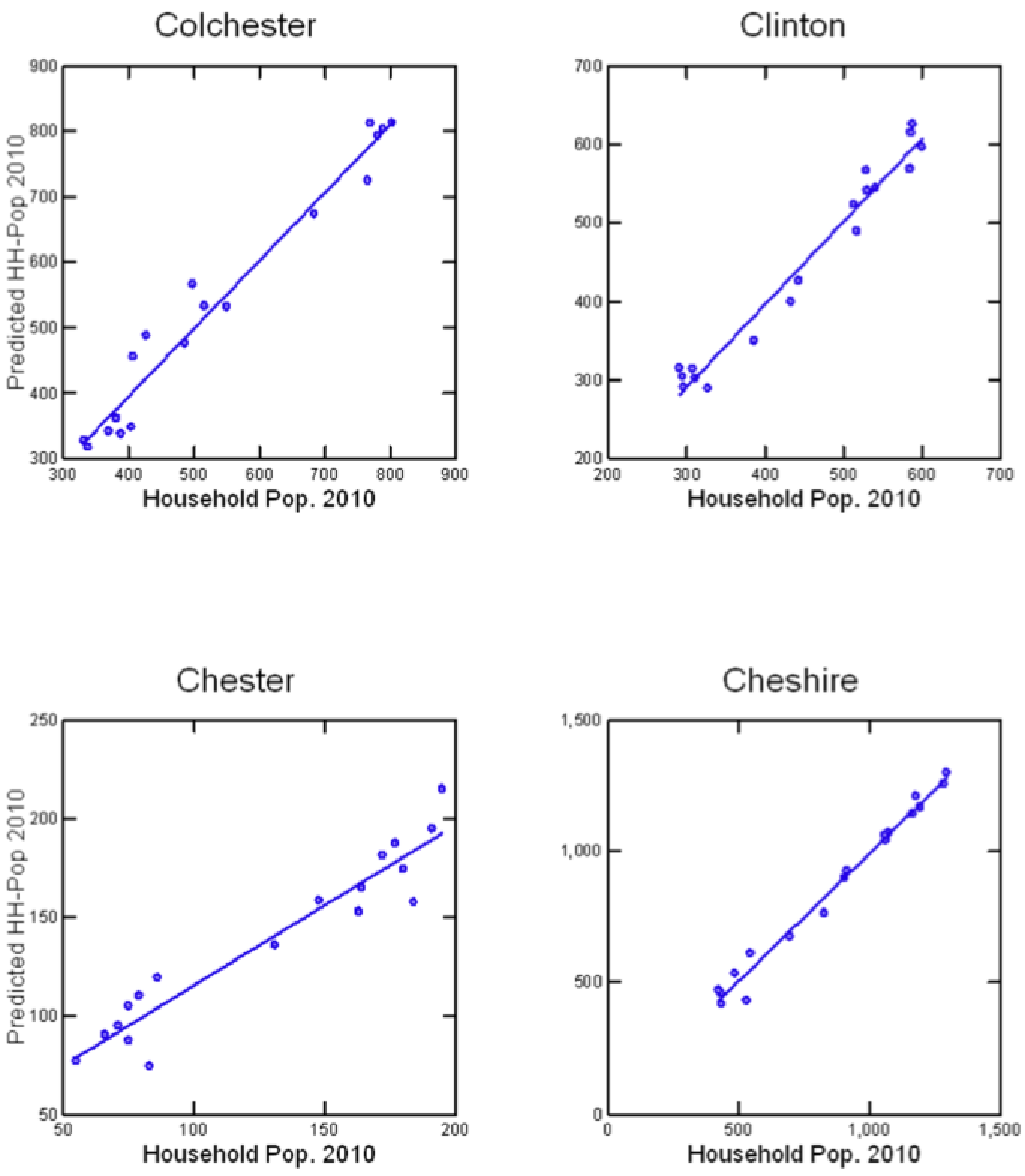
Pop Estimates Model for Ages 20-64 - 2010



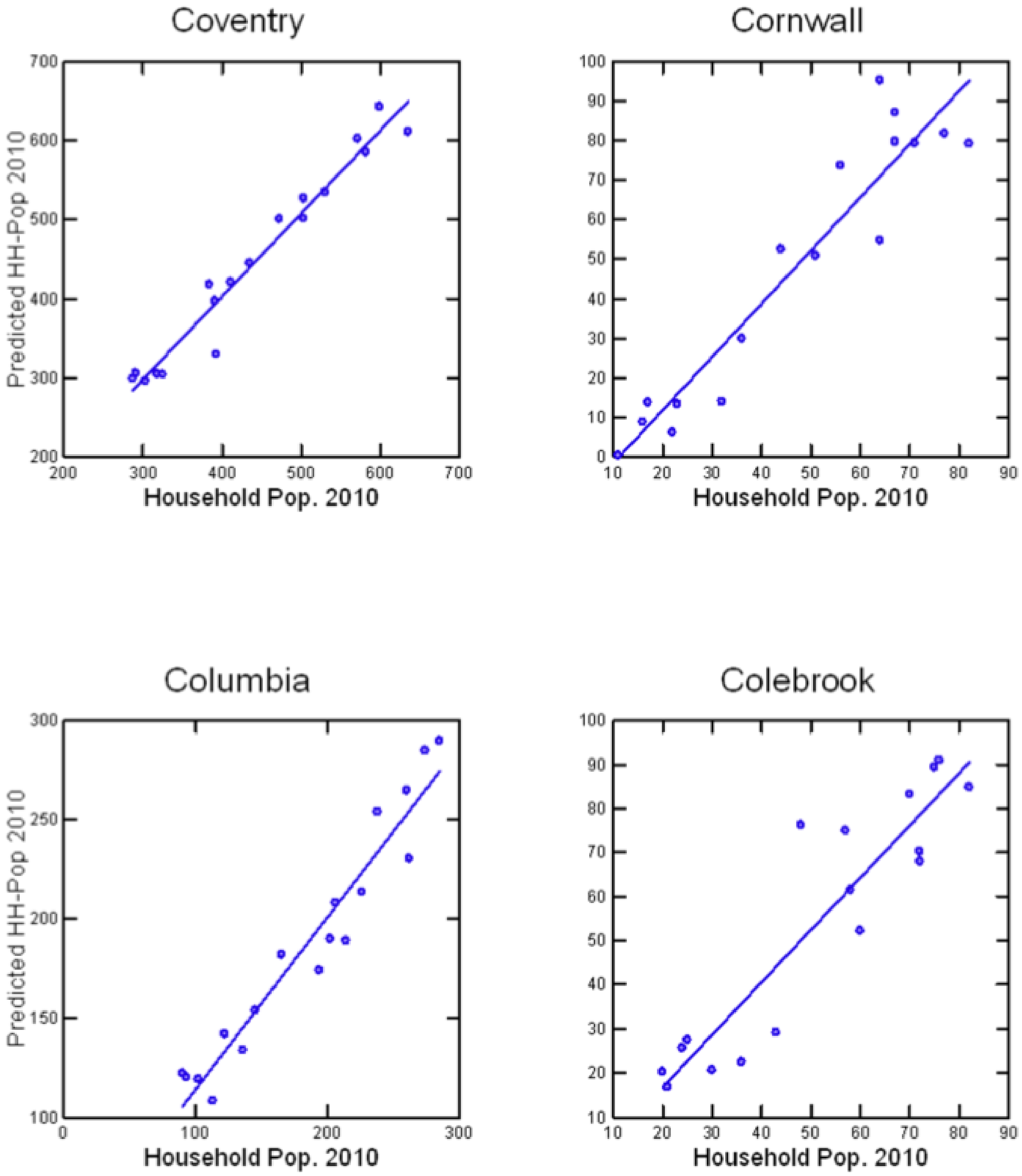
Pop Estimates Model for Ages 20-64 - 2010



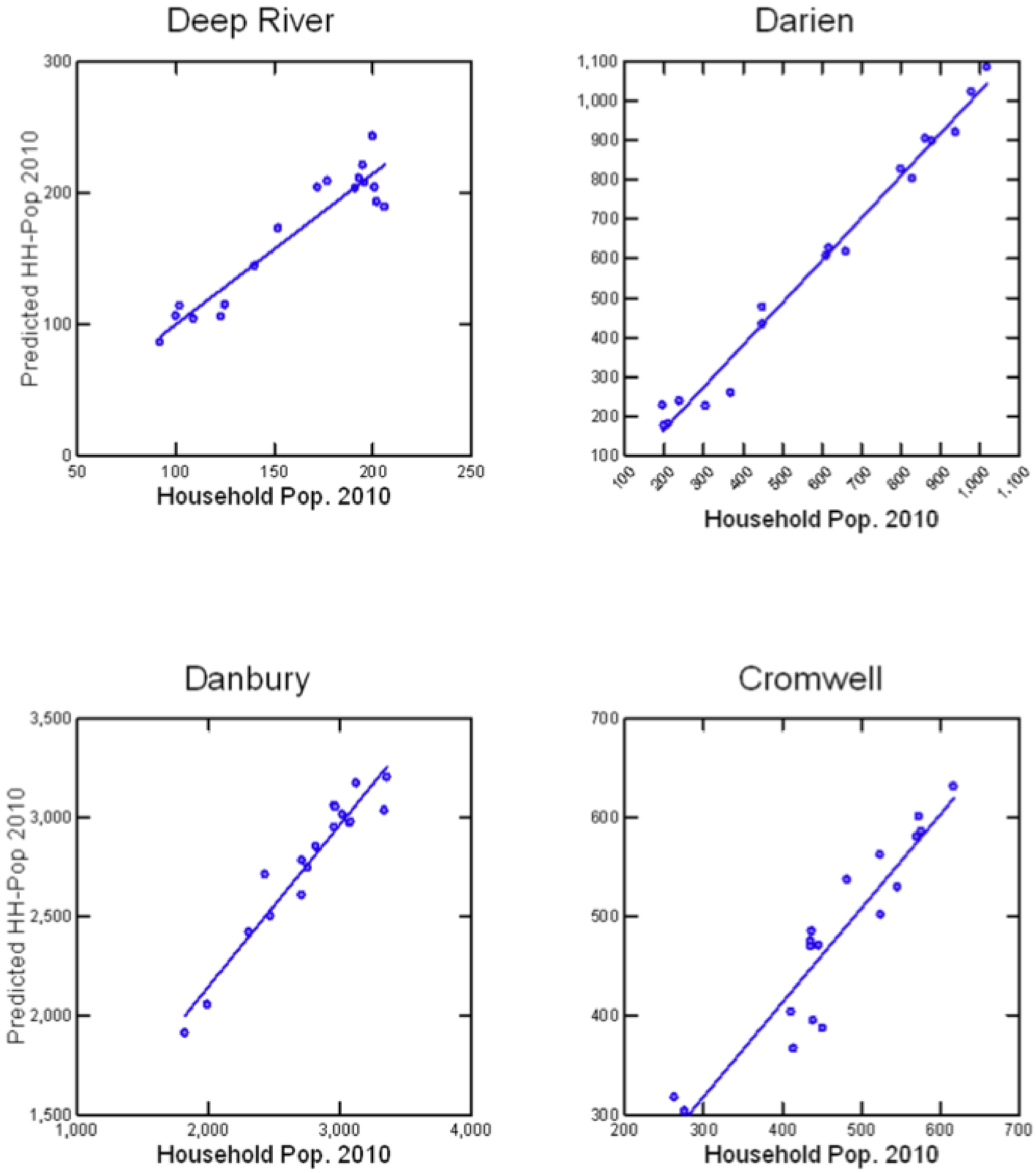
Pop Estimates Model for Ages 20-64 - 2010



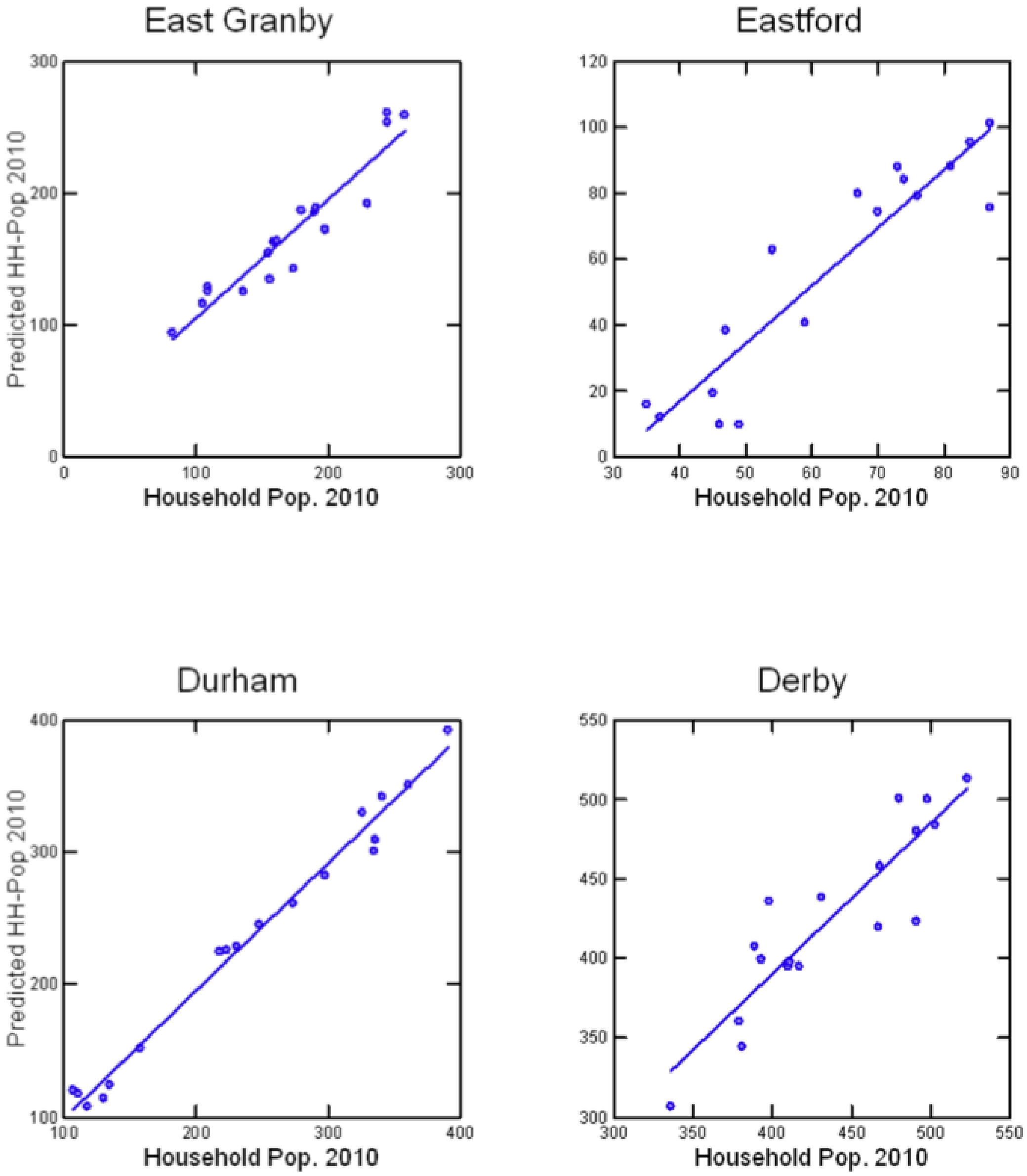
Pop Estimates Model for Ages 20-64 - 2010



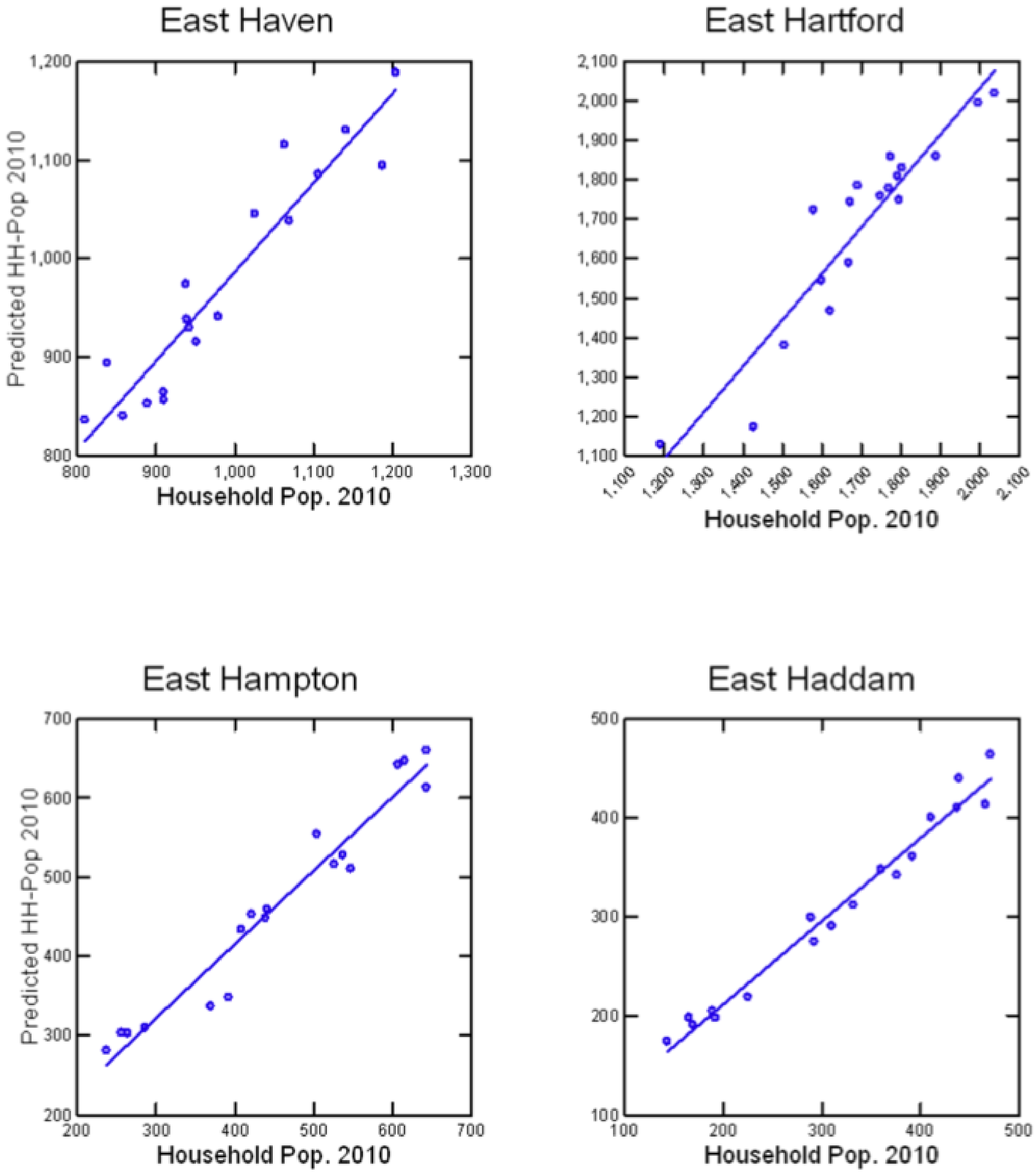
Pop Estimates Model for Ages 20-64 - 2010



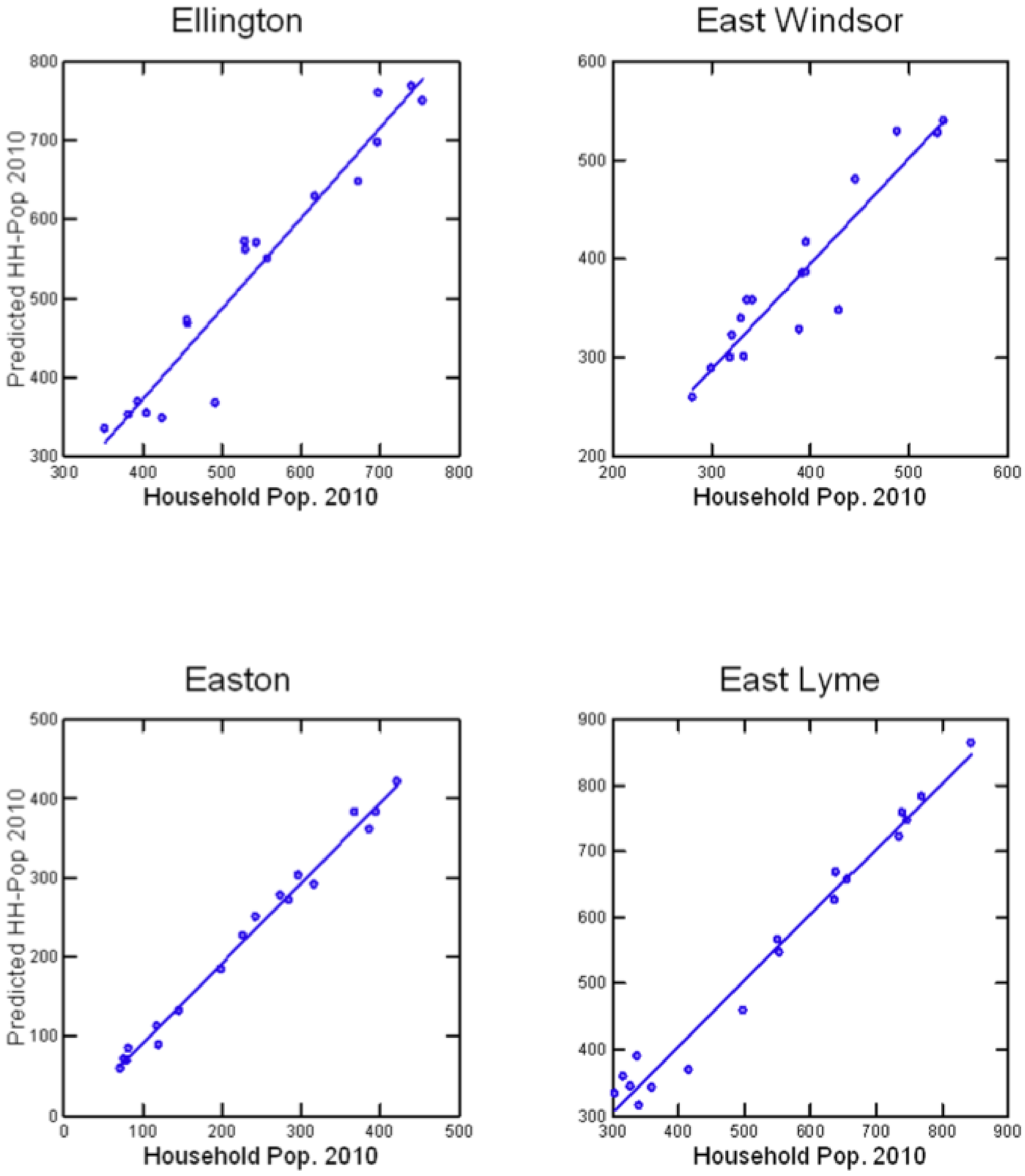
Pop Estimates Model for Ages 20-64 - 2010



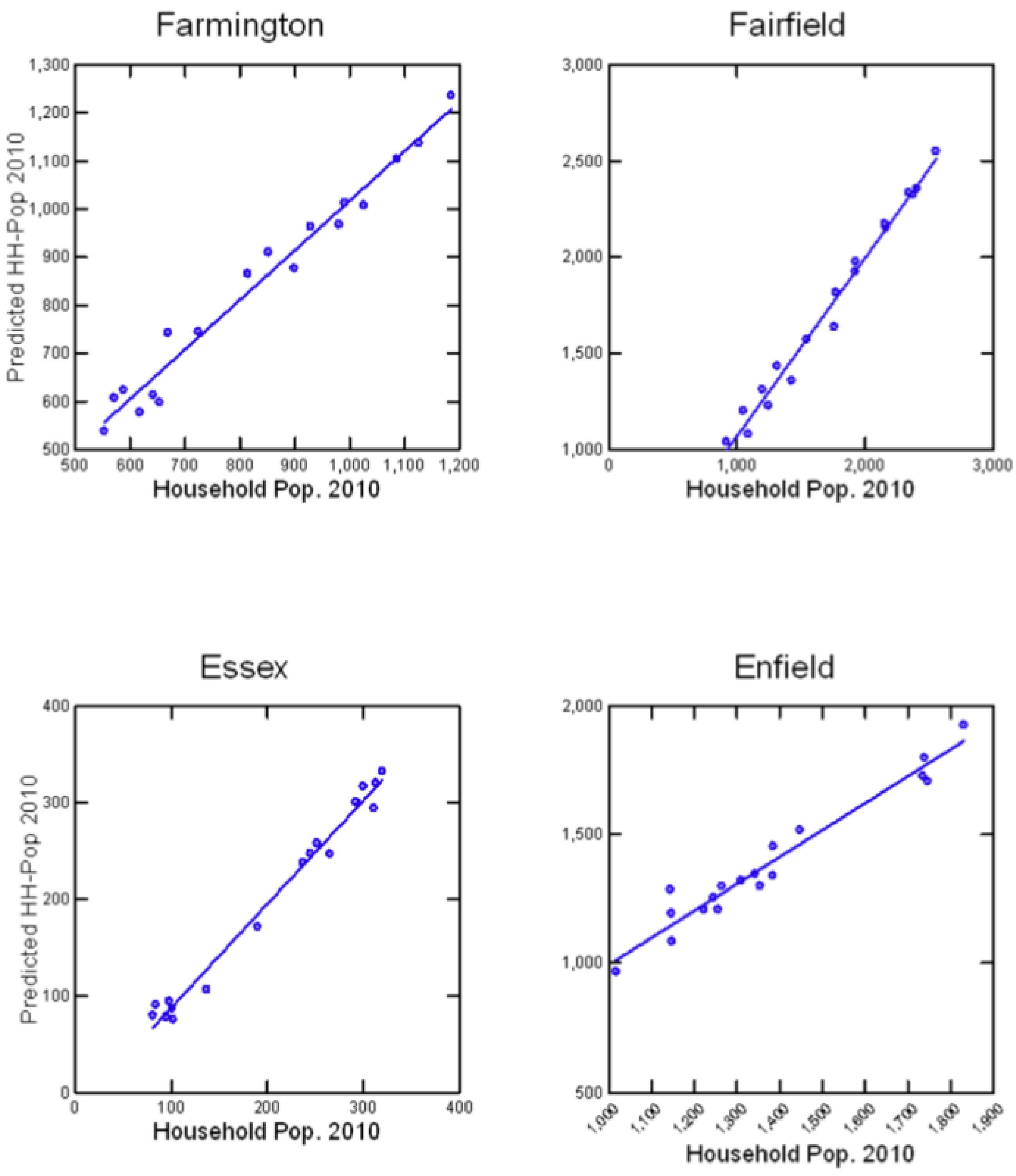
Pop Estimates Model for Ages 20-64 - 2010



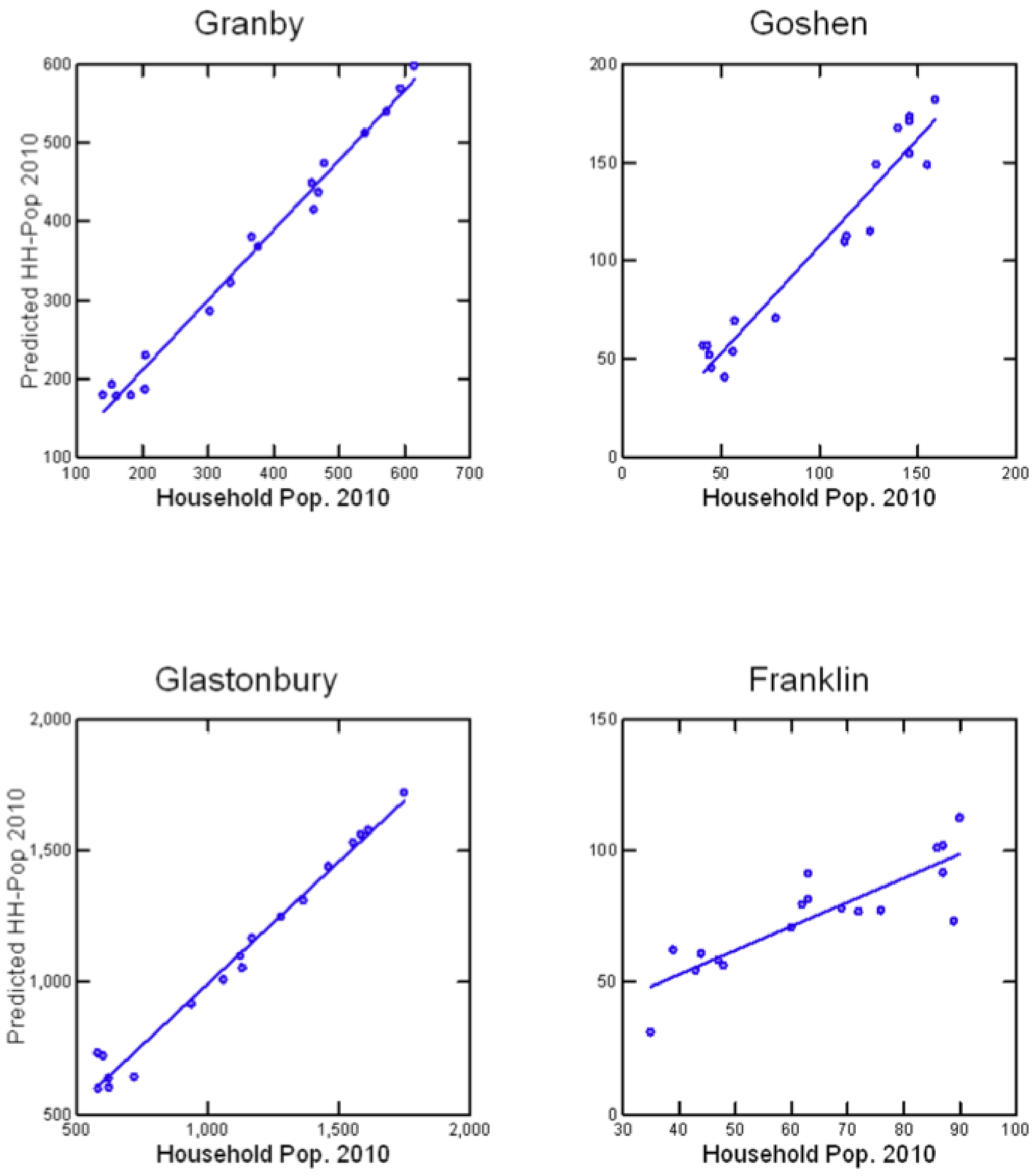
Pop Estimates Model for Ages 20-64 - 2010



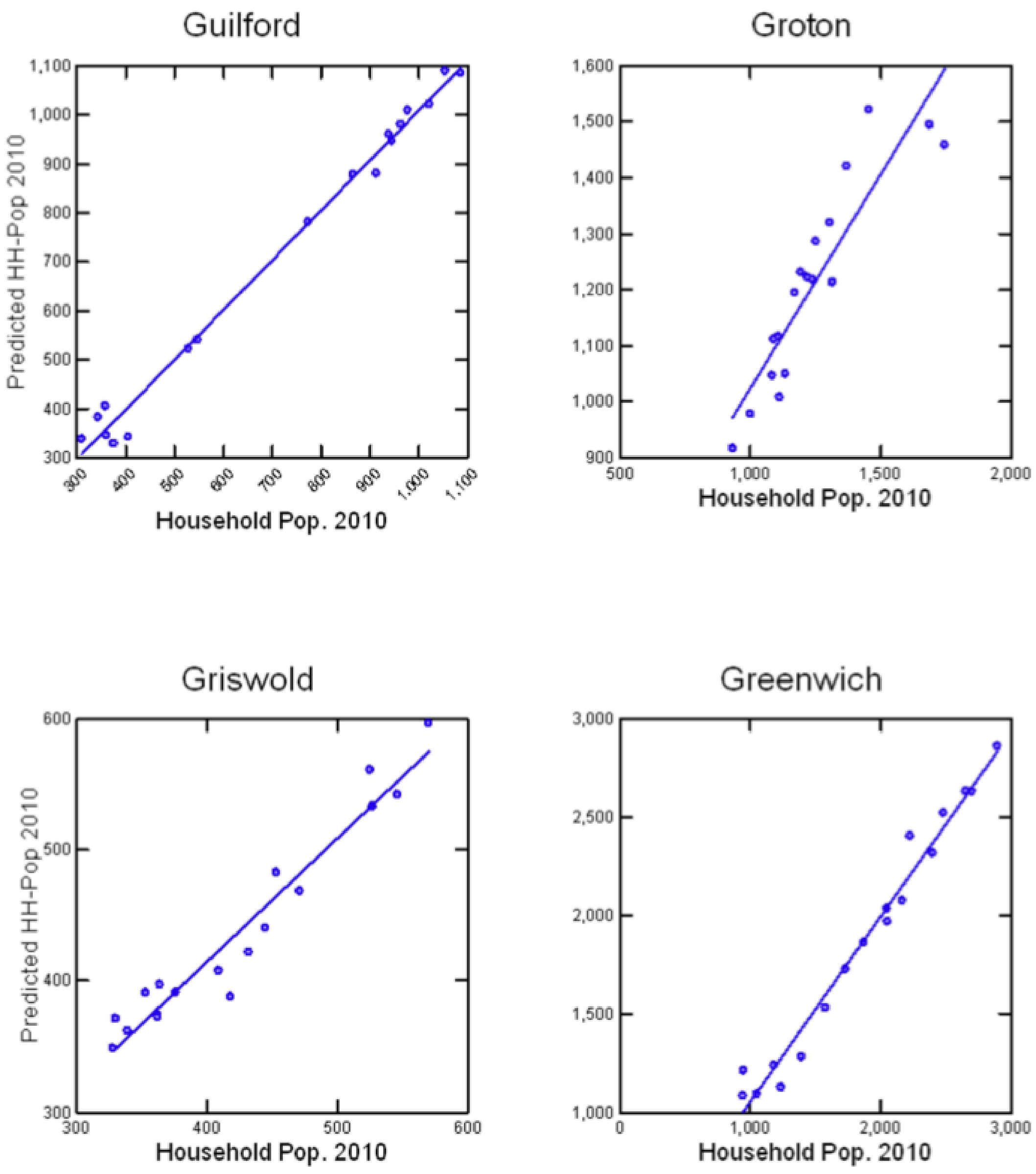
Pop Estimates Model for Ages 20-64 - 2010



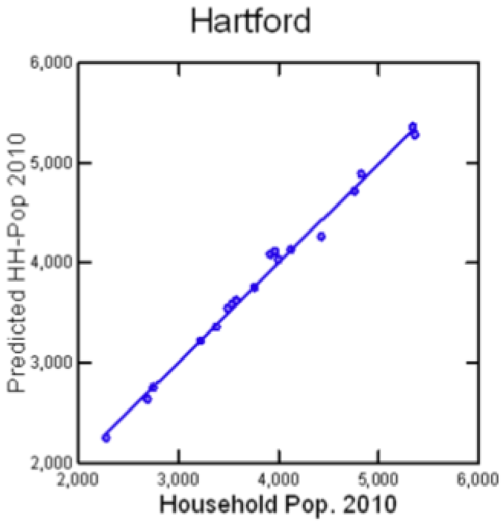
Pop Estimates Model for Ages 20-64 - 2010

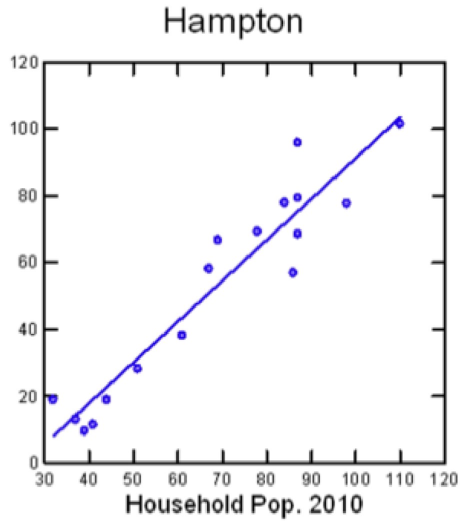


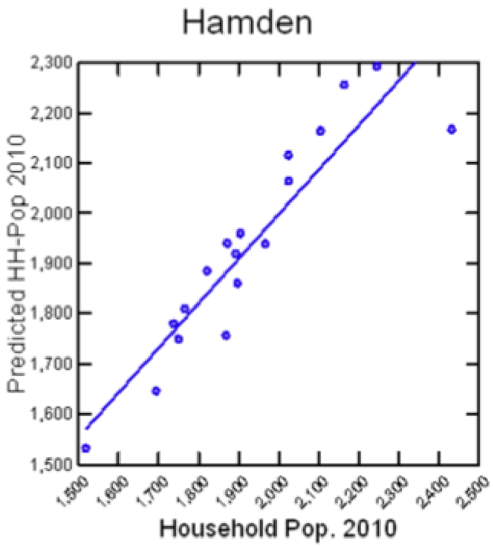
Pop Estimates Model for Ages 20-64 - 2010

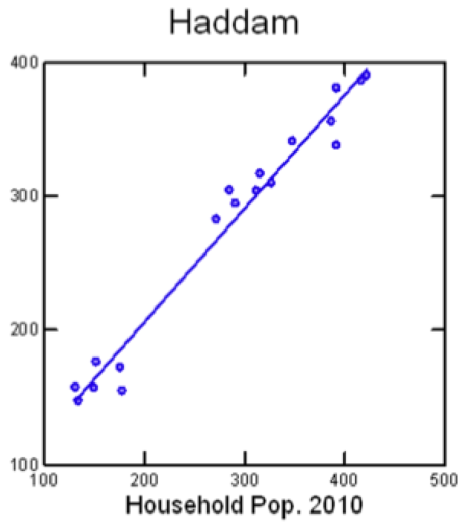


Pop Estimates Model for Ages 20-64 - 2010

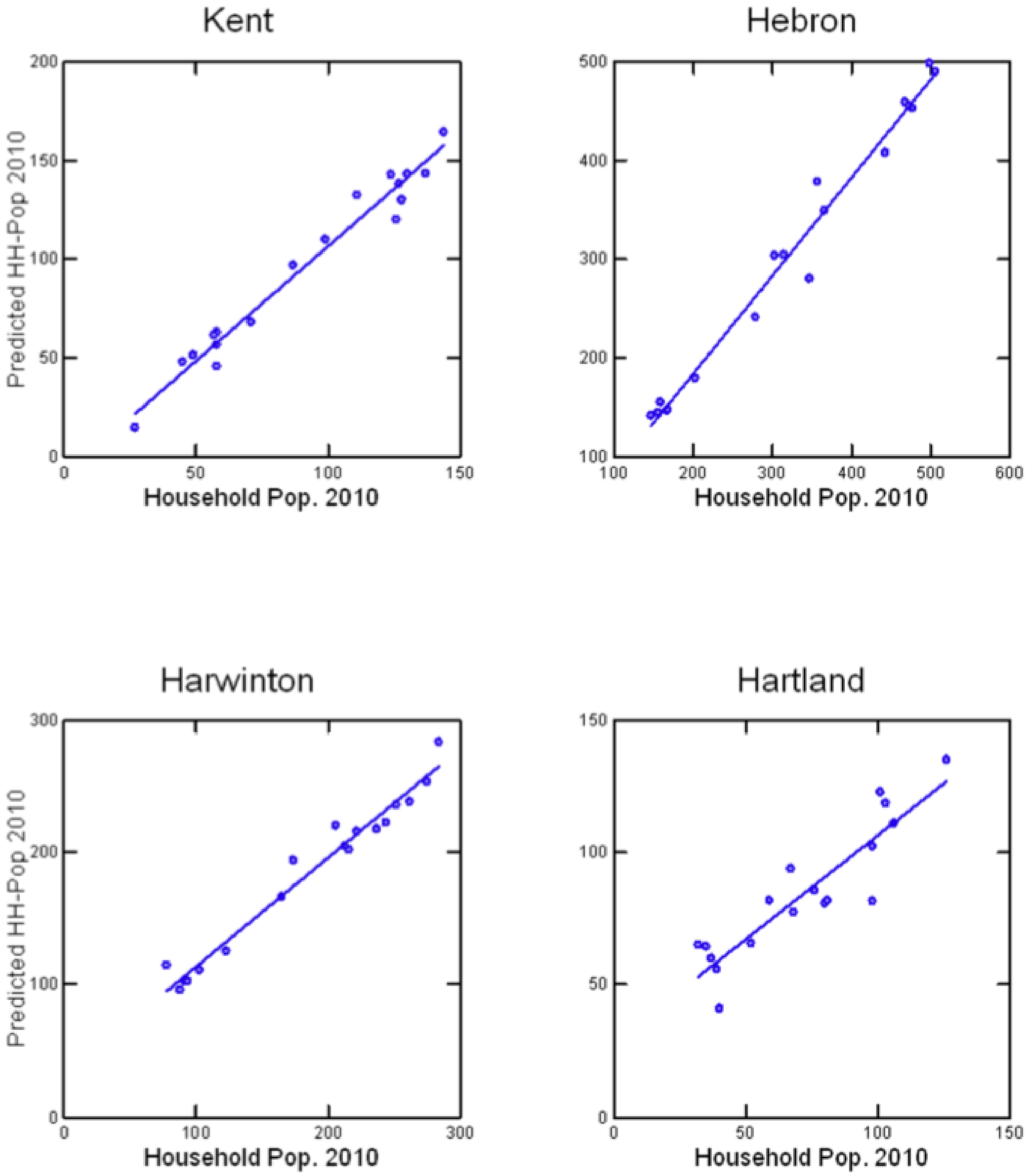




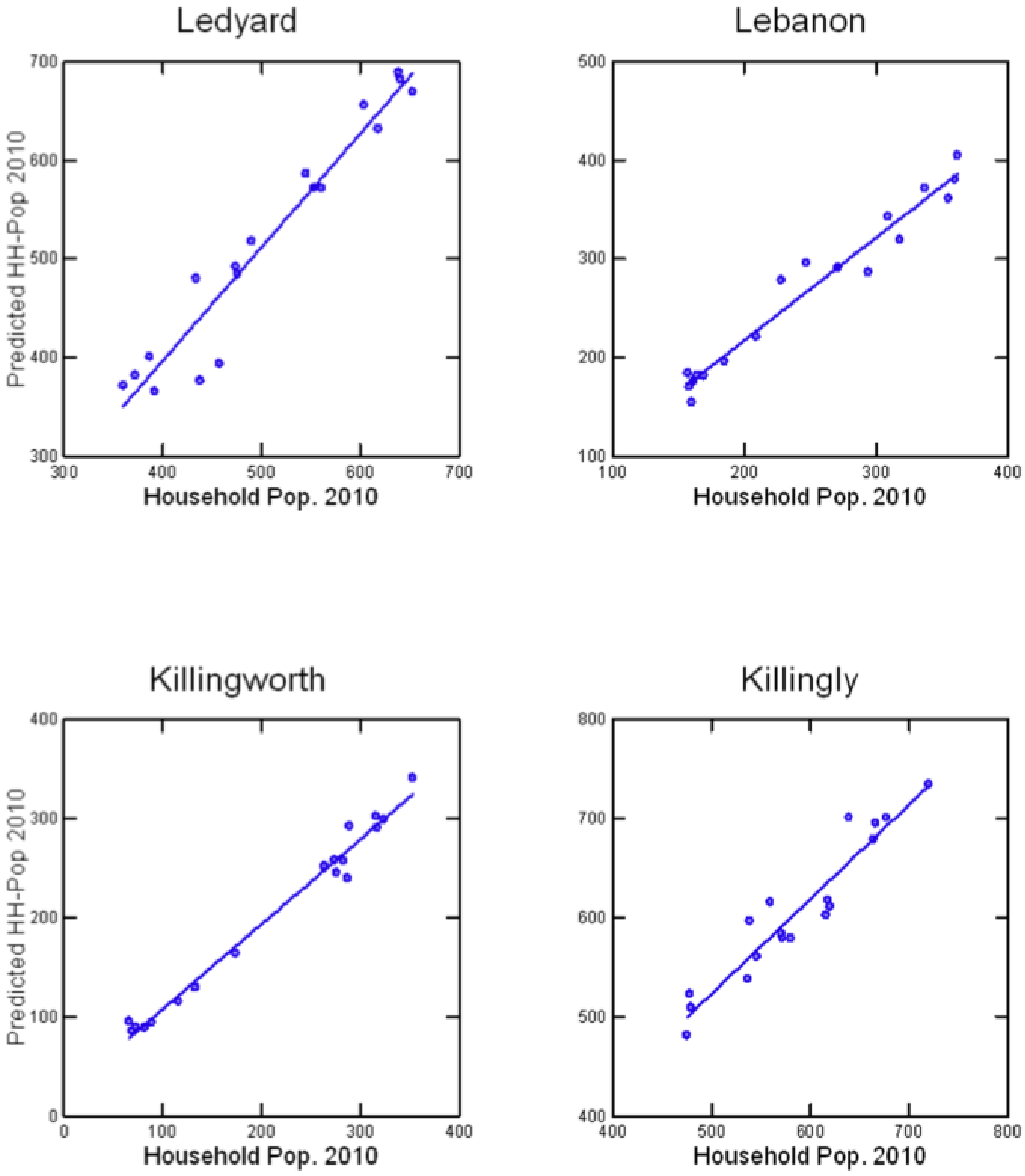




Pop Estimates Model for Ages 20-64 - 2010



Pop Estimates Model for Ages 20-64 - 2010



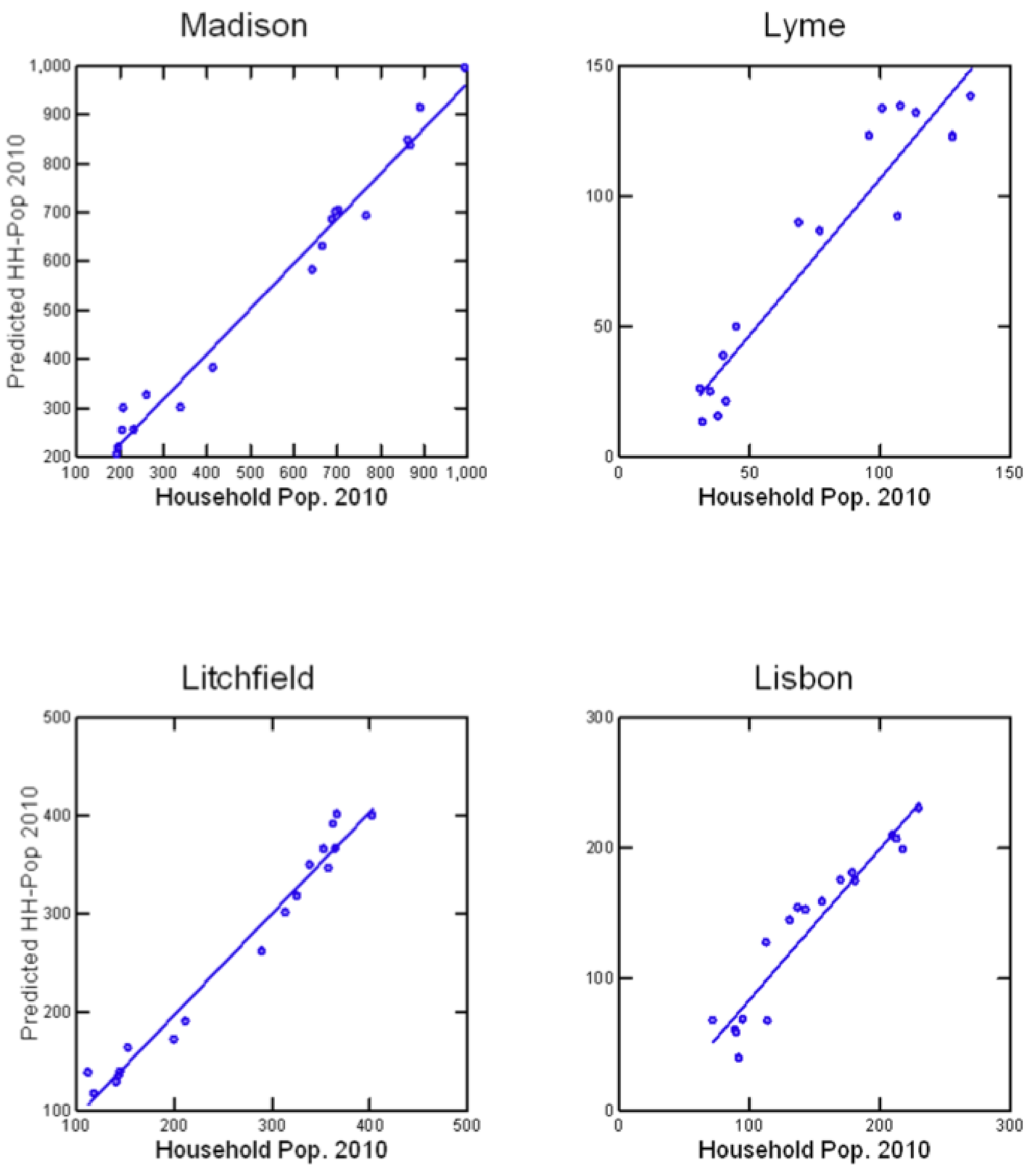
Ledyard

Lebanon

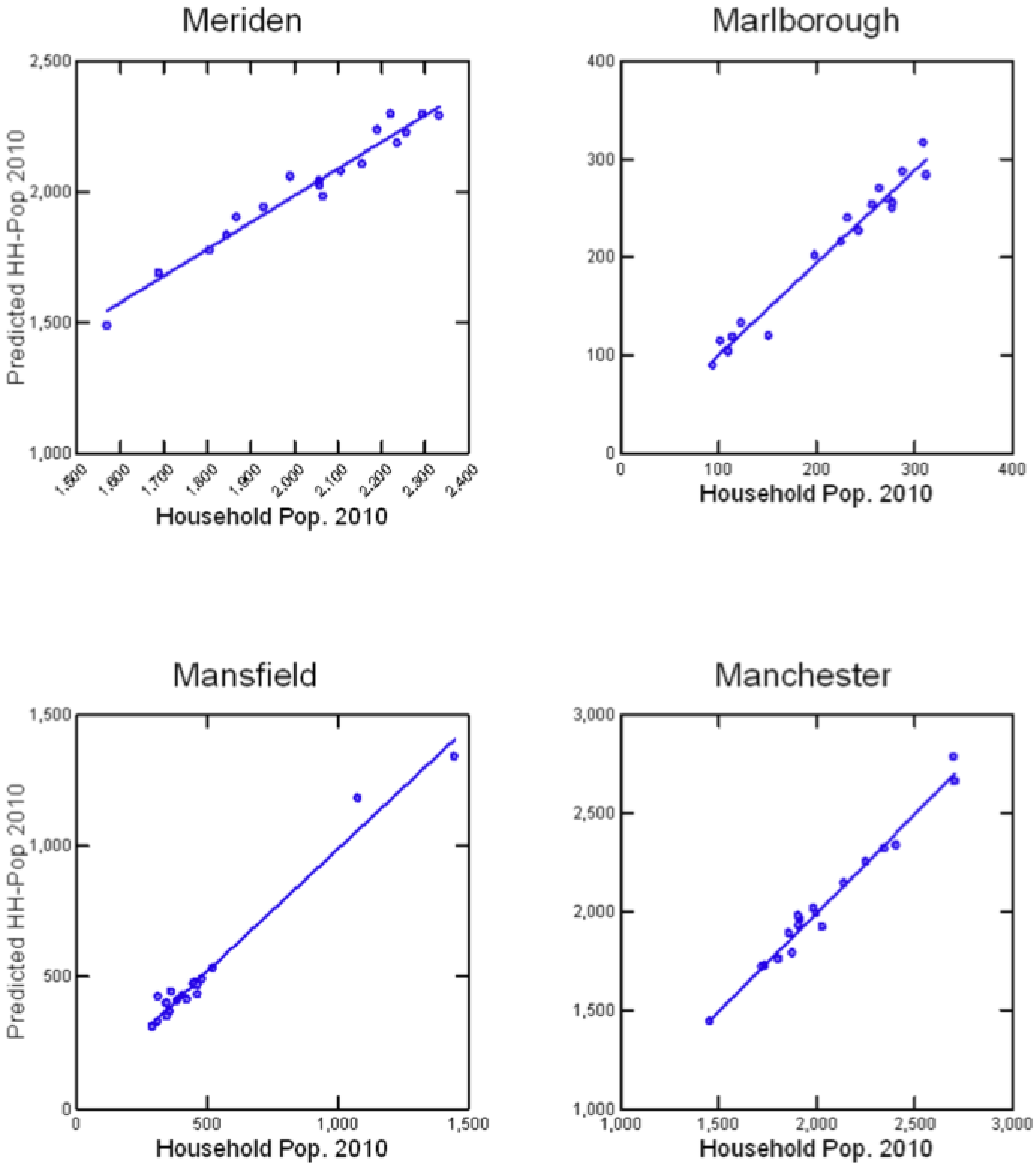
Killingworth

Killingly

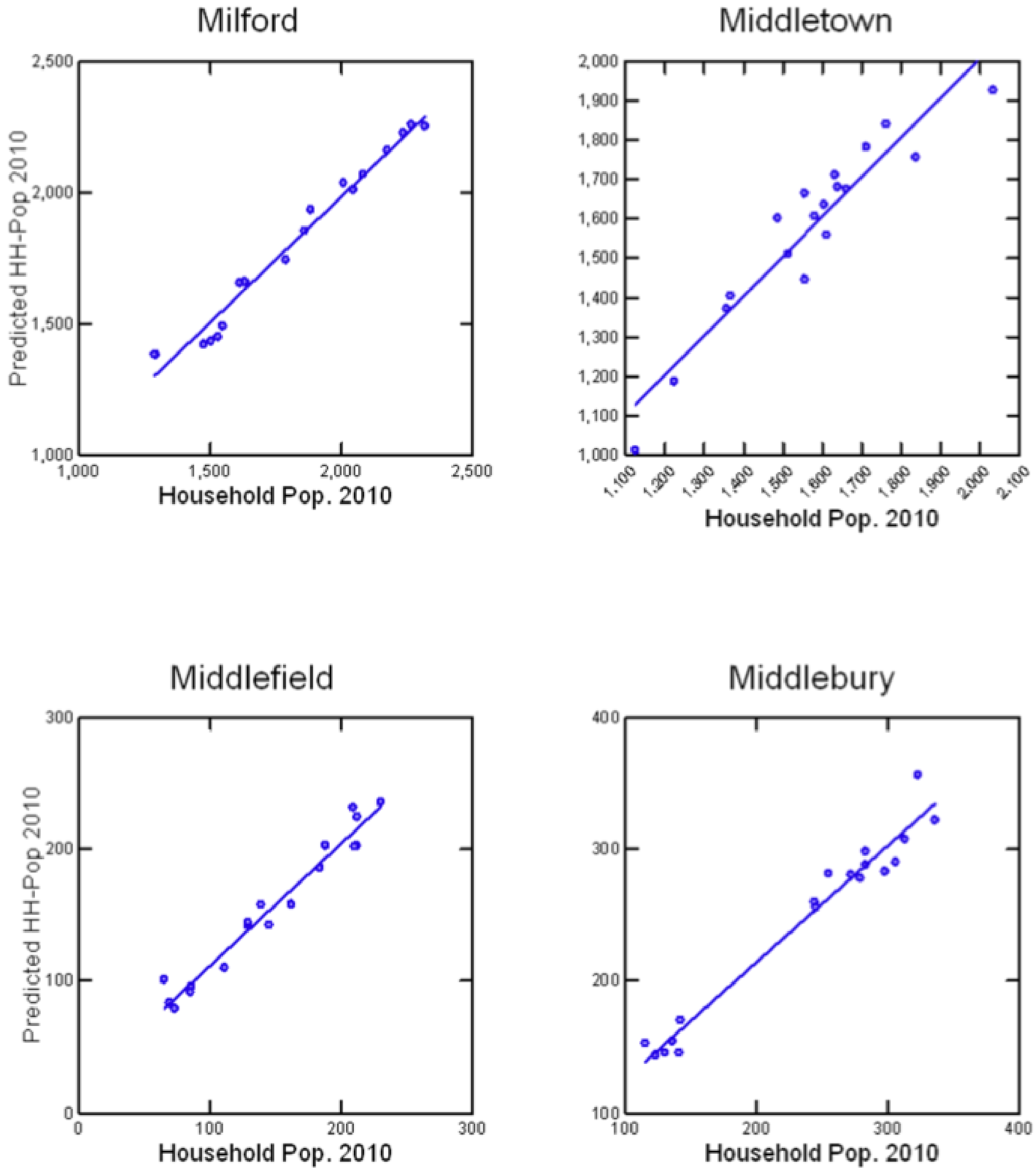
Pop Estimates Model for Ages 20-64 - 2010



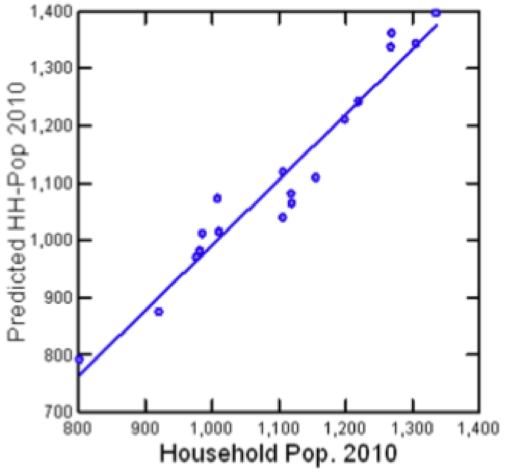
Pop Estimates Model for Ages 20-64 - 2010

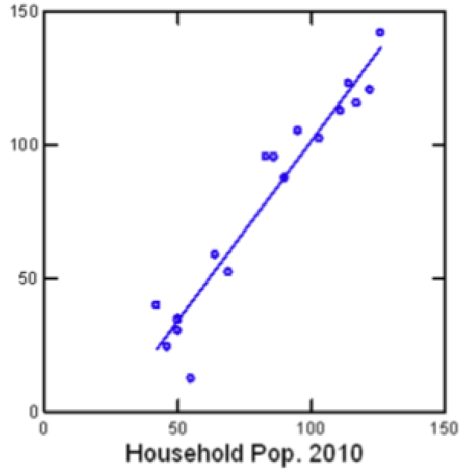


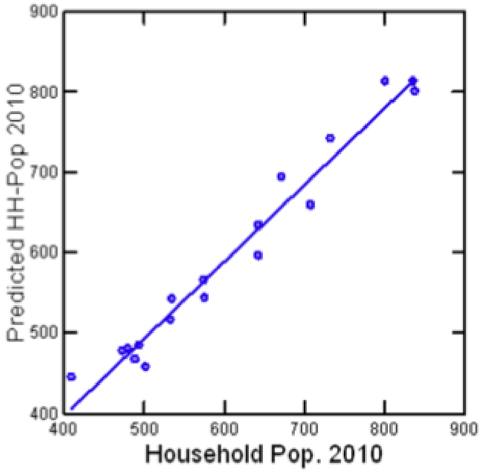
Pop Estimates Model for Ages 20-64 - 2010

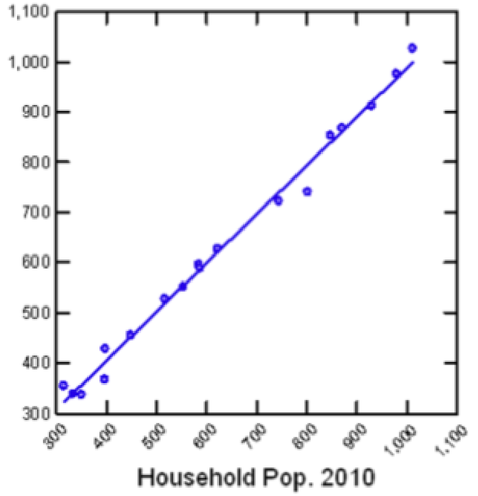


Pop Estimates Model for Ages 20-64 - 2010

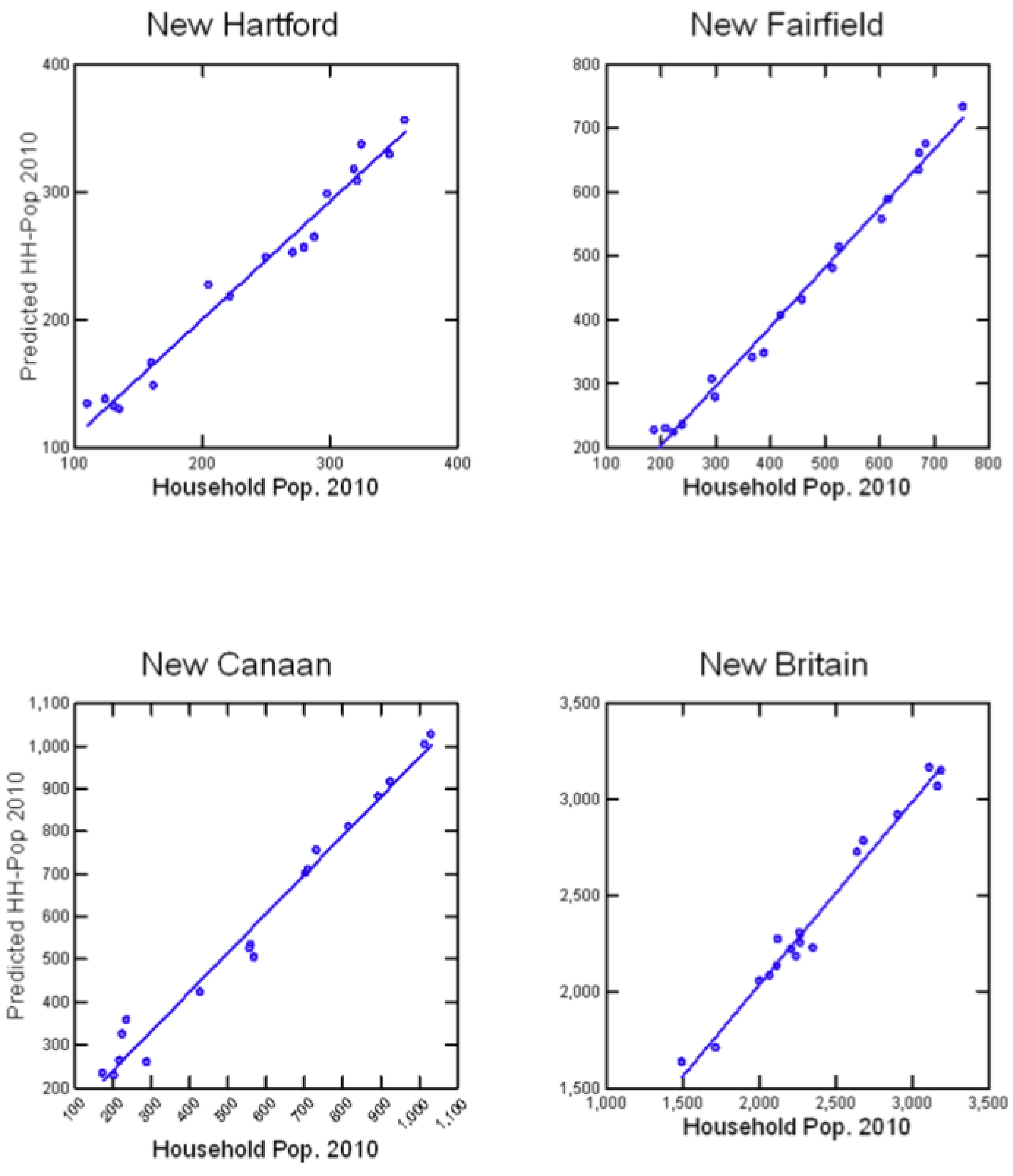

Naugatuck


Morris

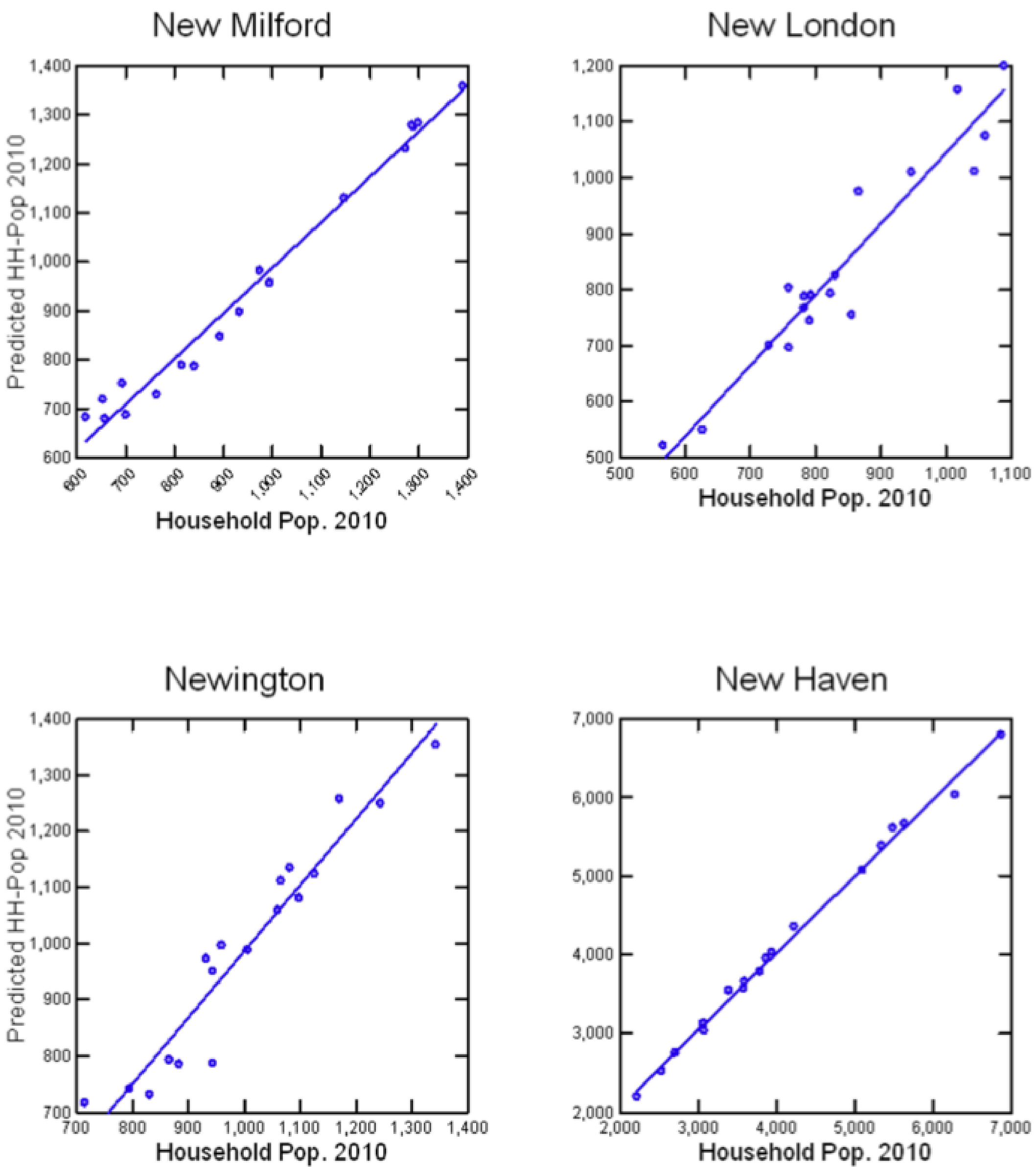

Montville


Monroe

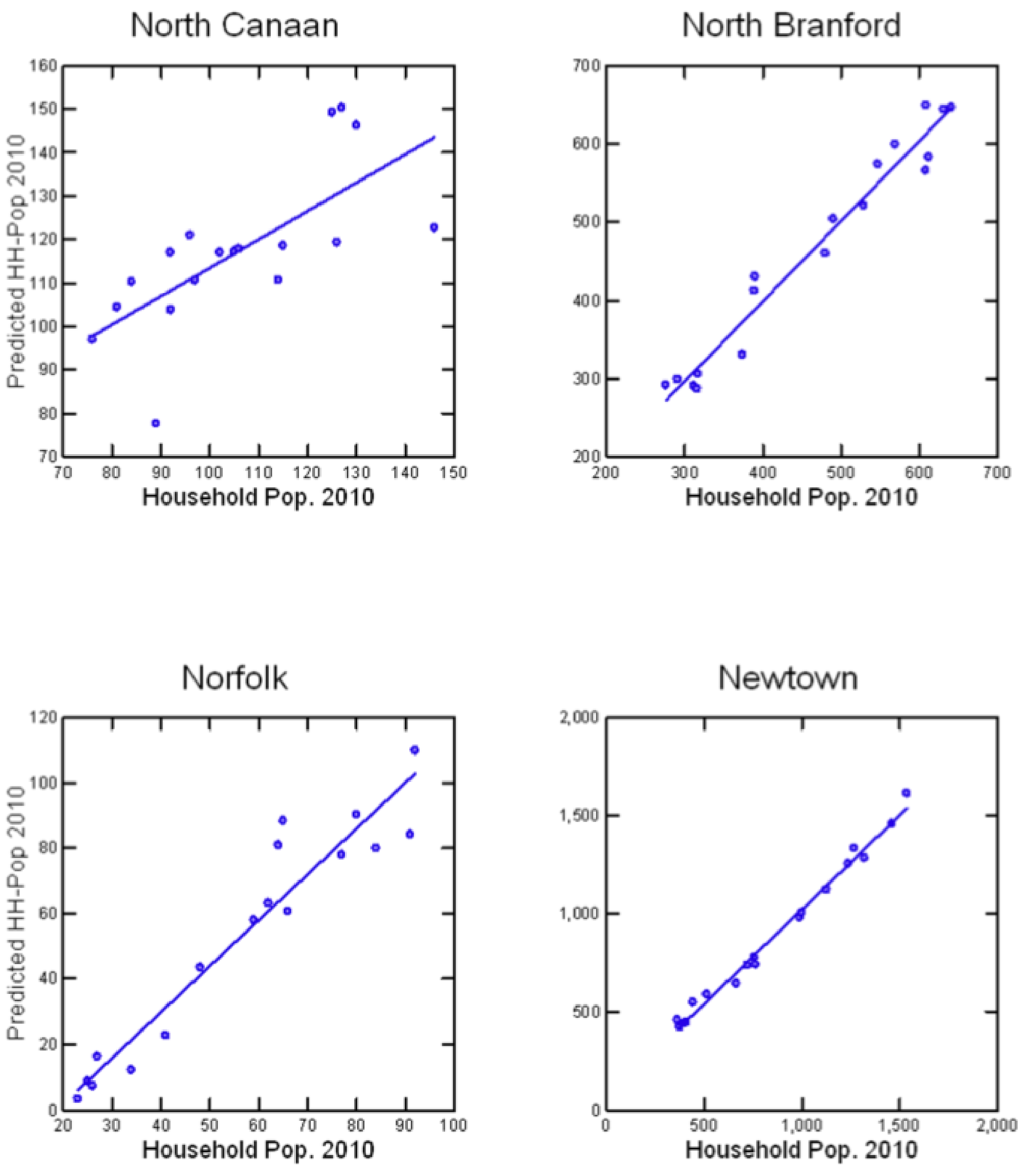
Pop Estimates Model for Ages 20-64 - 2010



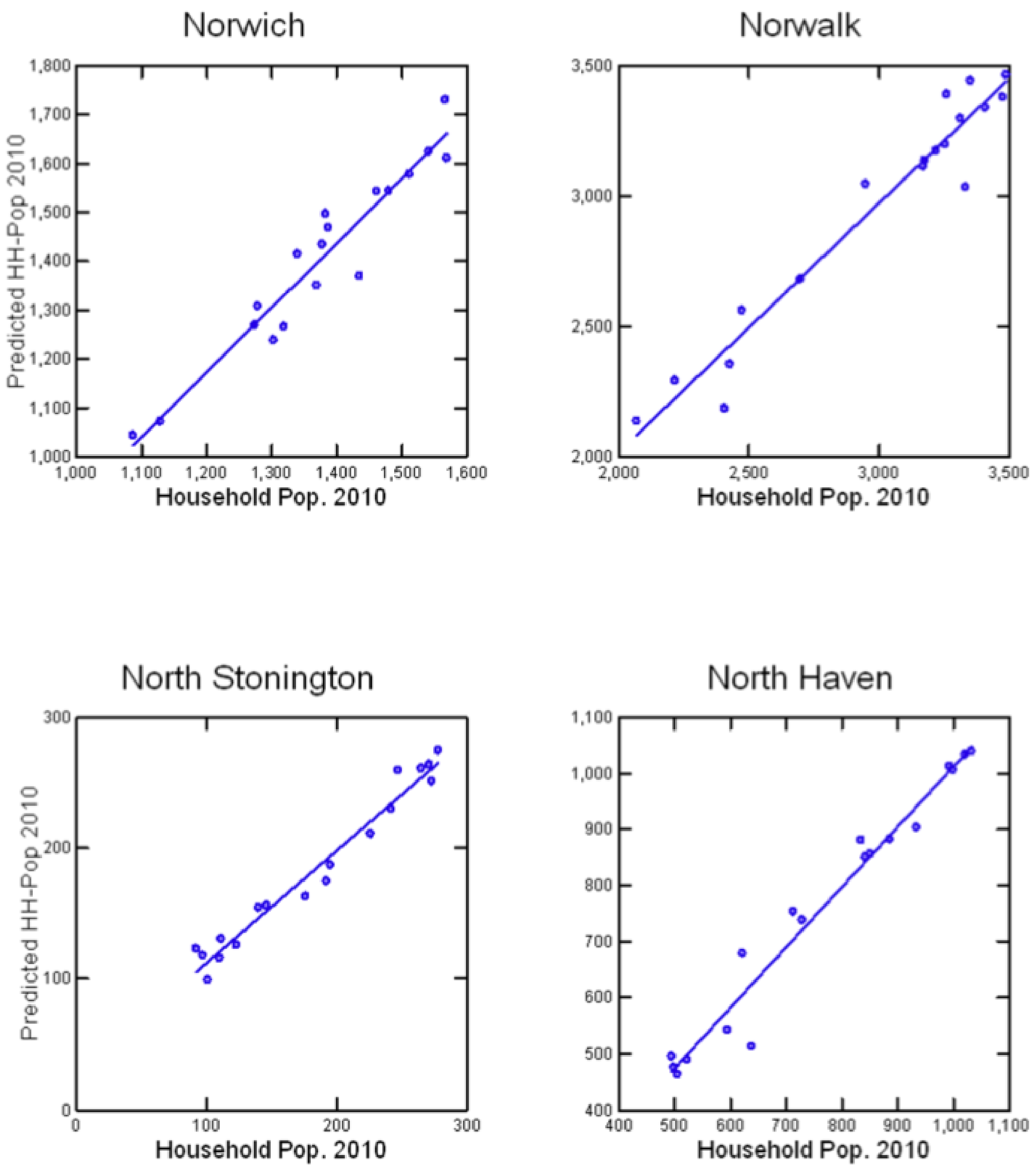
Pop Estimates Model for Ages 20-64 - 2010



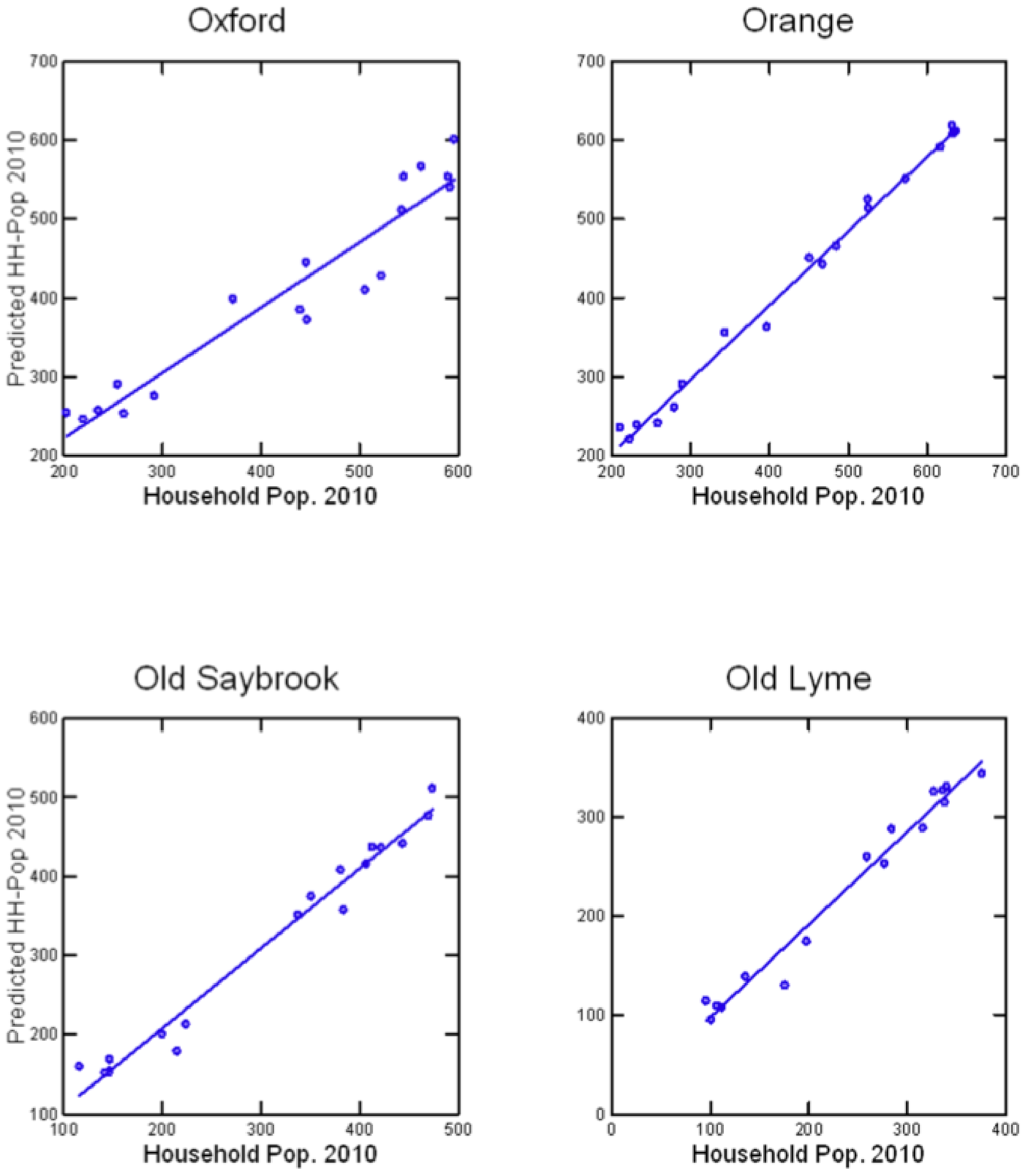
Pop Estimates Model for Ages 20-64 - 2010



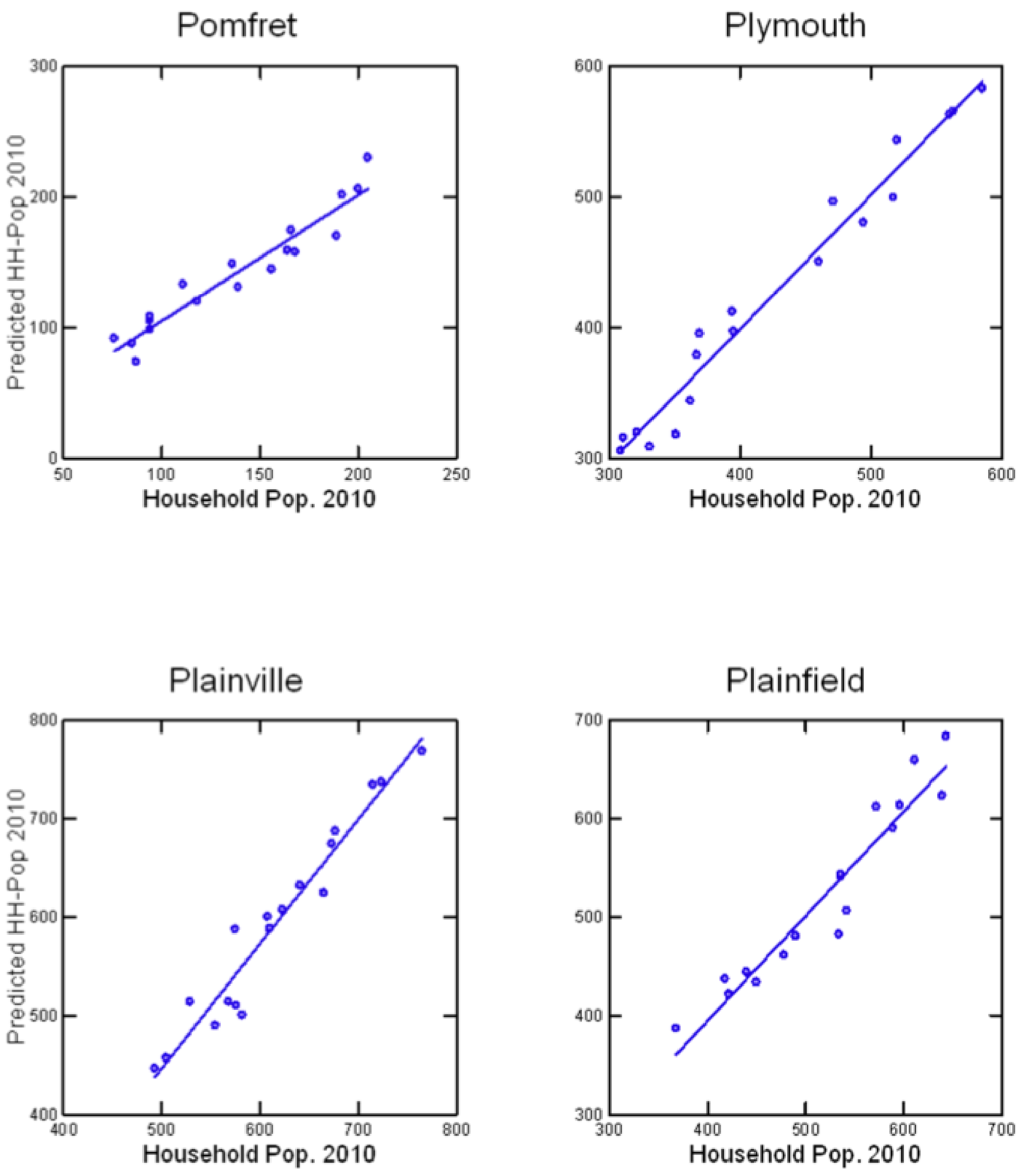
Pop Estimates Model for Ages 20-64 - 2010



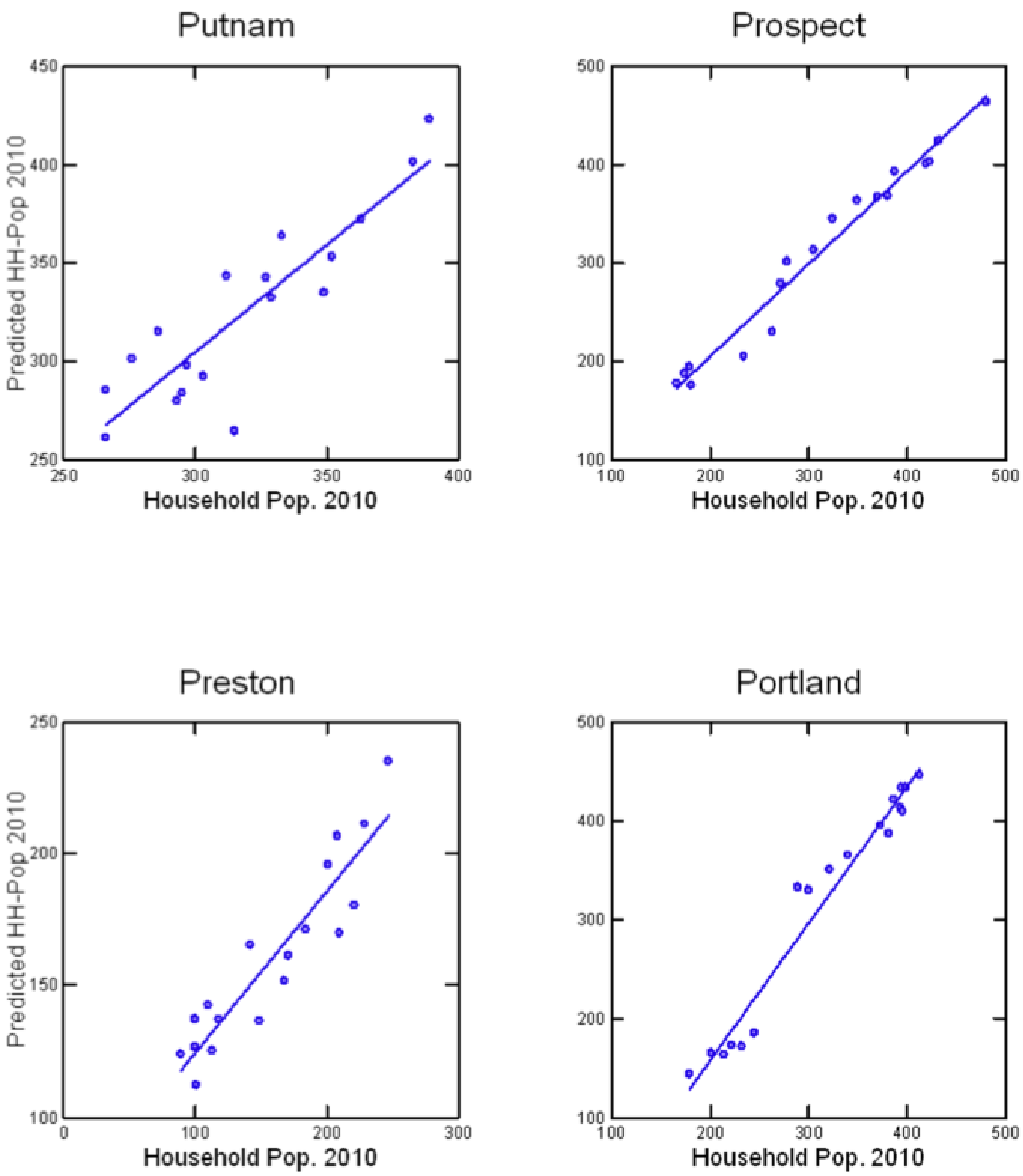
Pop Estimates Model for Ages 20-64 - 2010



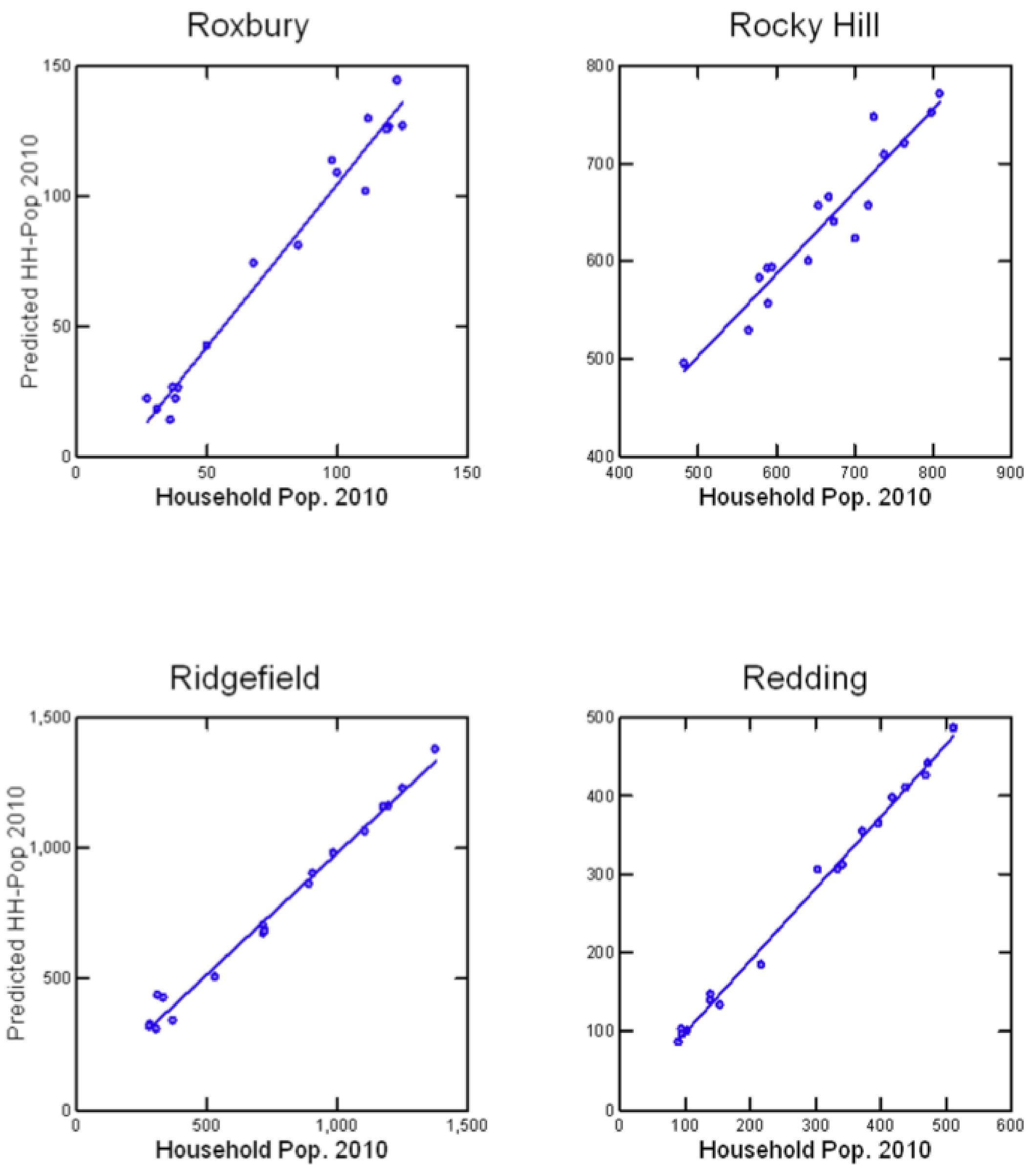
Pop Estimates Model for Ages 20-64 - 2010



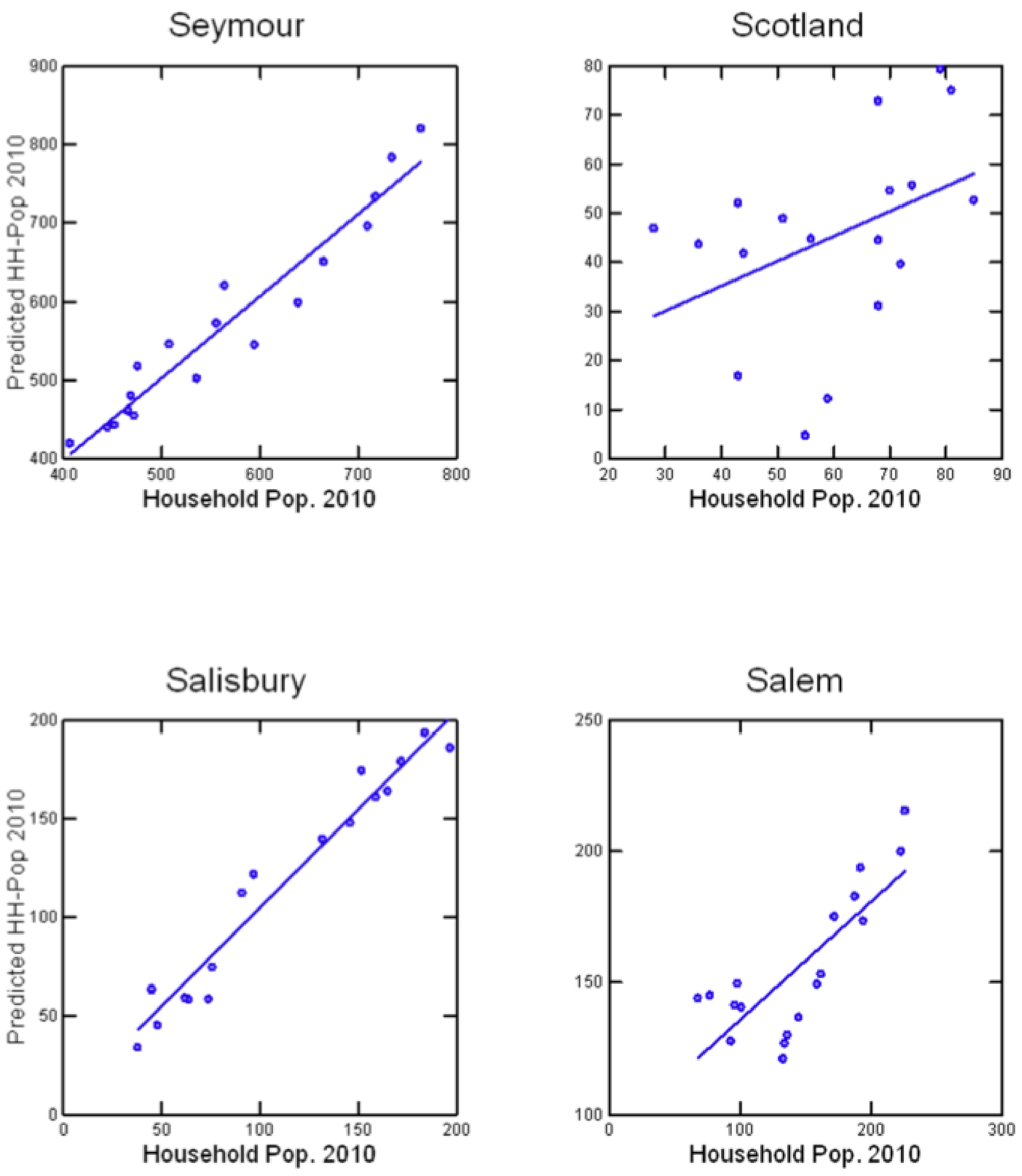
Pop Estimates Model for Ages 20-64 - 2010



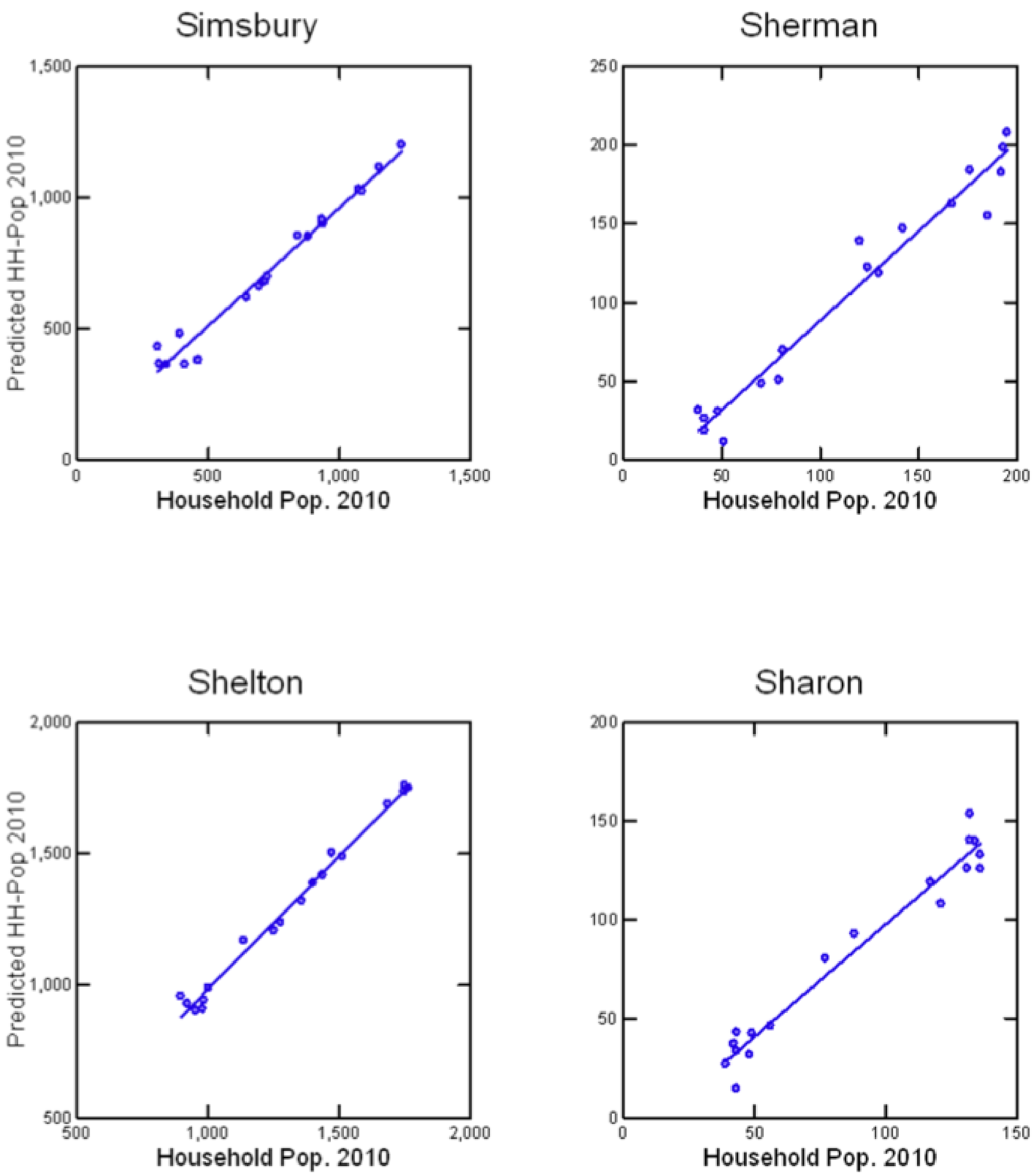
Pop Estimates Model for Ages 20-64 - 2010



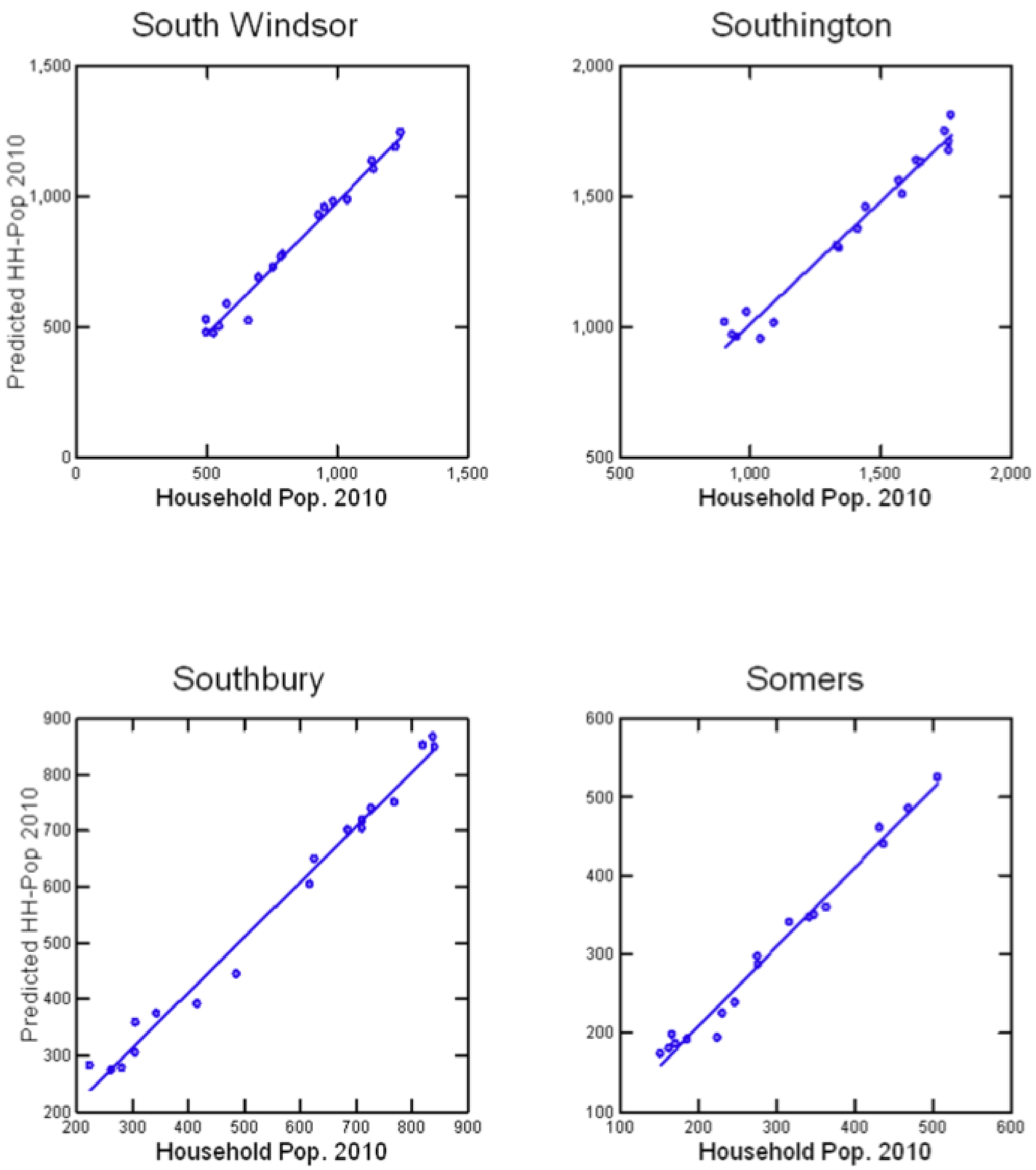
Pop Estimates Model for Ages 20-64 - 2010



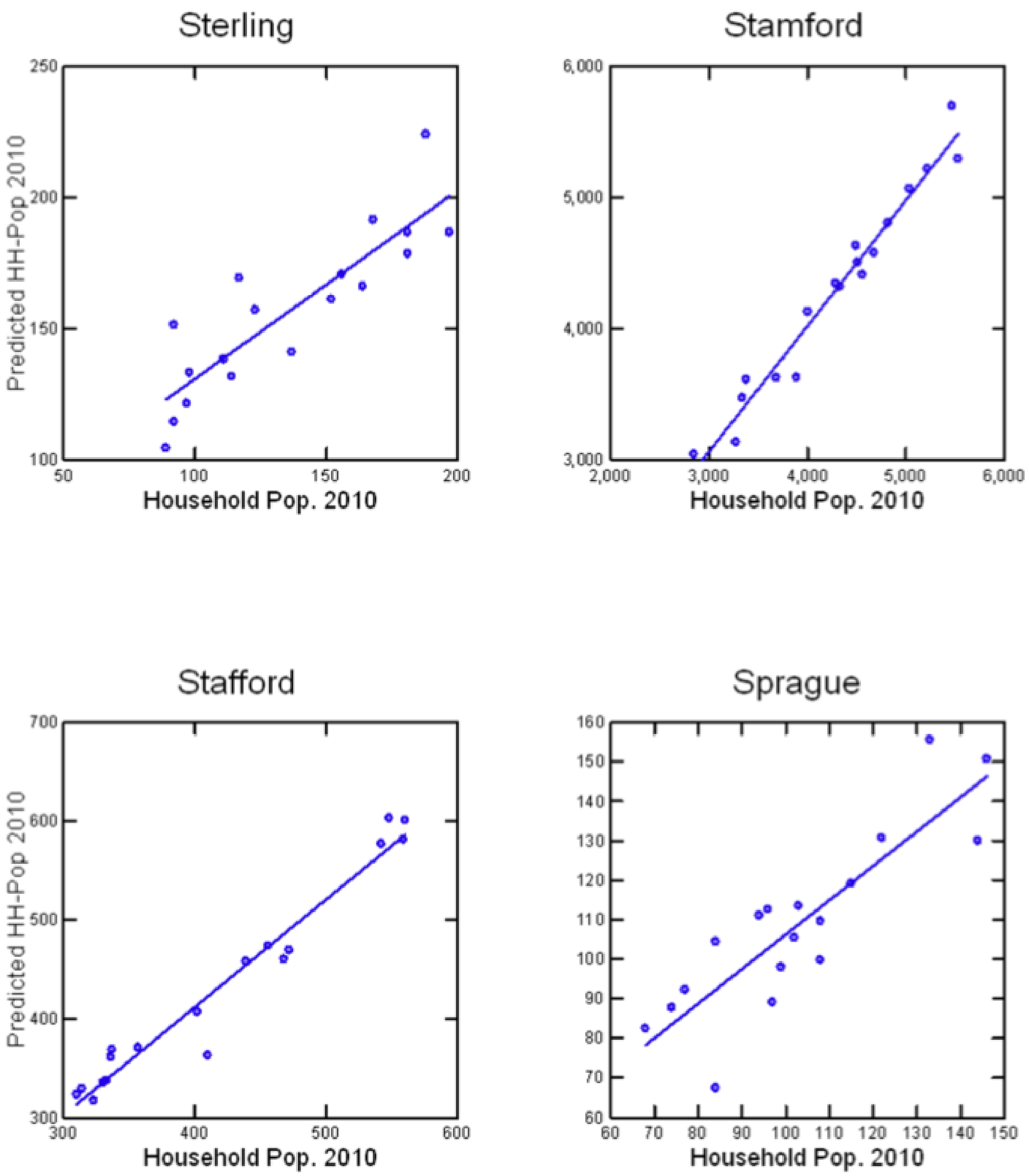
Pop Estimates Model for Ages 20-64 - 2010



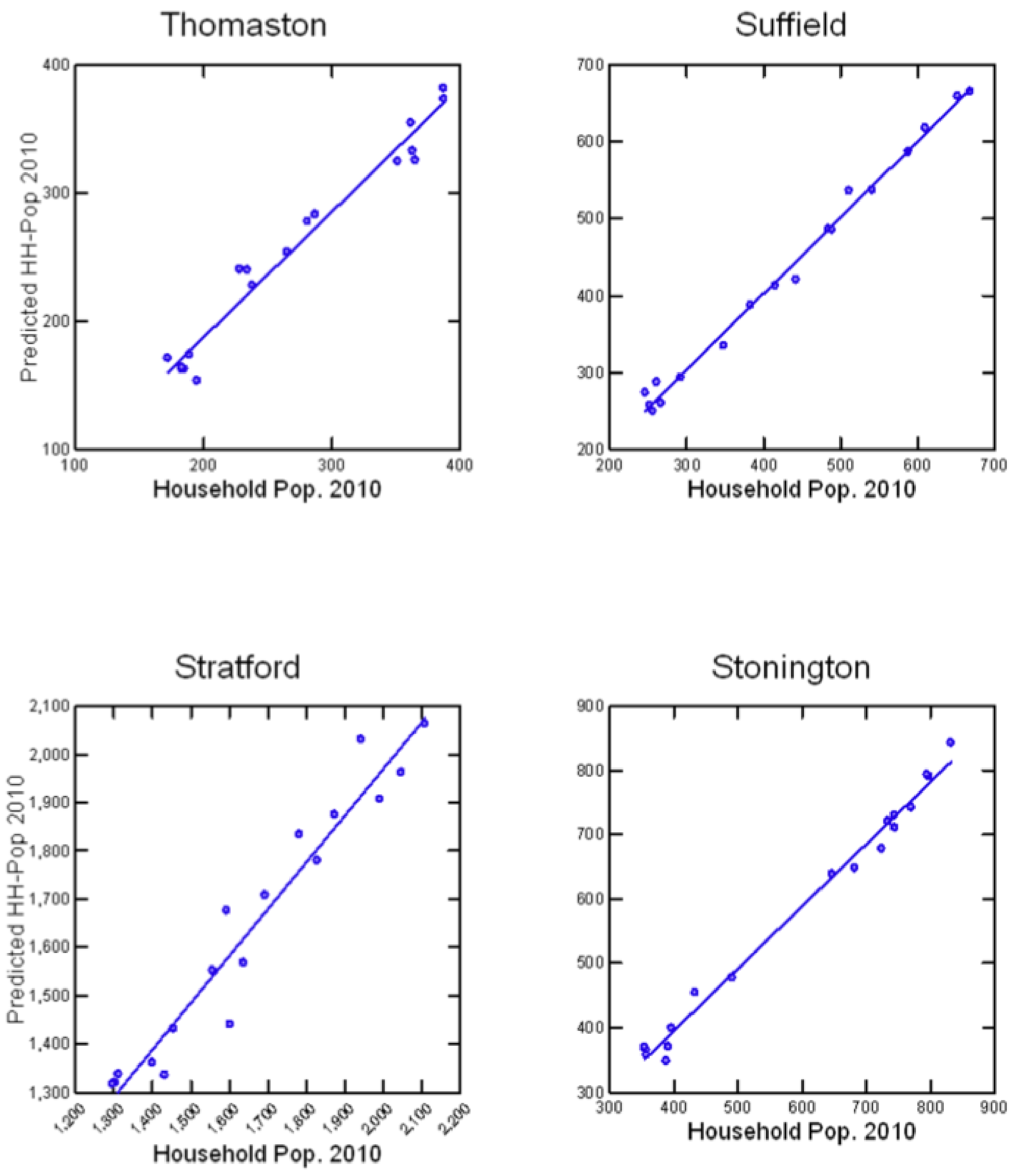
Pop Estimates Model for Ages 20-64 - 2010



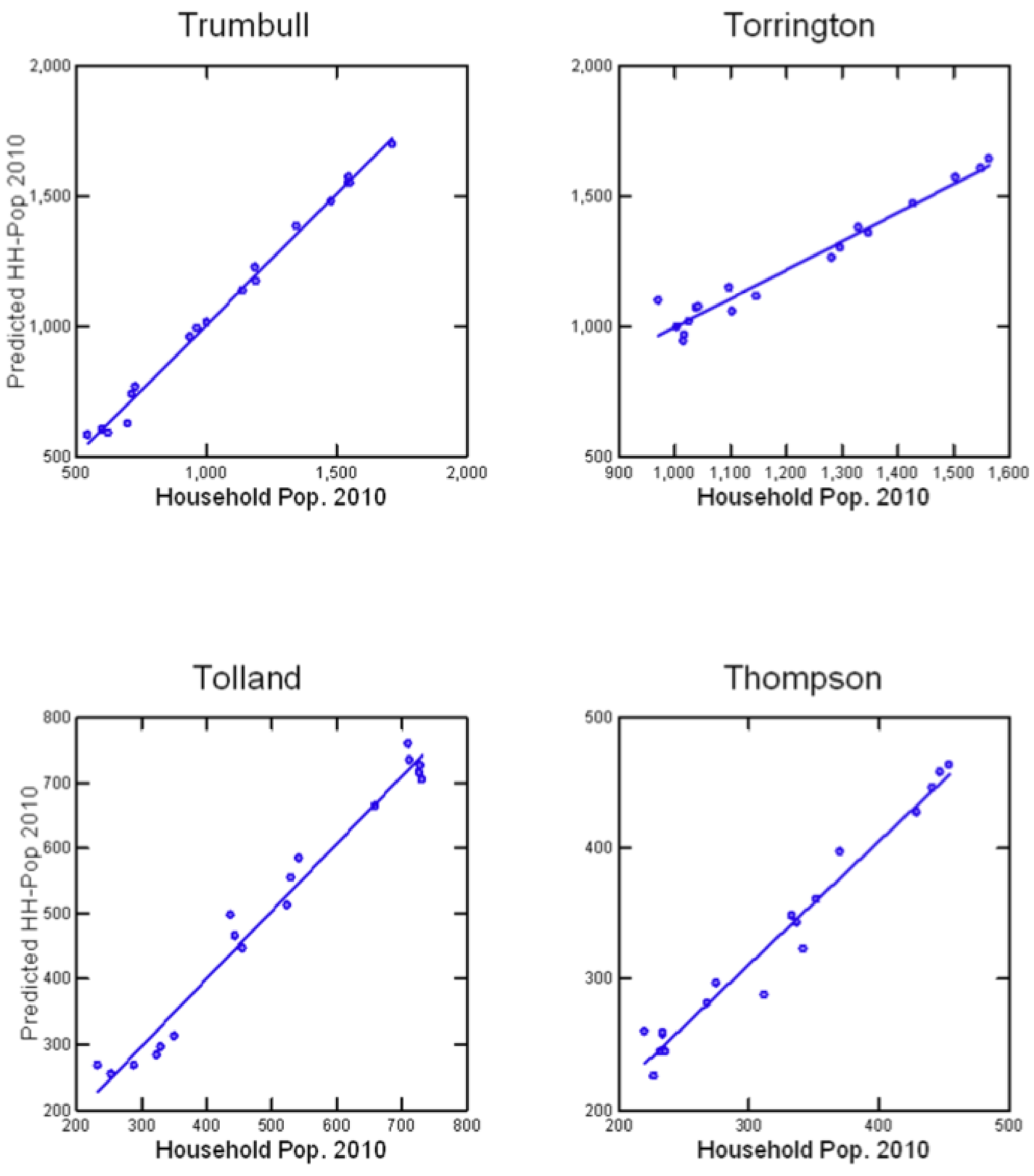
Pop Estimates Model for Ages 20-64 - 2010



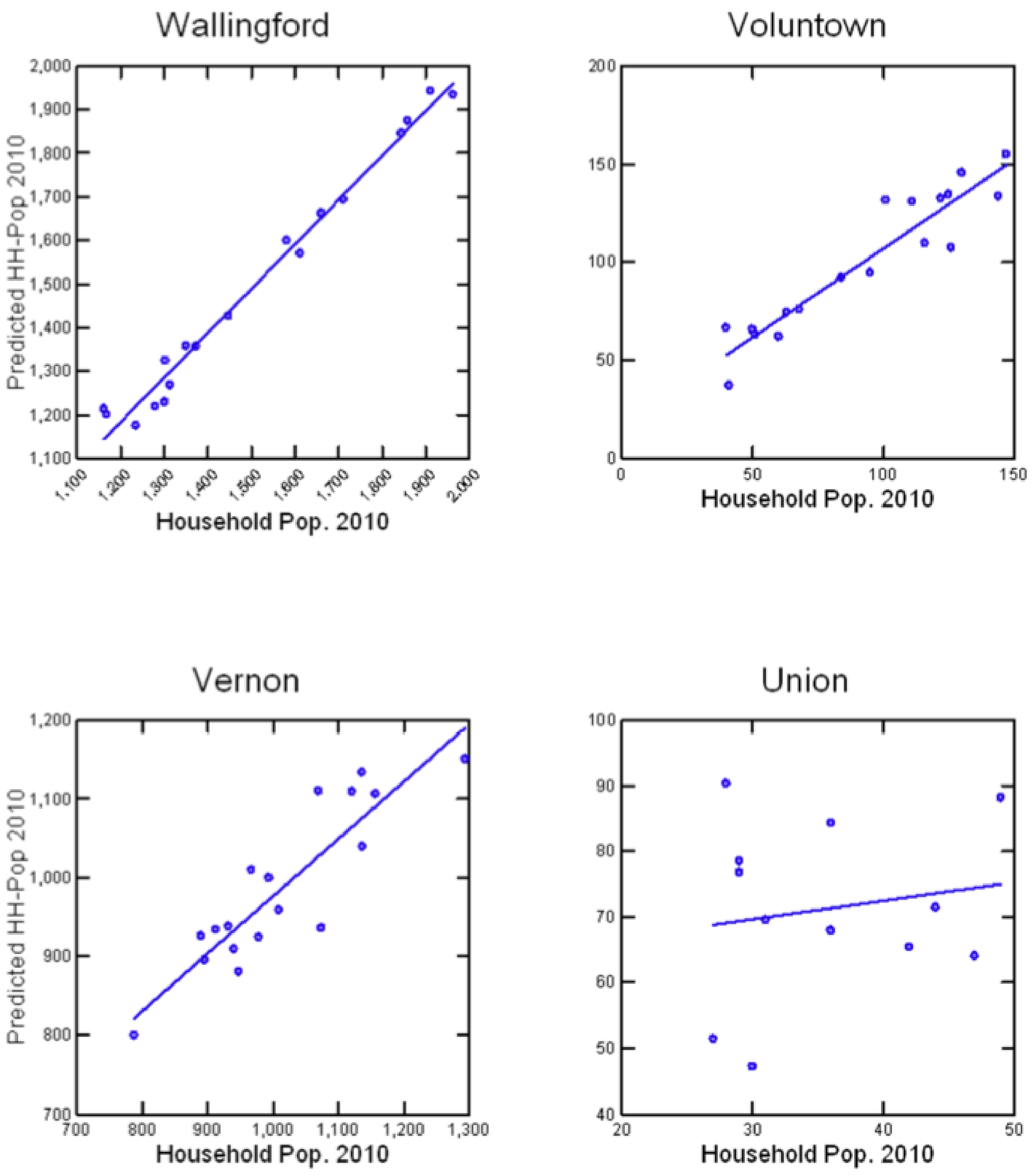
Pop Estimates Model for Ages 20-64 - 2010



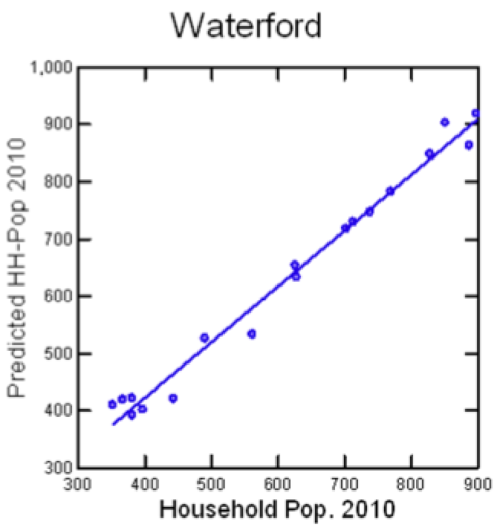
Pop Estimates Model for Ages 20-64 - 2010

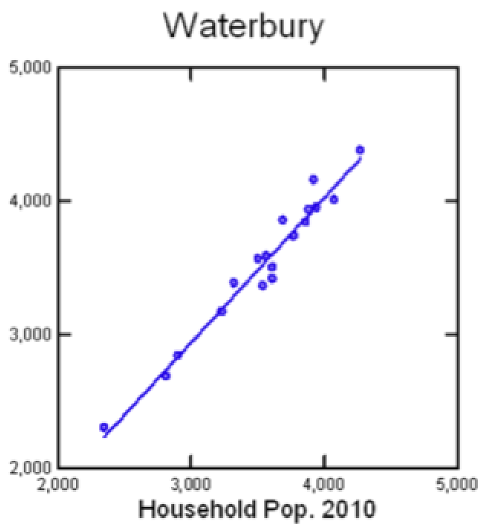


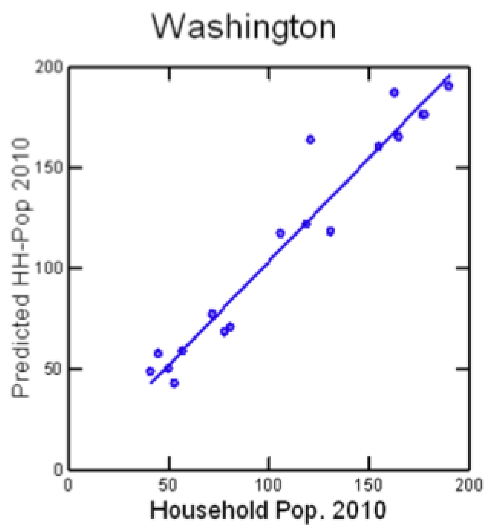
Pop Estimates Model for Ages 20-64 - 2010

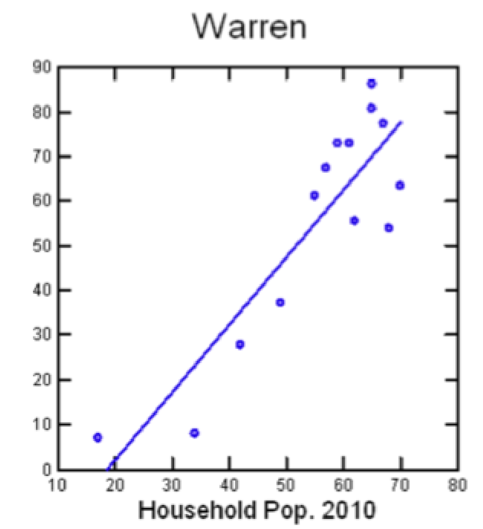


Pop Estimates Model for Ages 20-64 - 2010

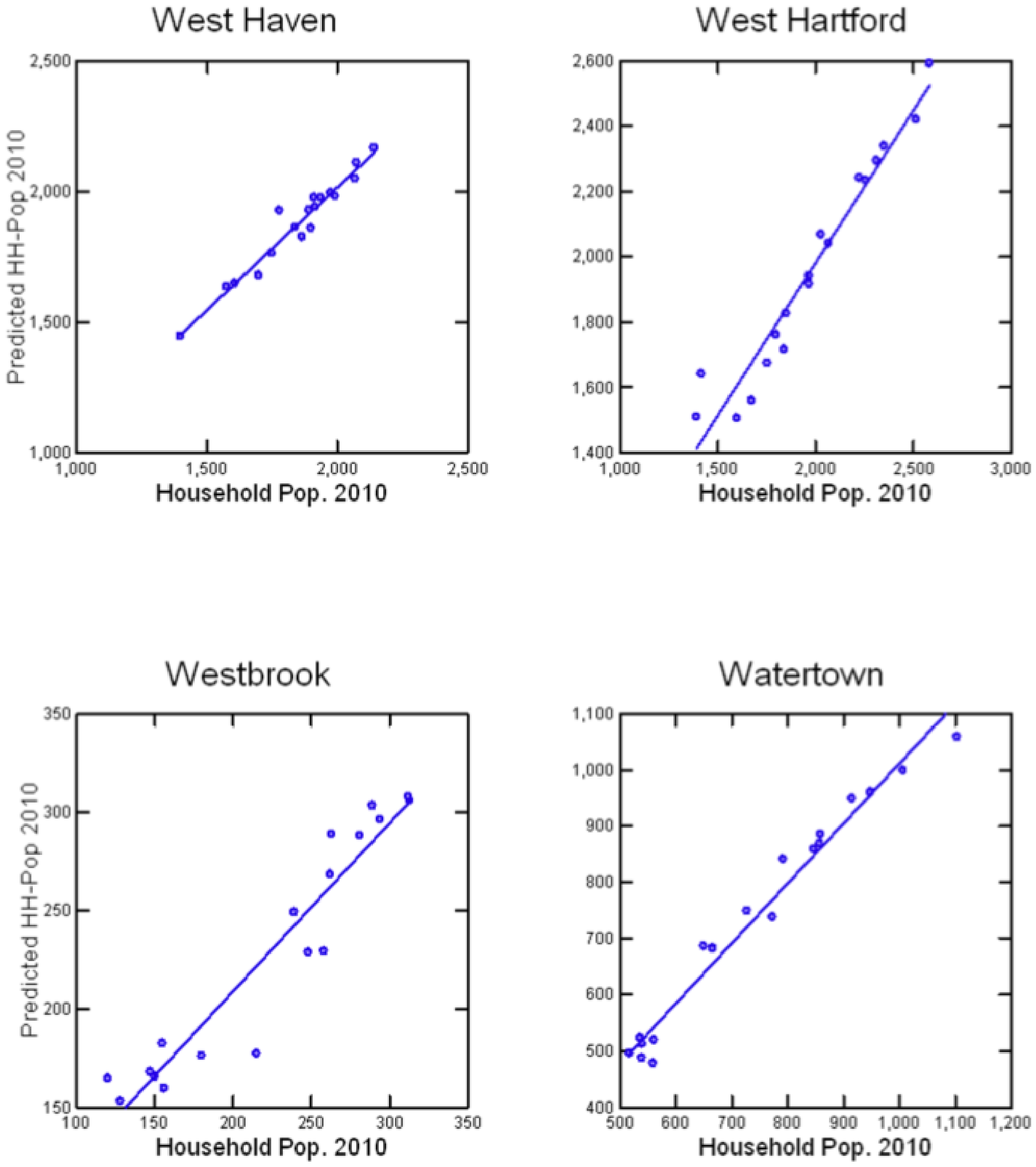




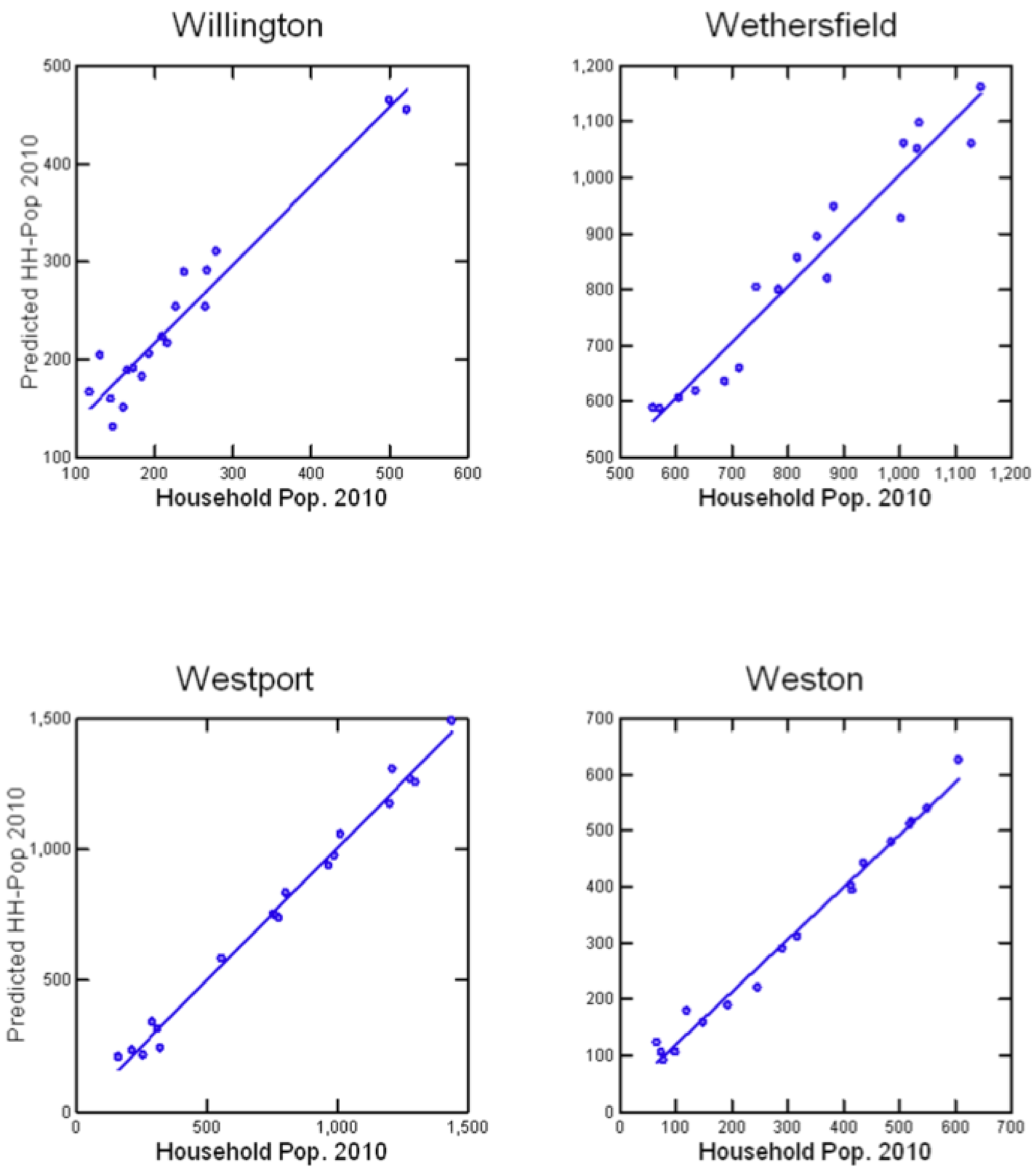




Pop Estimates Model for Ages 20-64 - 2010



Pop Estimates Model for Ages 20-64 - 2010



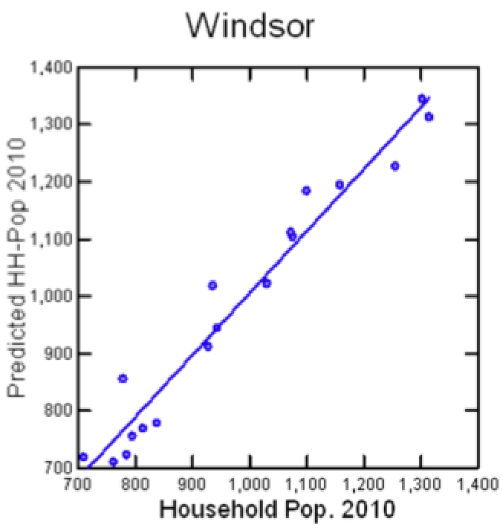
Willington

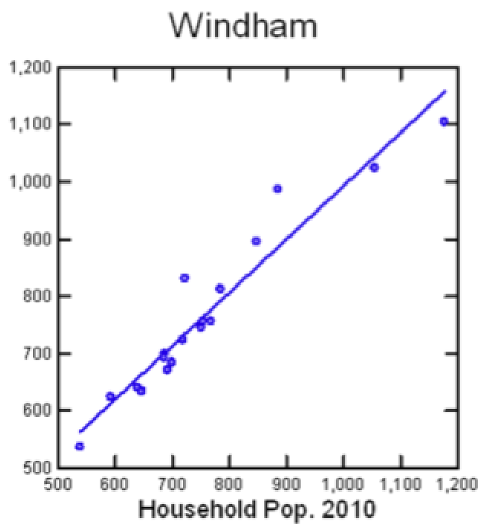
Wethersfield

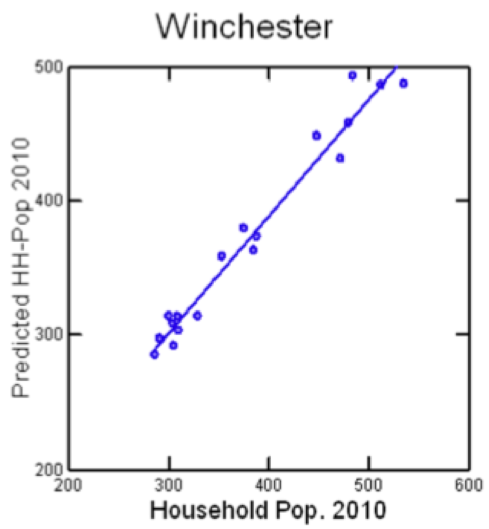
Westport

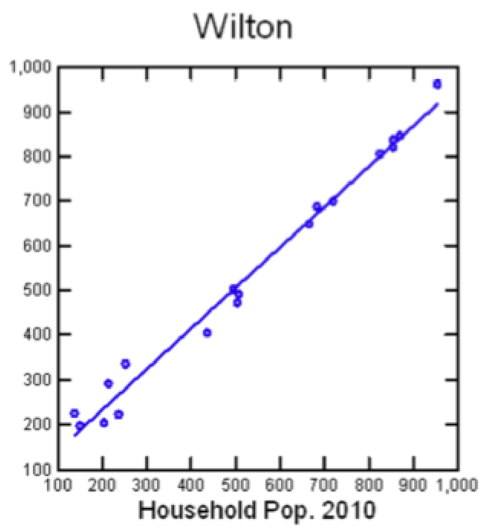
Weston

Pop Estimates Model for Ages 20-64 - 2010

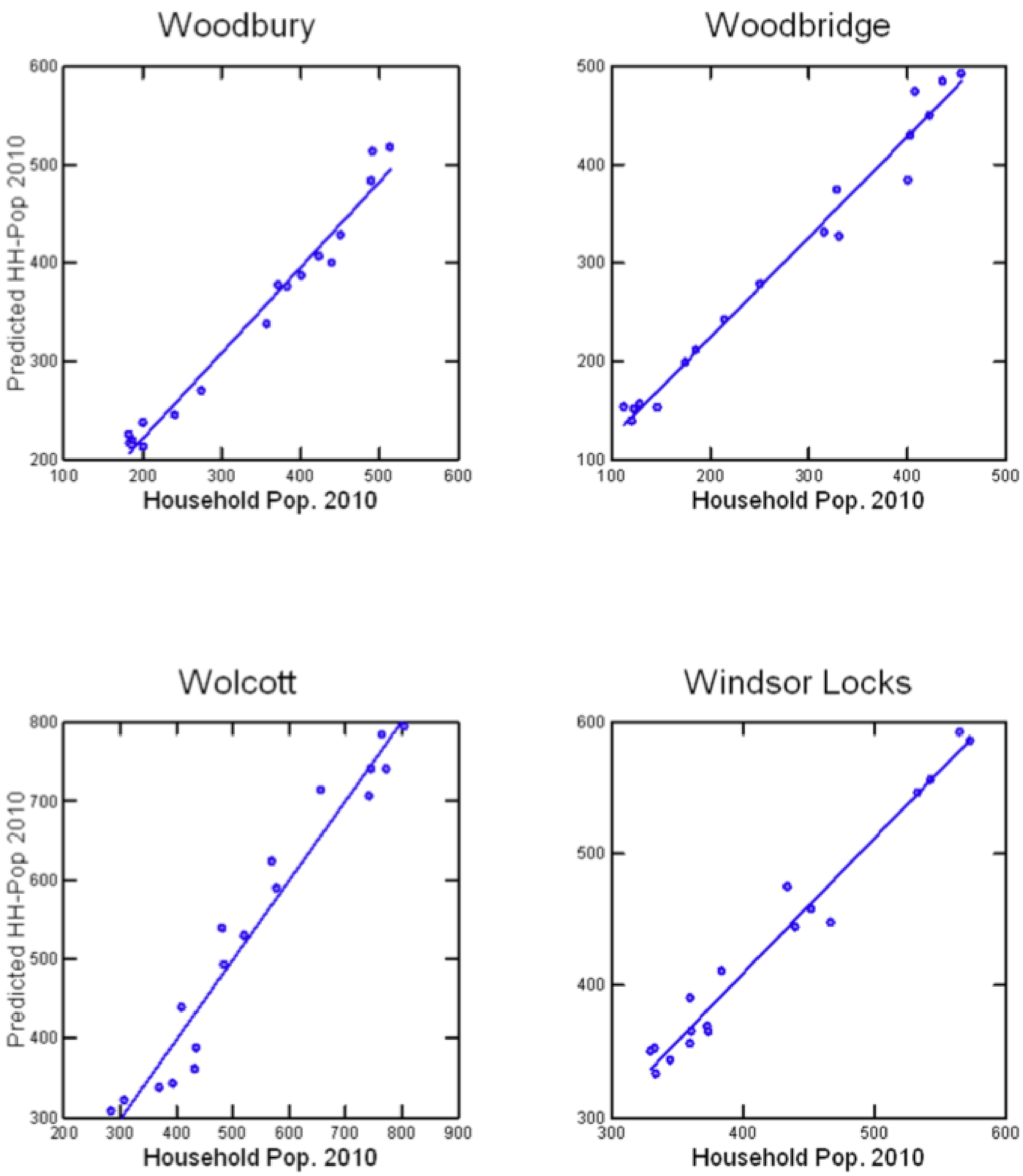

Windsor


Windham

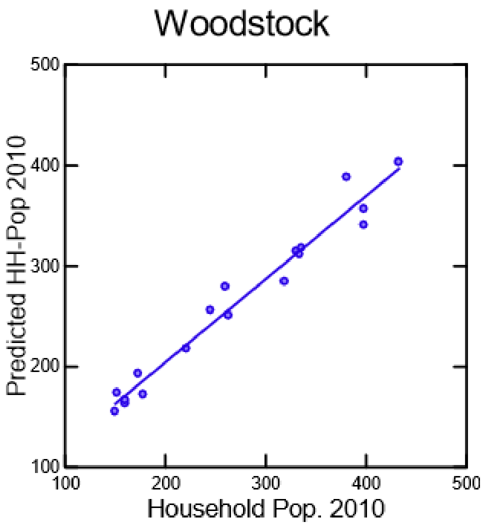

Winchester


Wilton

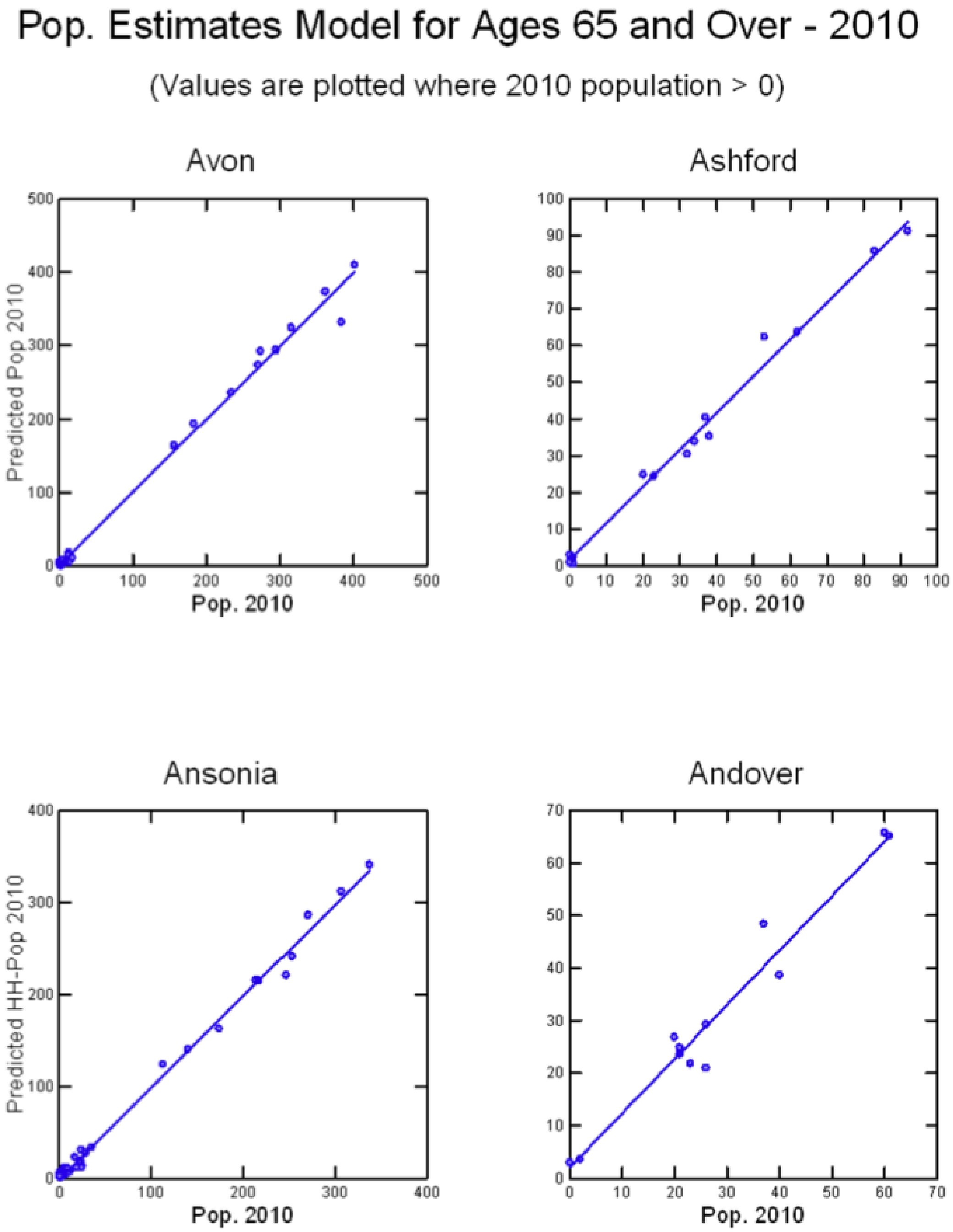
Pop Estimates Model for Ages 20-64 - 2010



Pop Estimates Model for Ages 20-64 - 2010

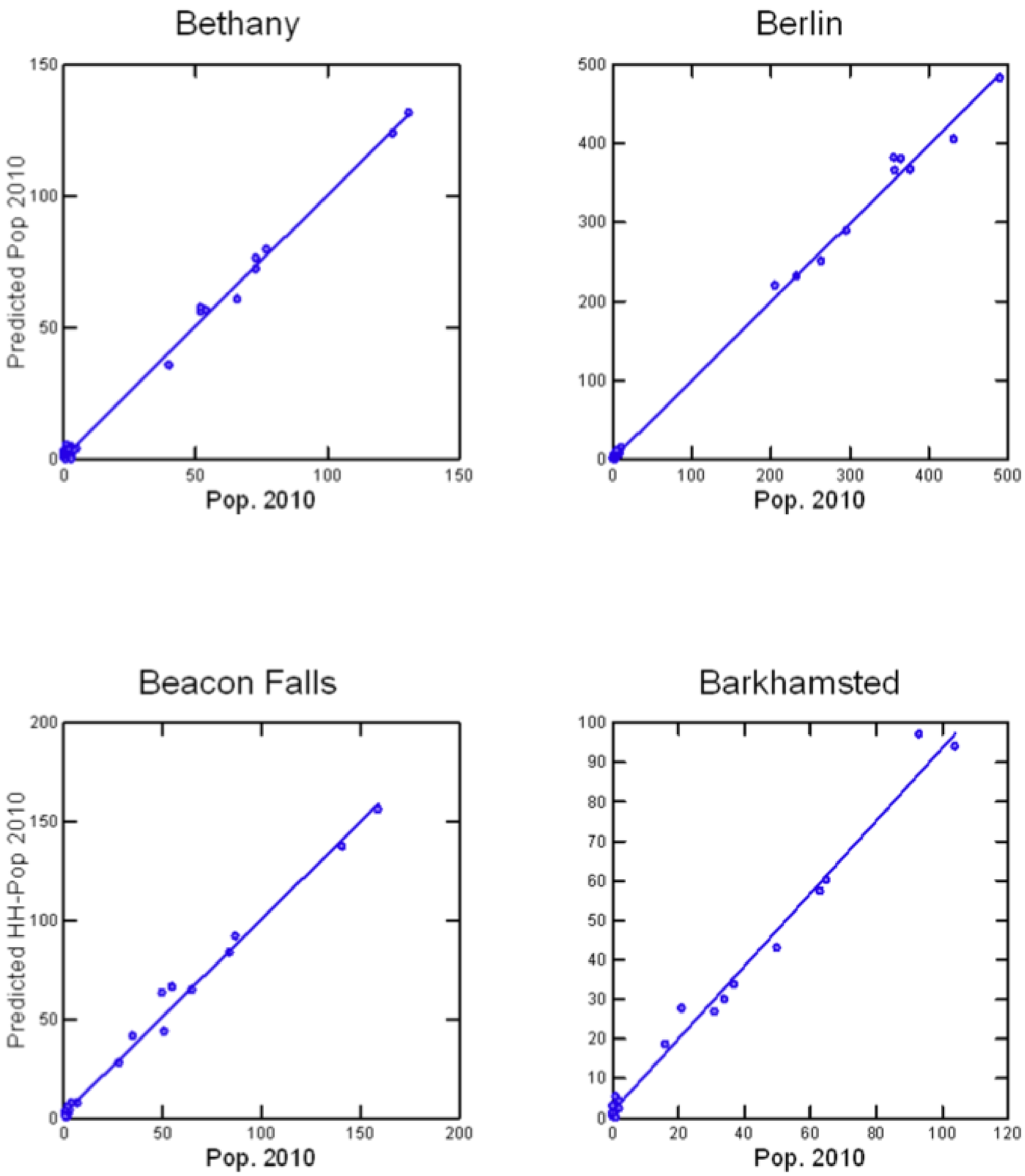


Appendix B: Population Estimates Model for Ages 65 and Over – 2010 by Town



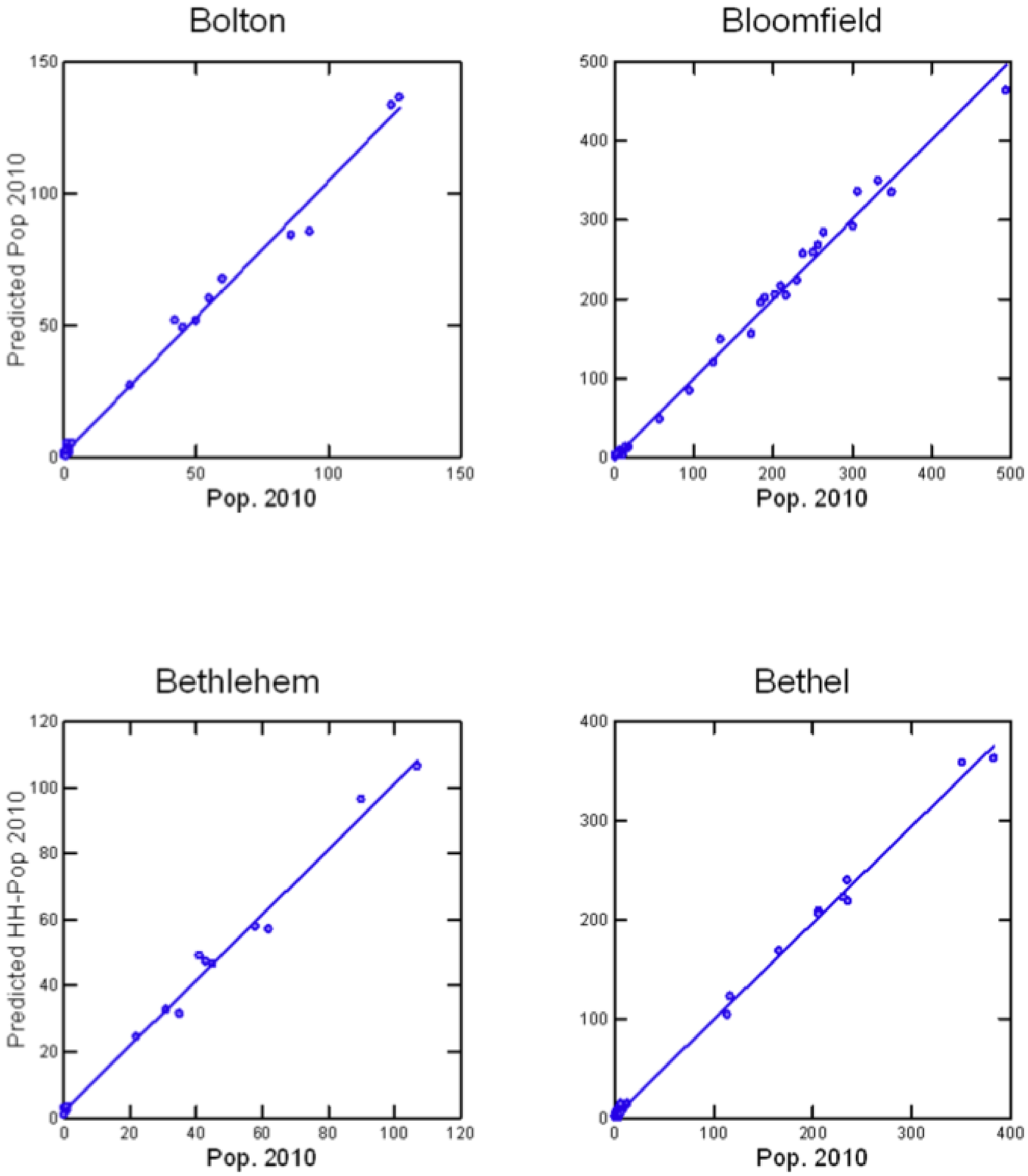
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



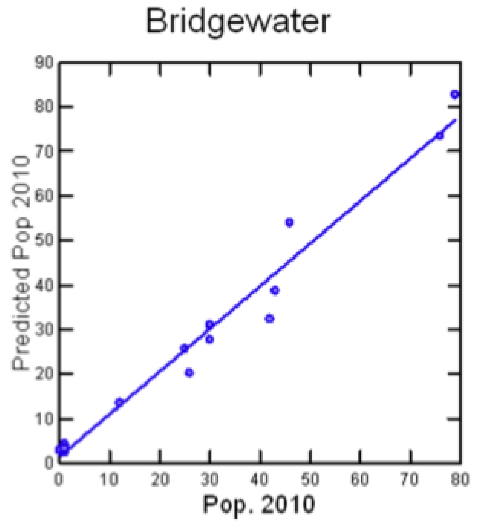
Pop. Estimates Model for Ages 65 and Over - 2010

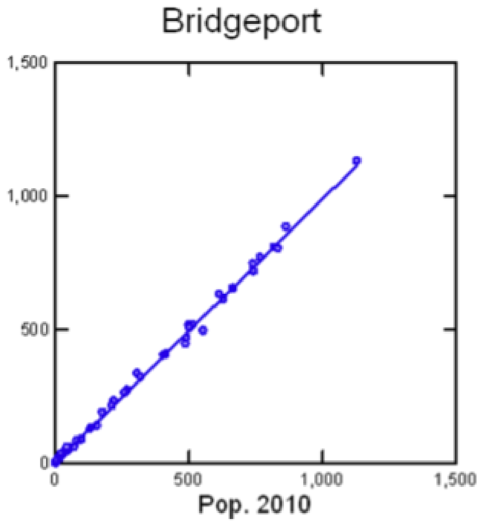
(Values are plotted where 2010 population > 0)

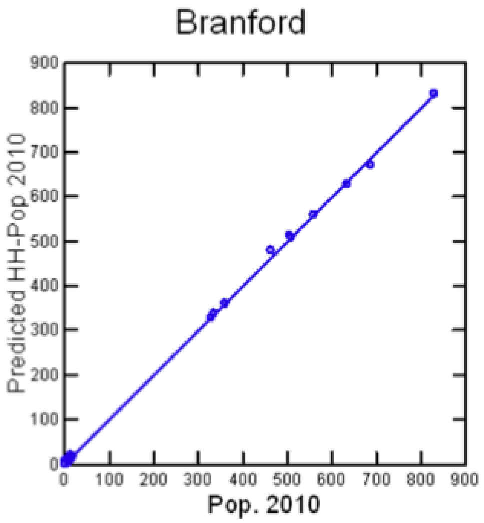


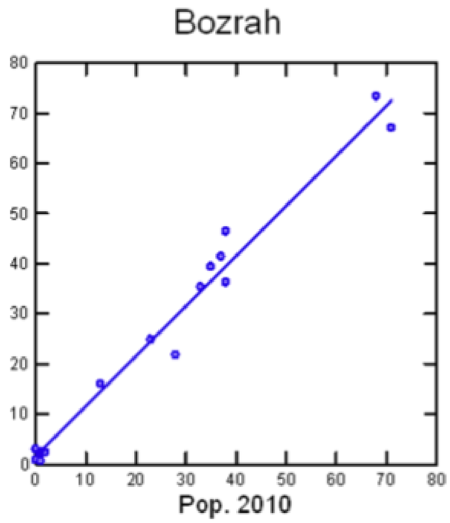
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



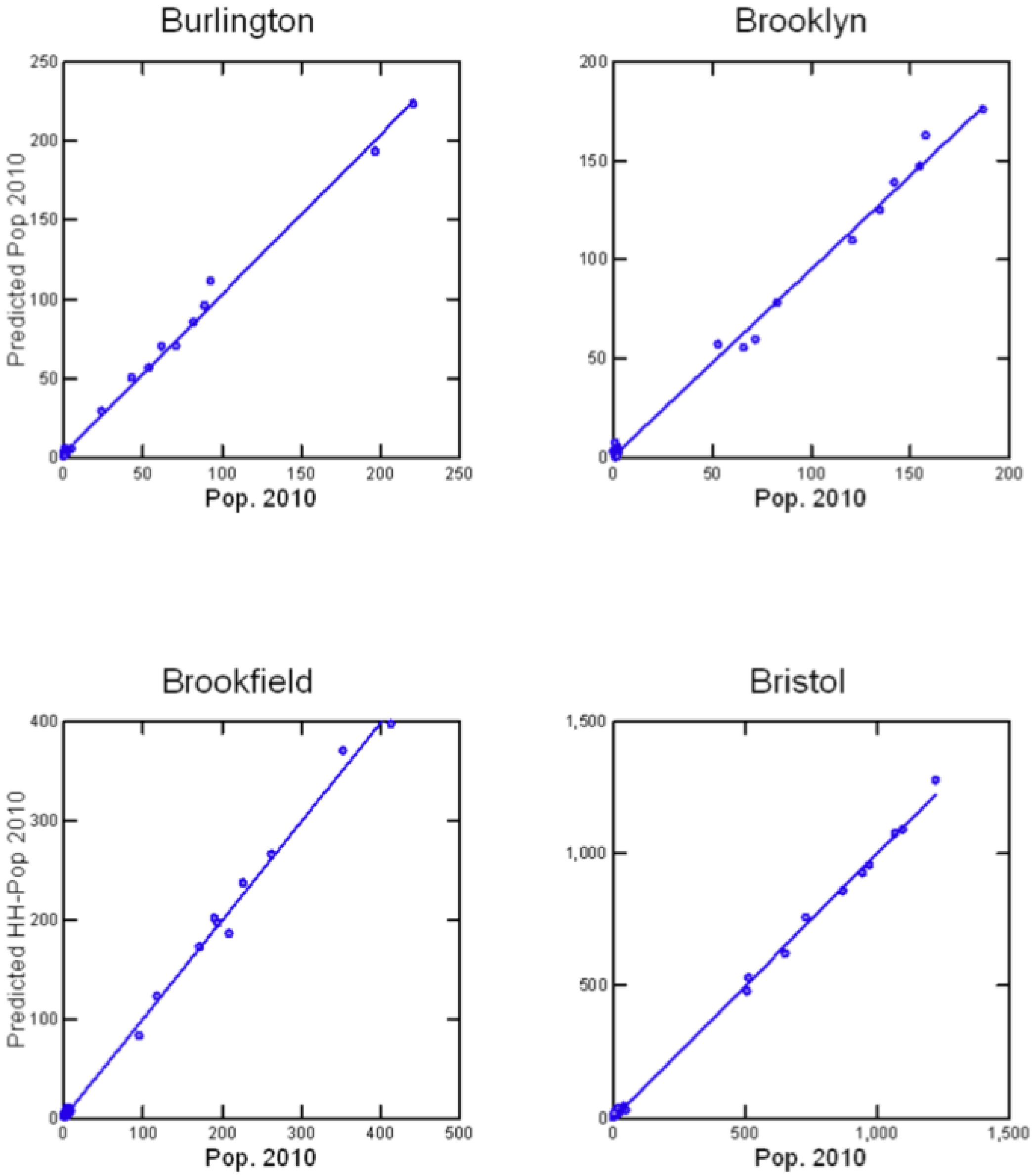






Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)


Burlington

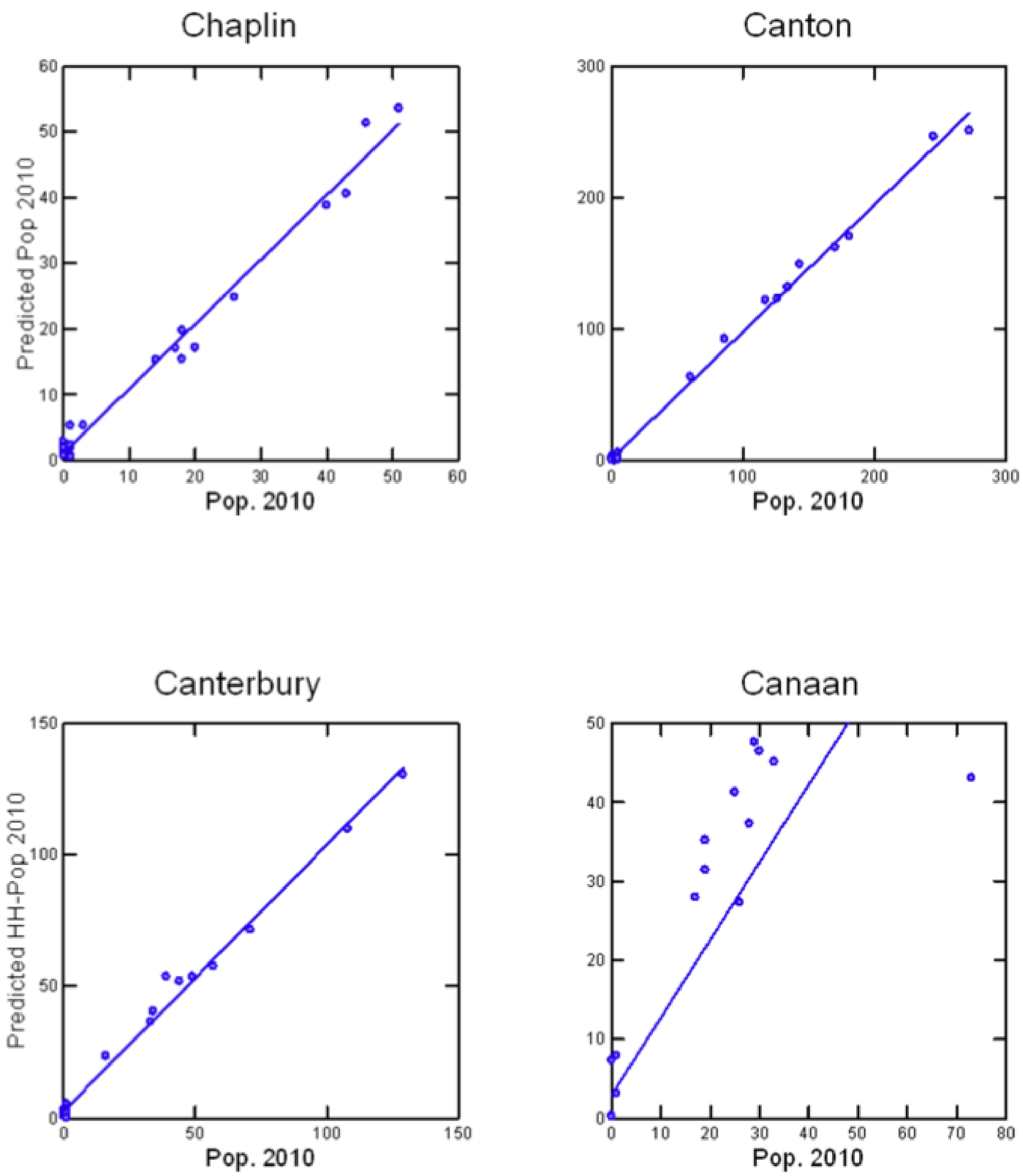
Brooklyn

Brookfield

Bristol

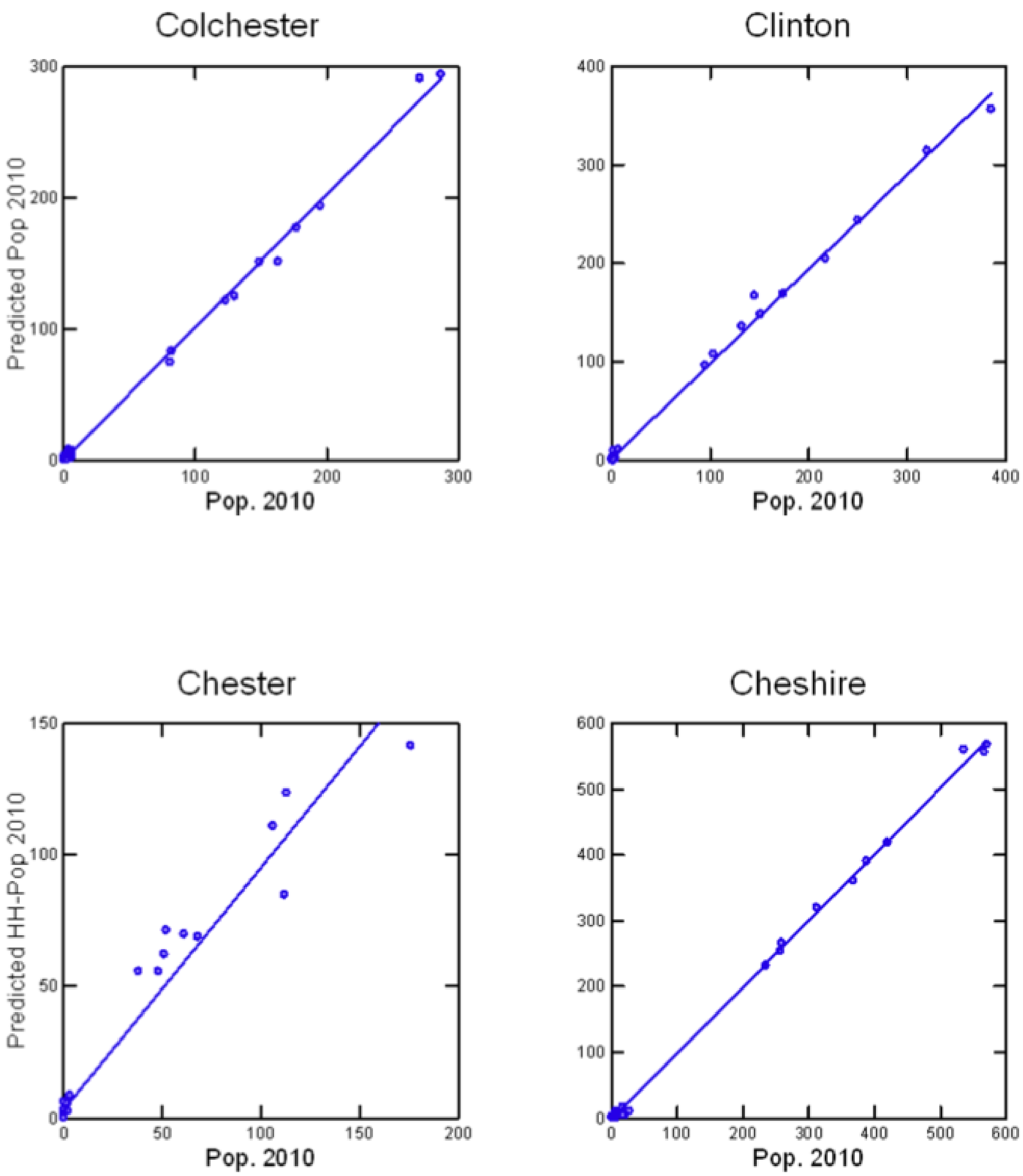
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



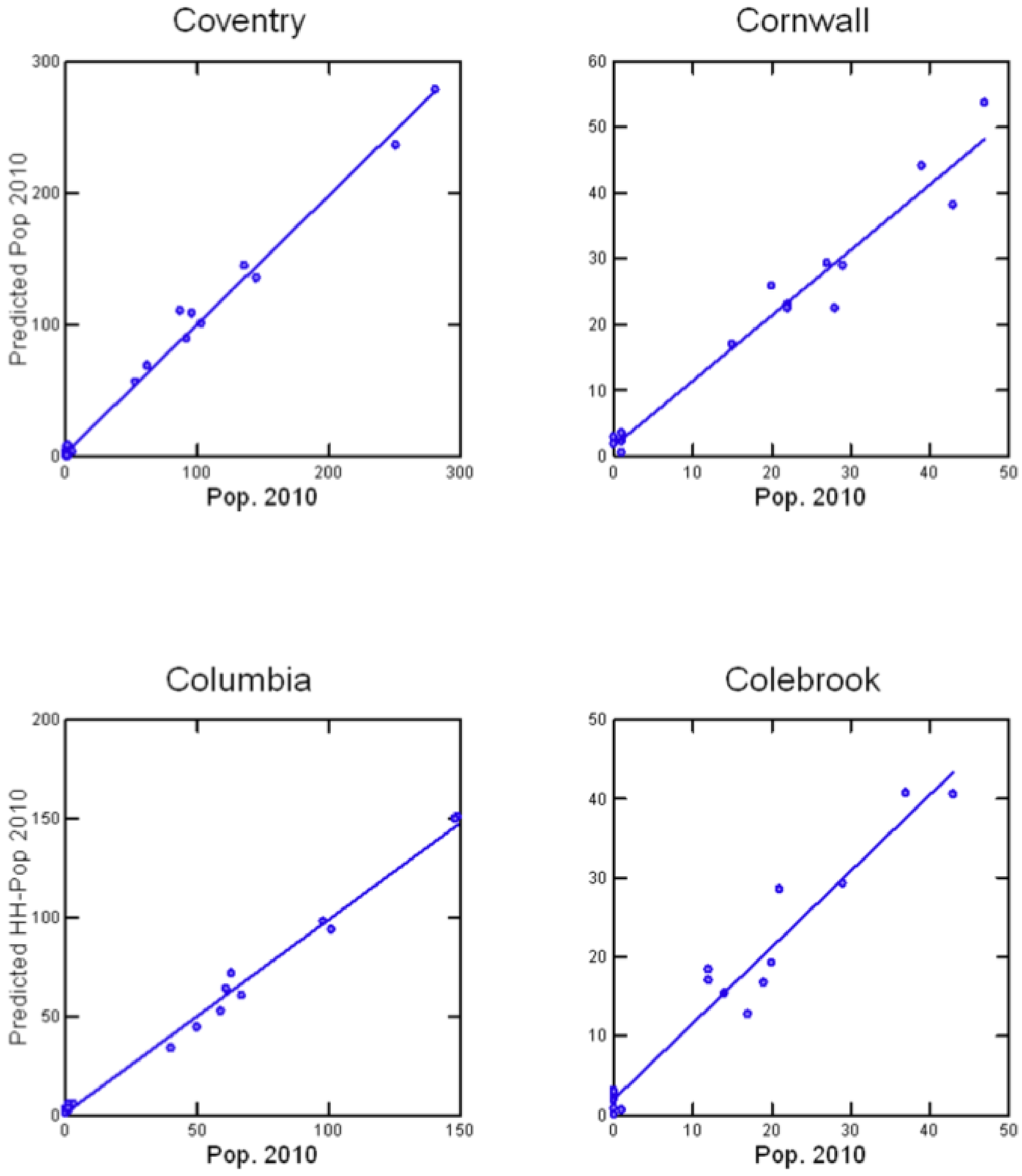
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



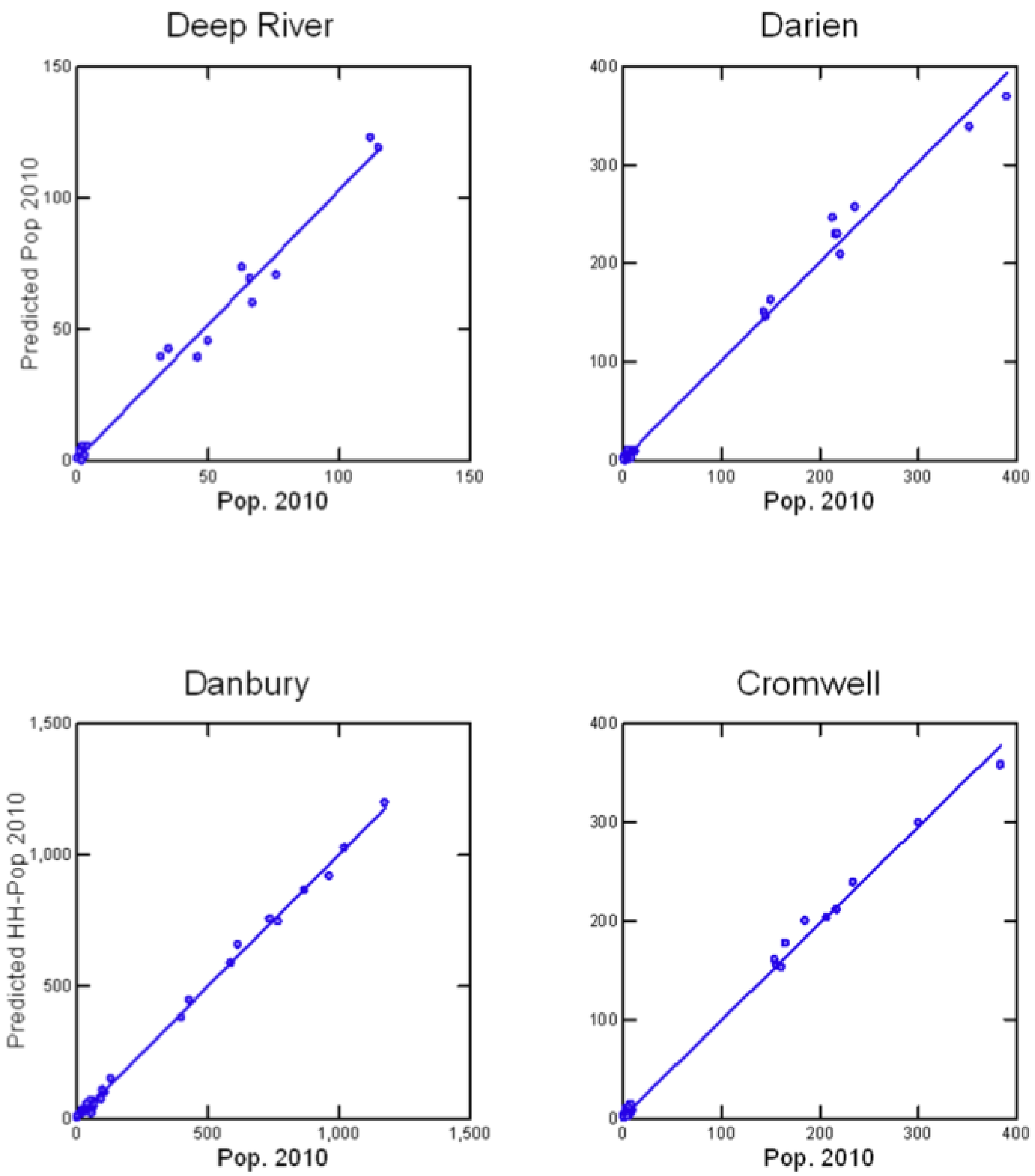
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



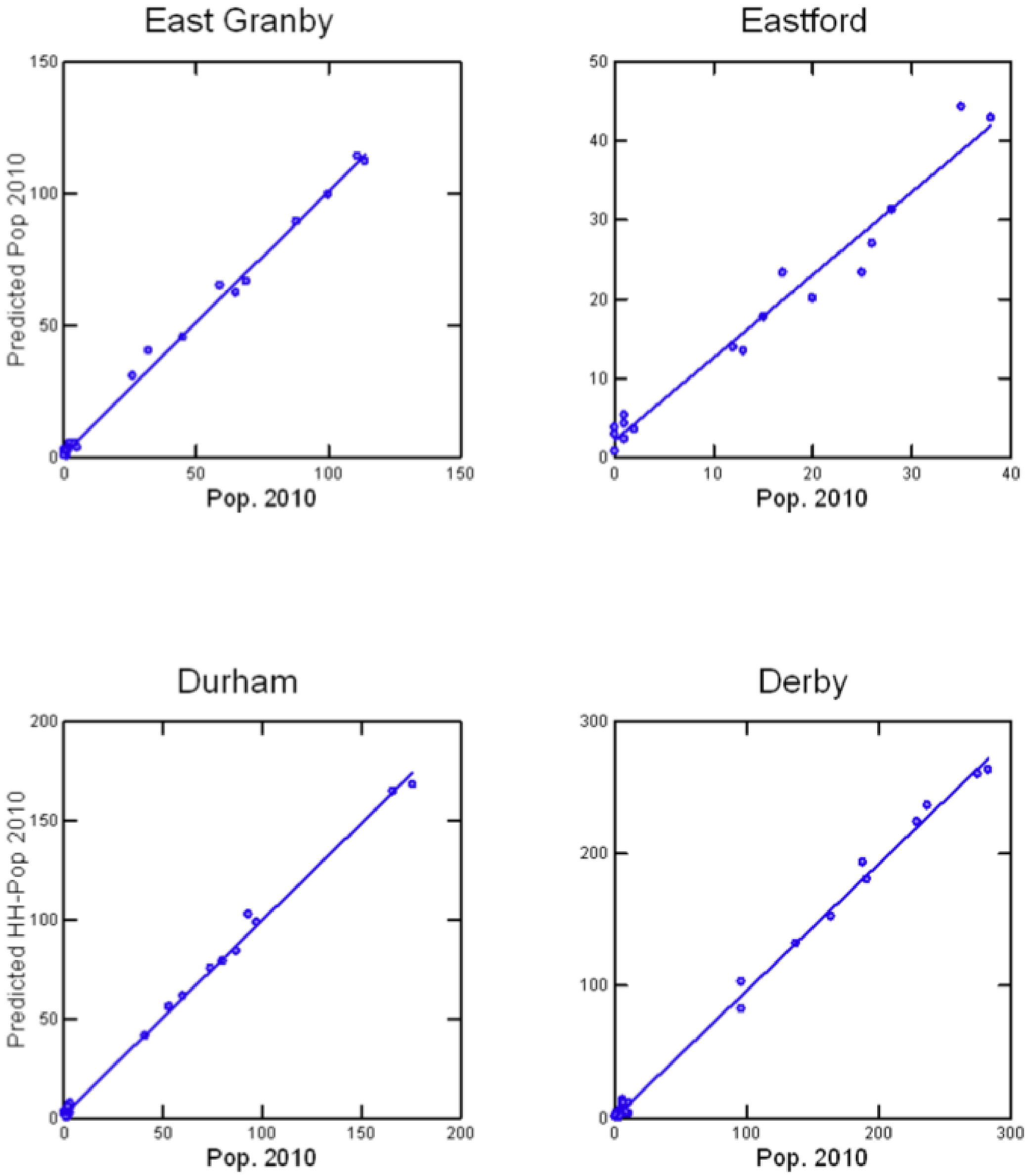
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



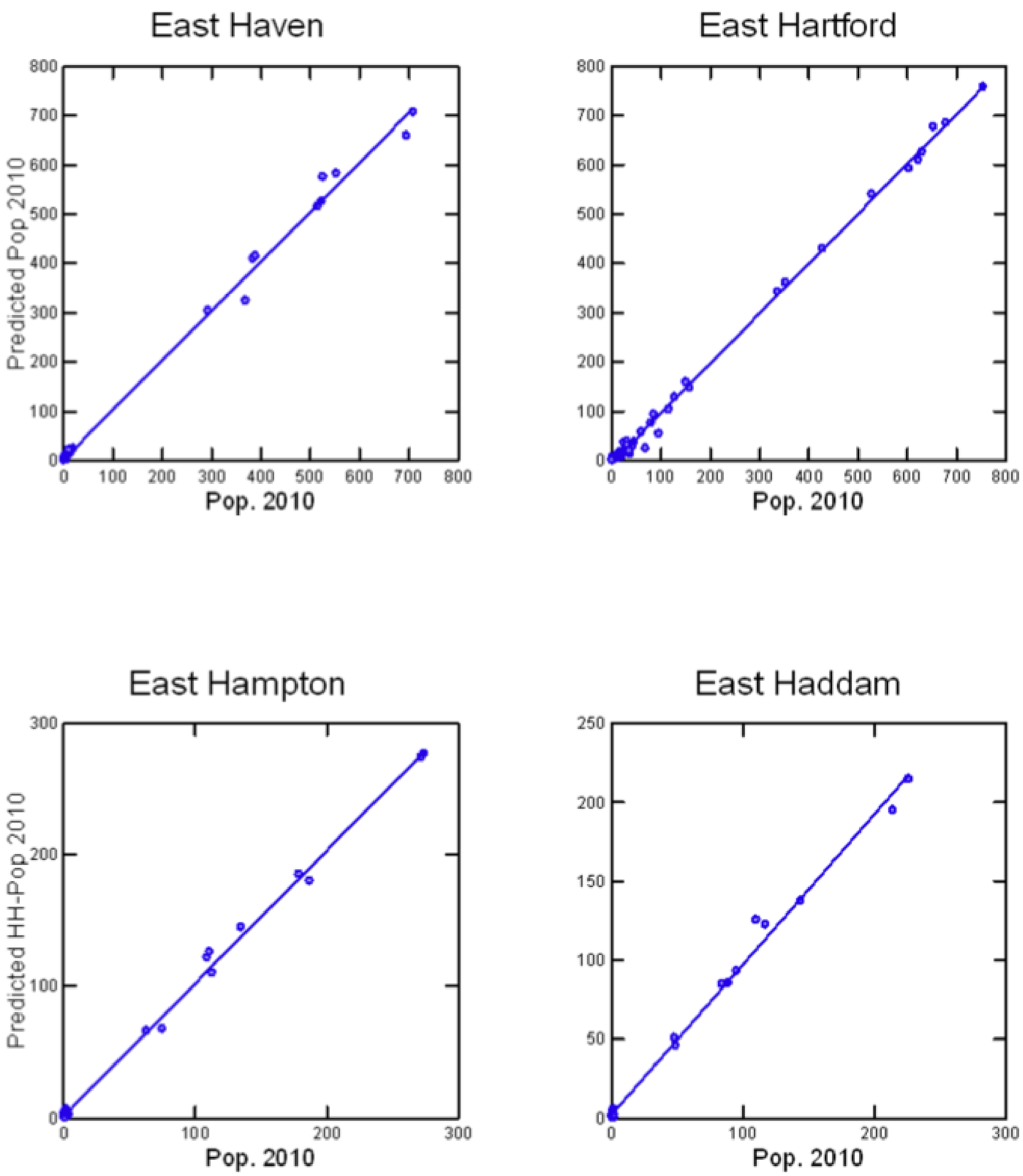
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



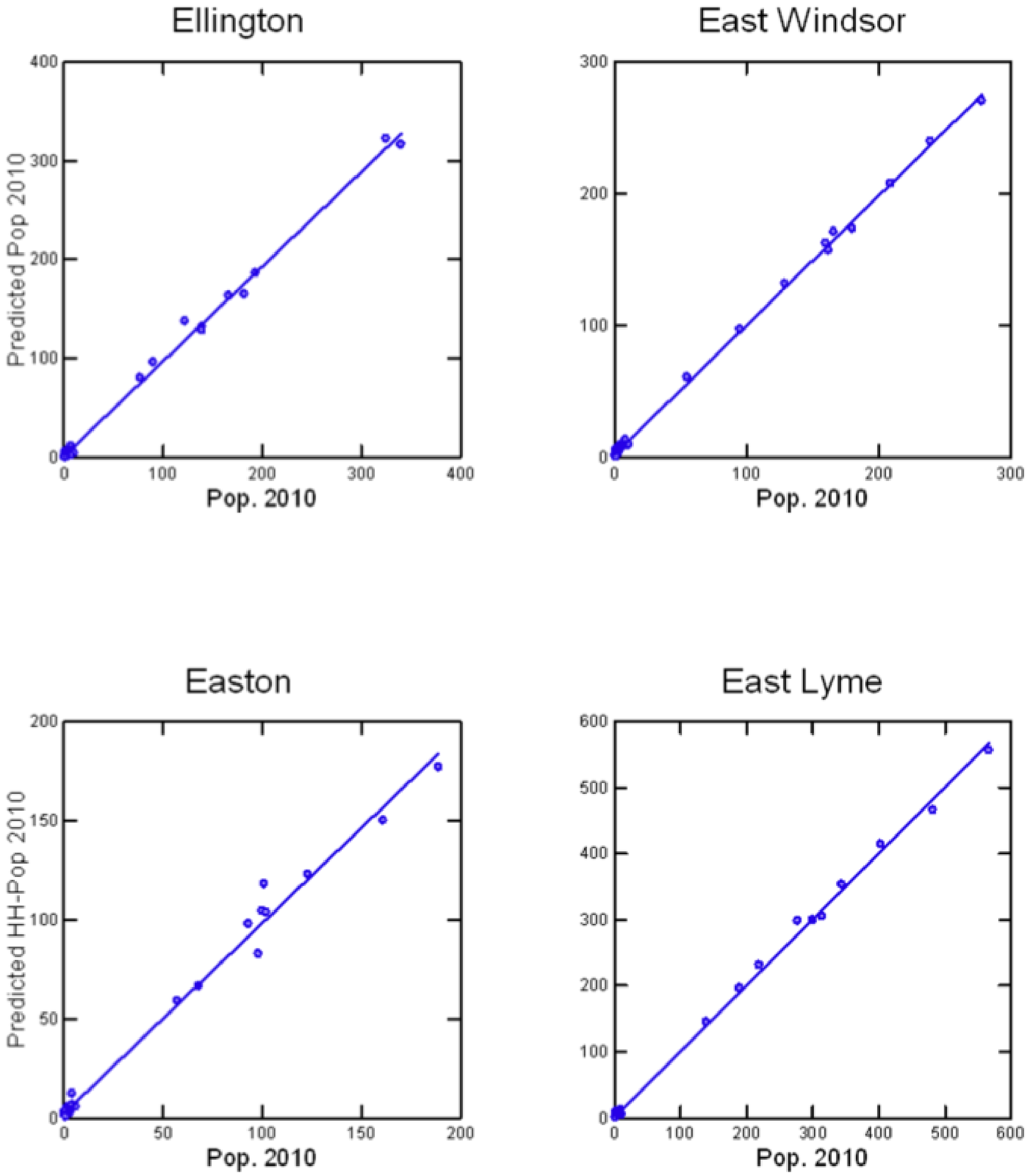
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



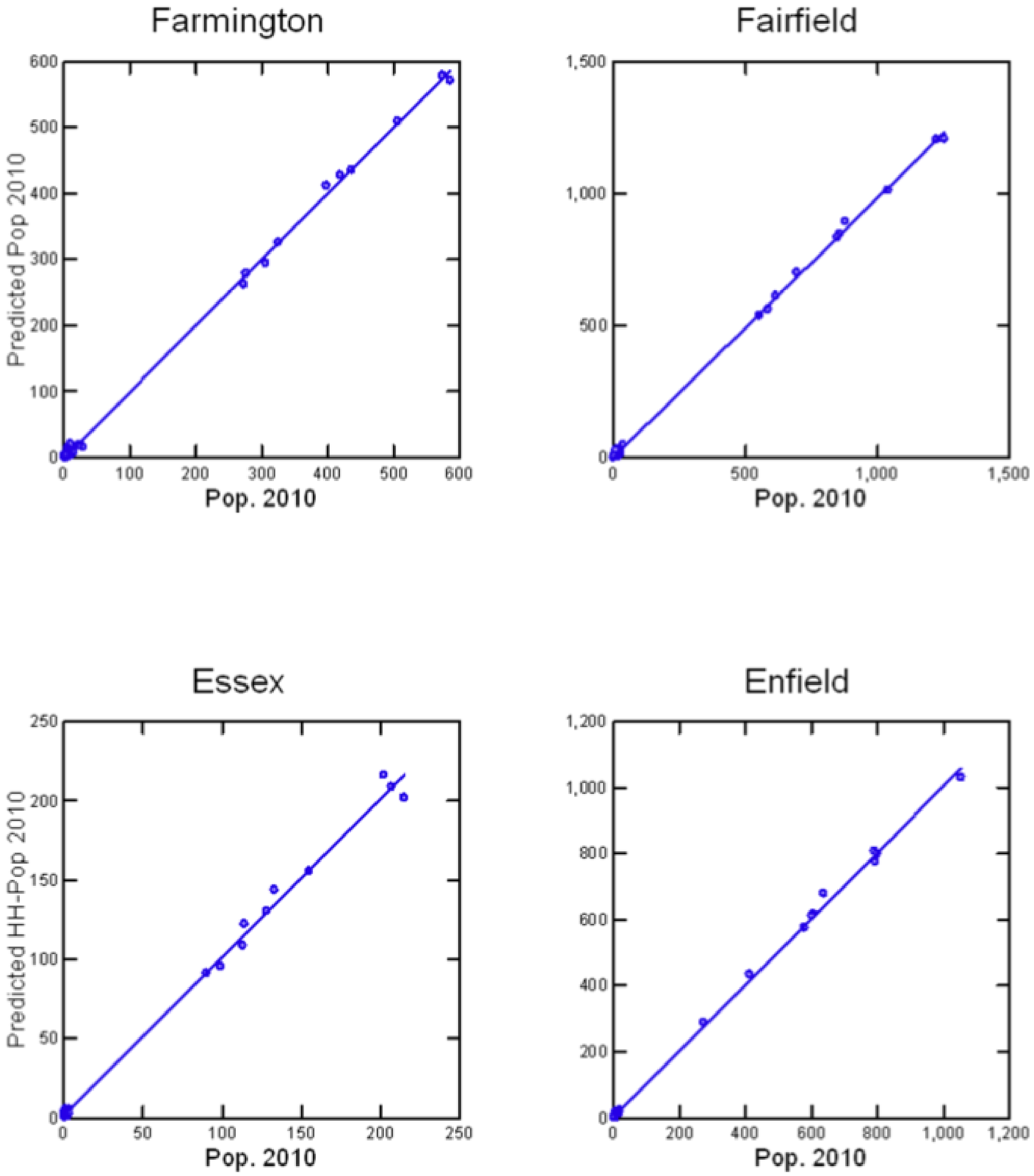
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Farmington

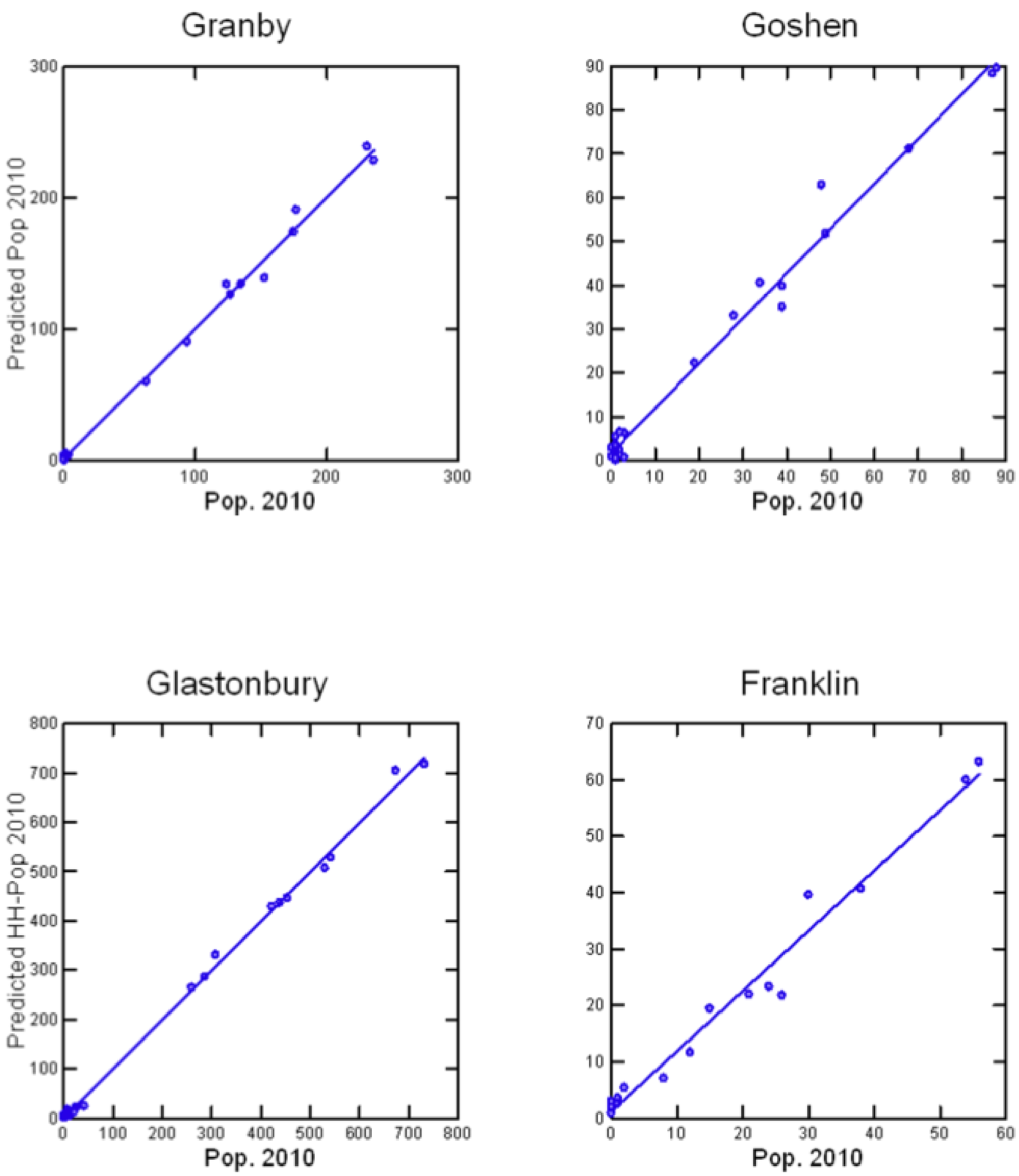
Fairfield

Essex

Enfield

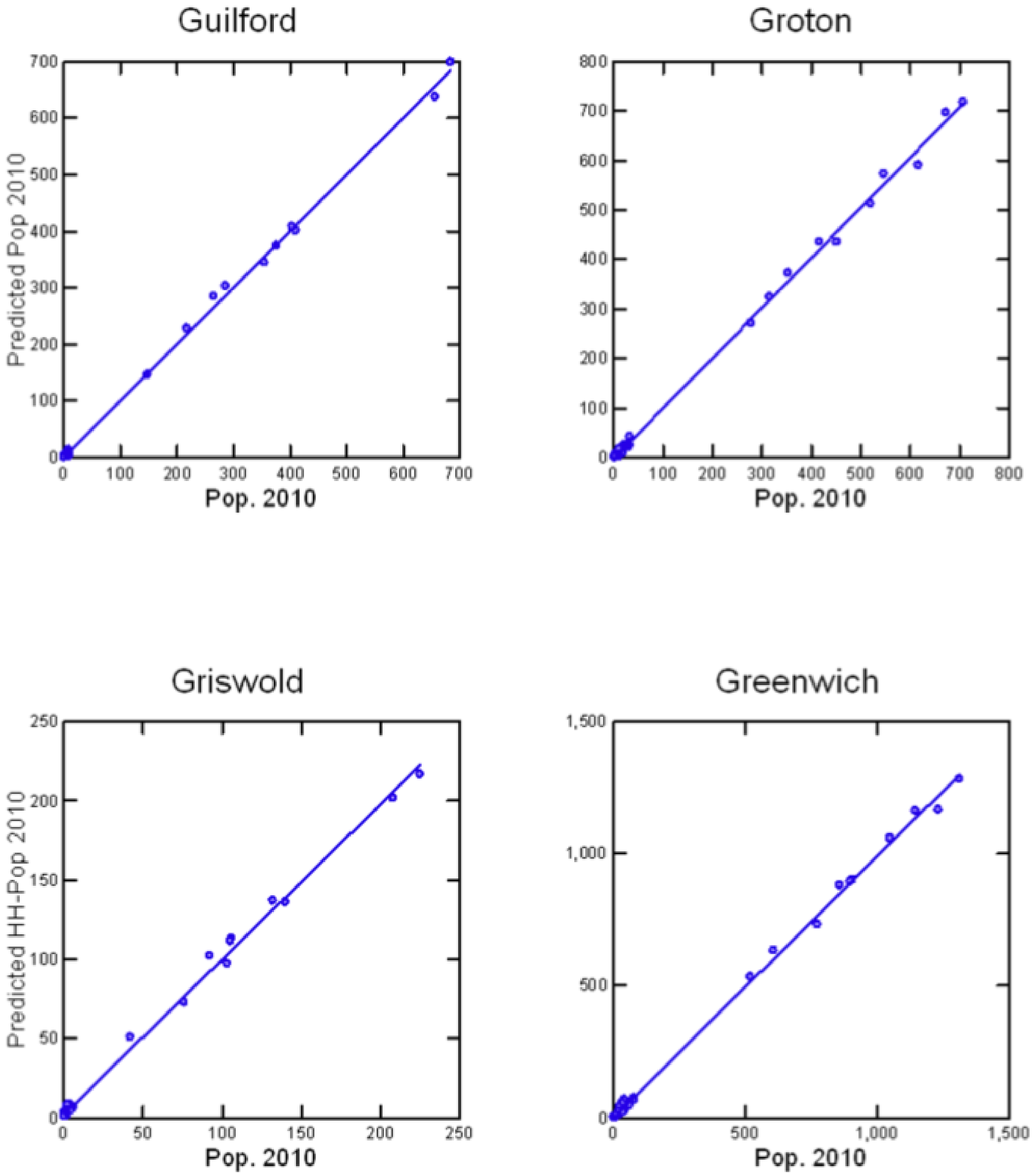
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Guilford

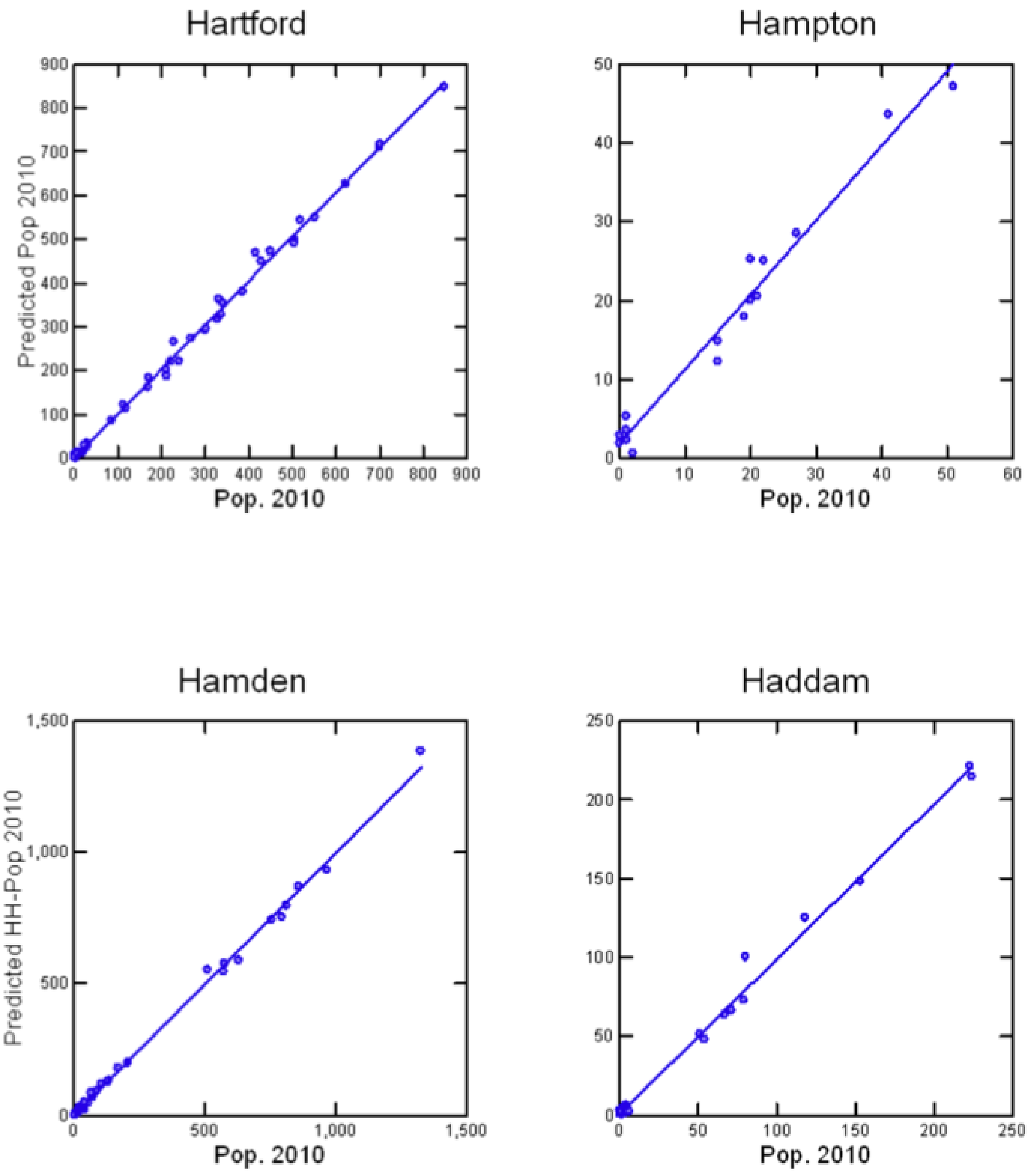
Groton

Griswold

Greenwich

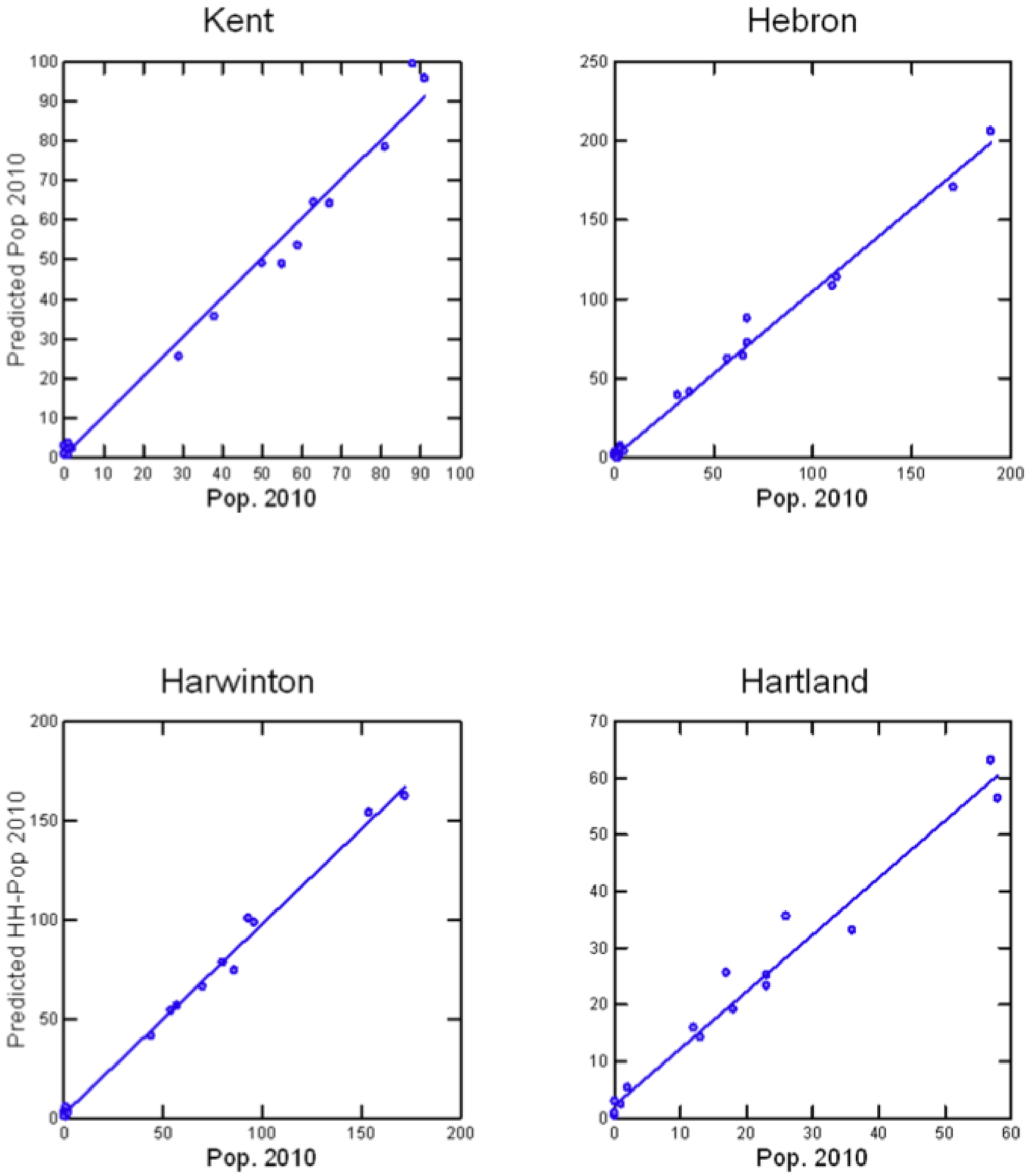
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



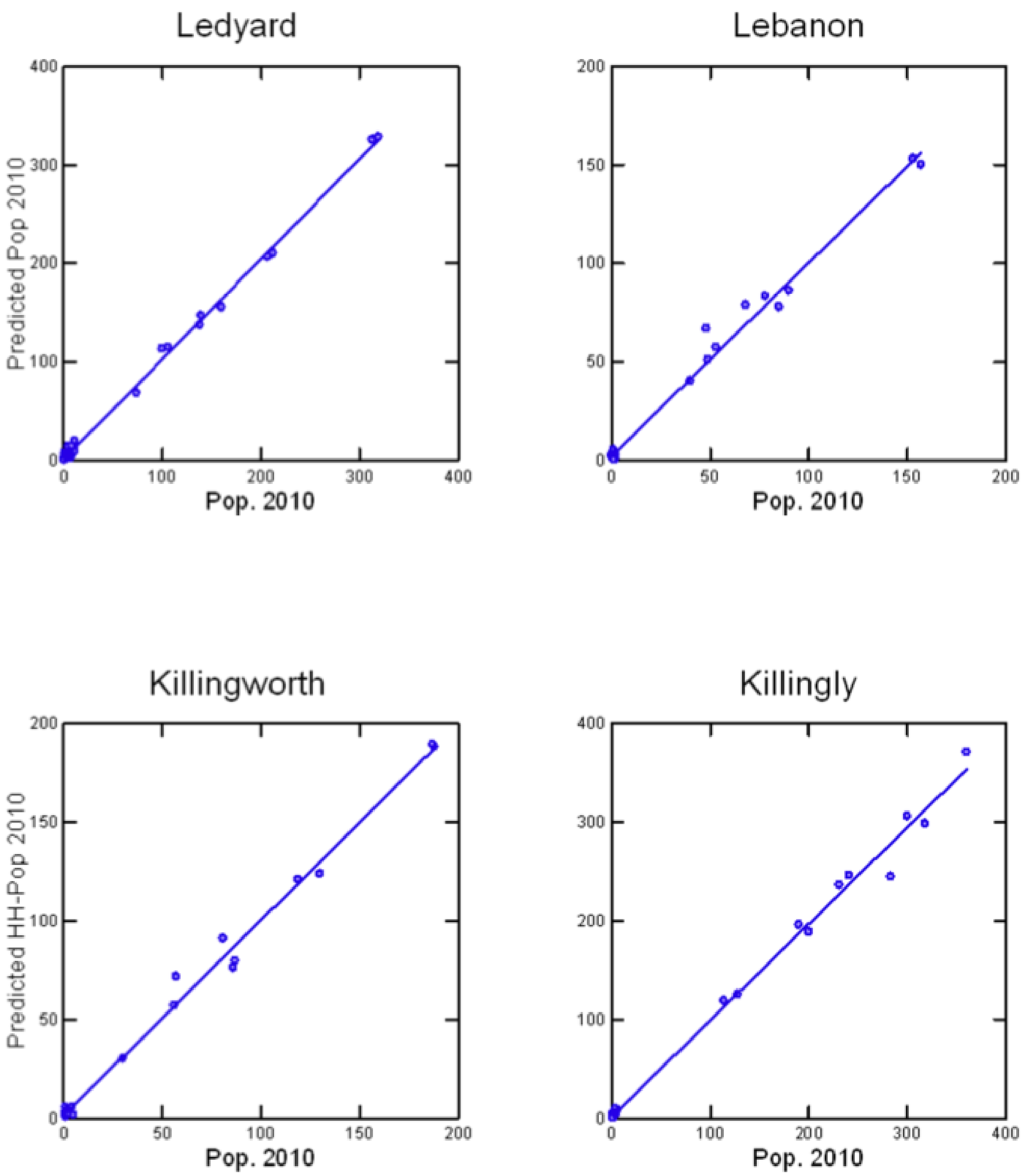
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



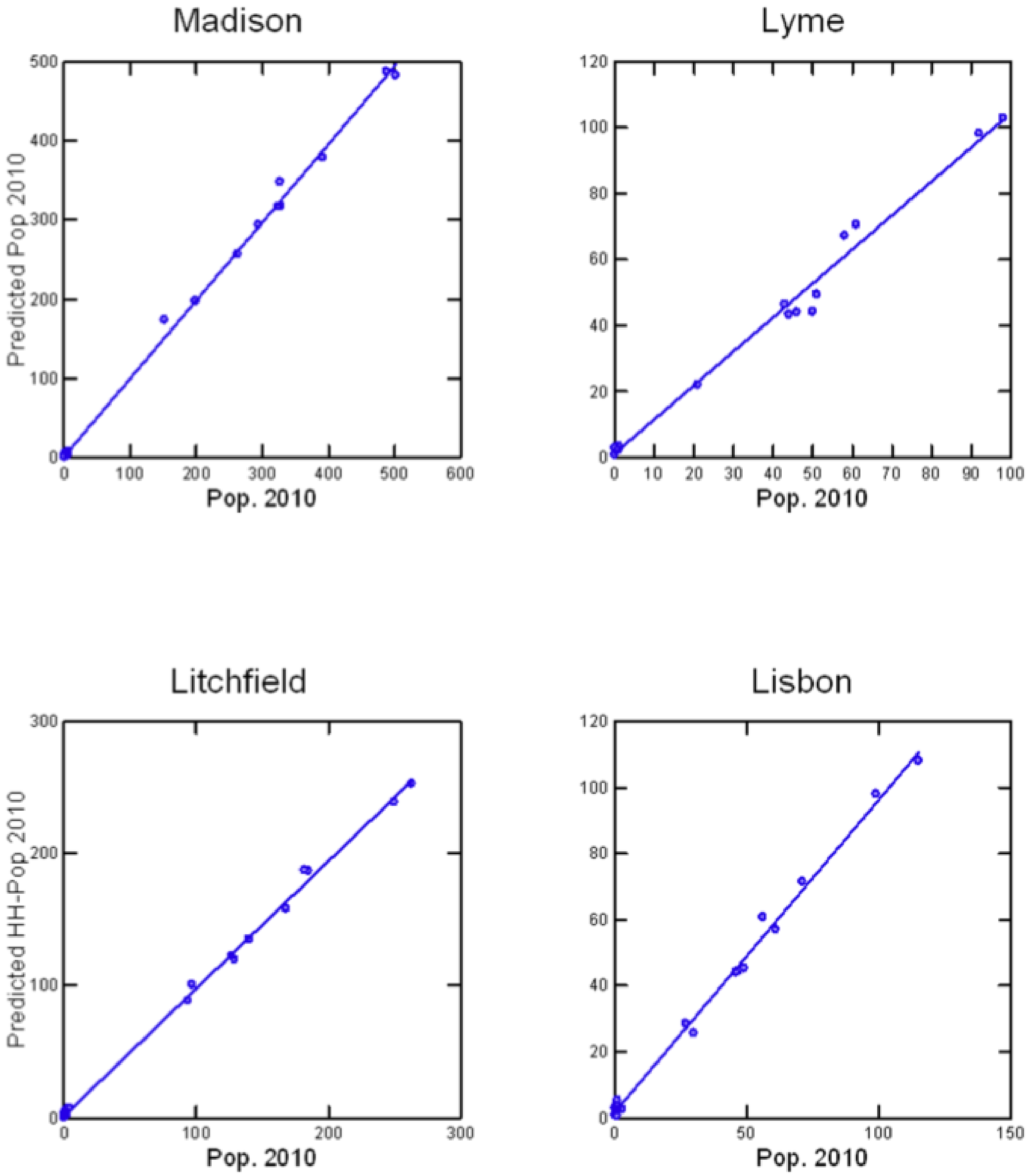
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



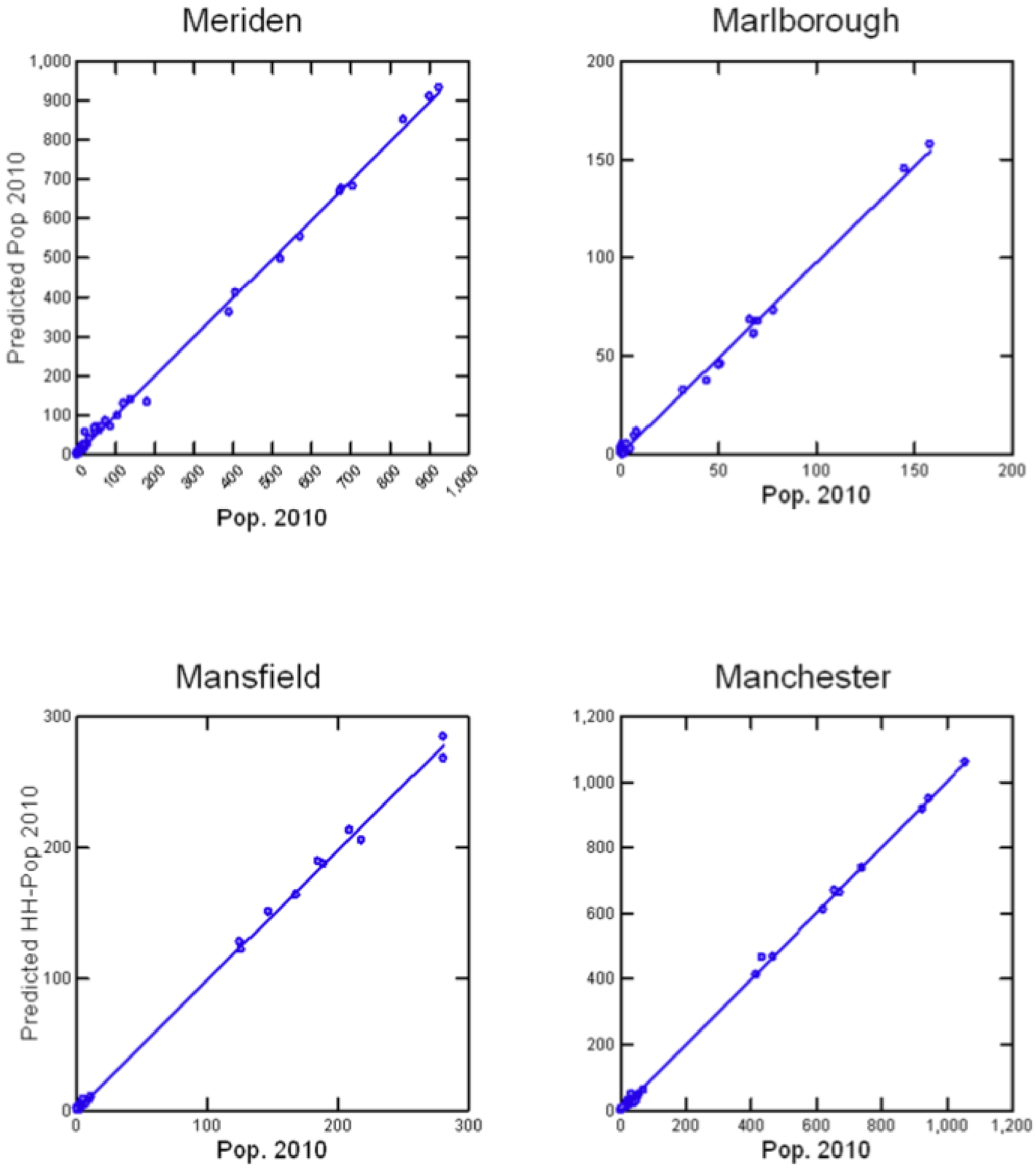
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



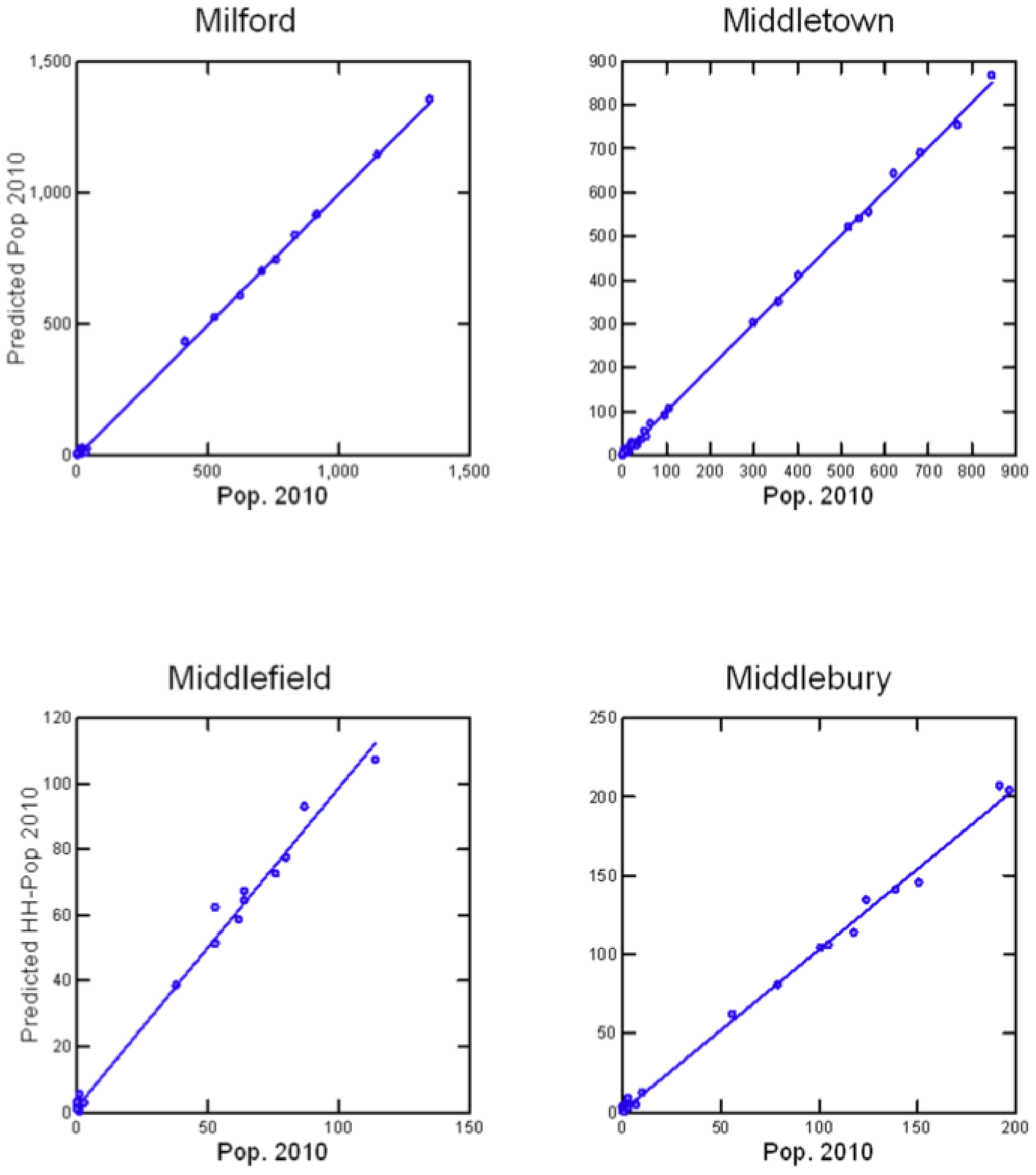
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



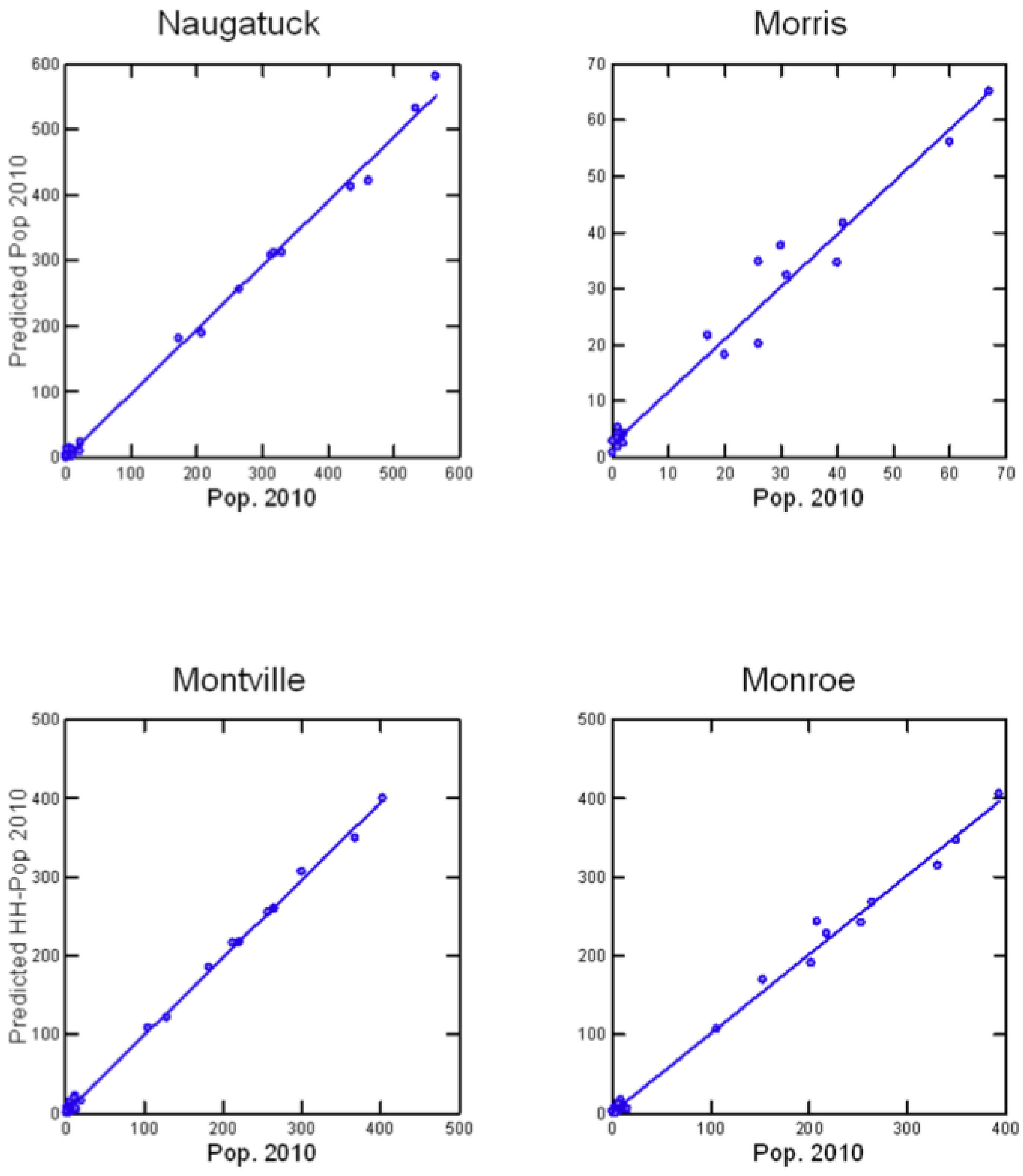
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



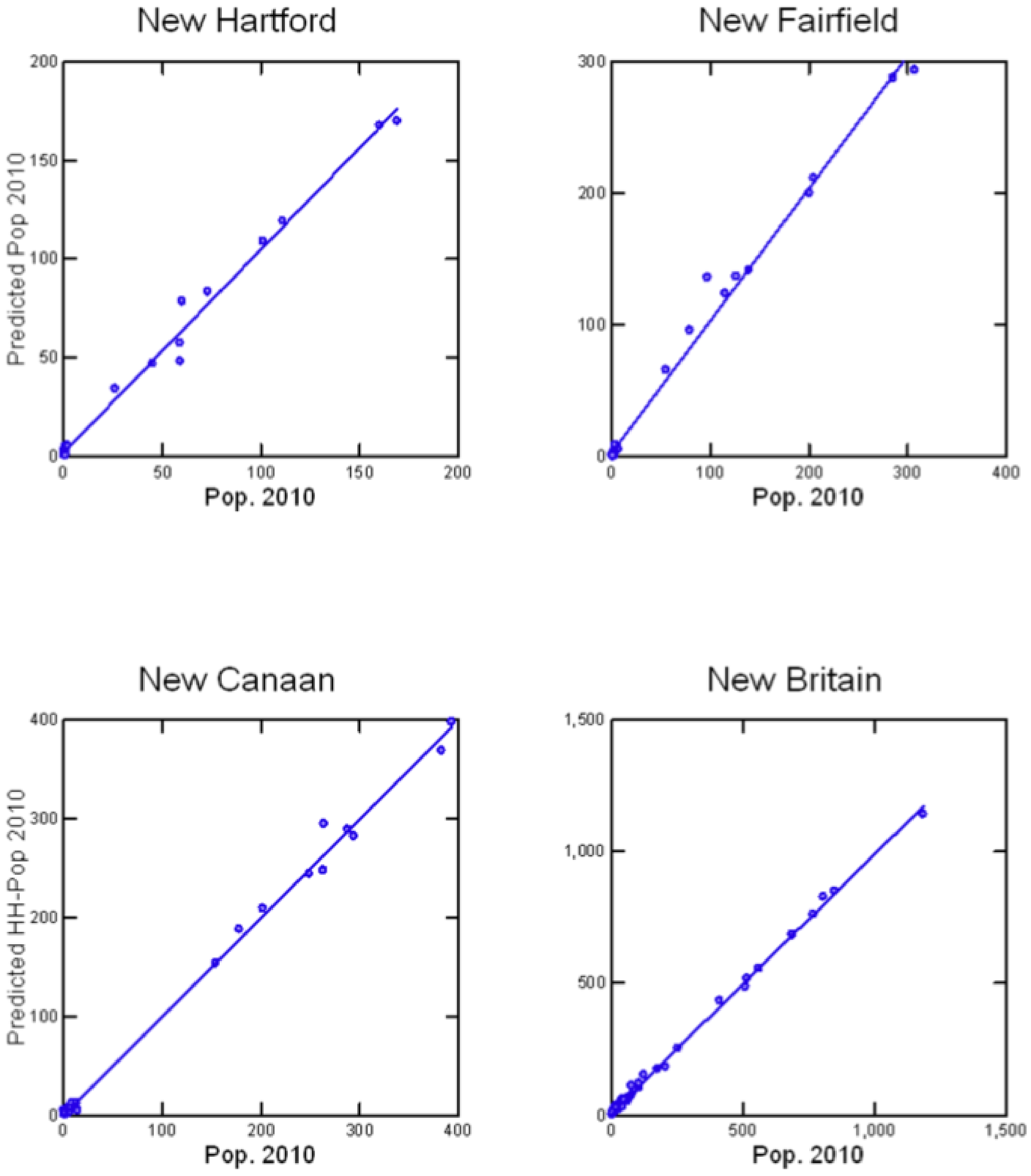
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



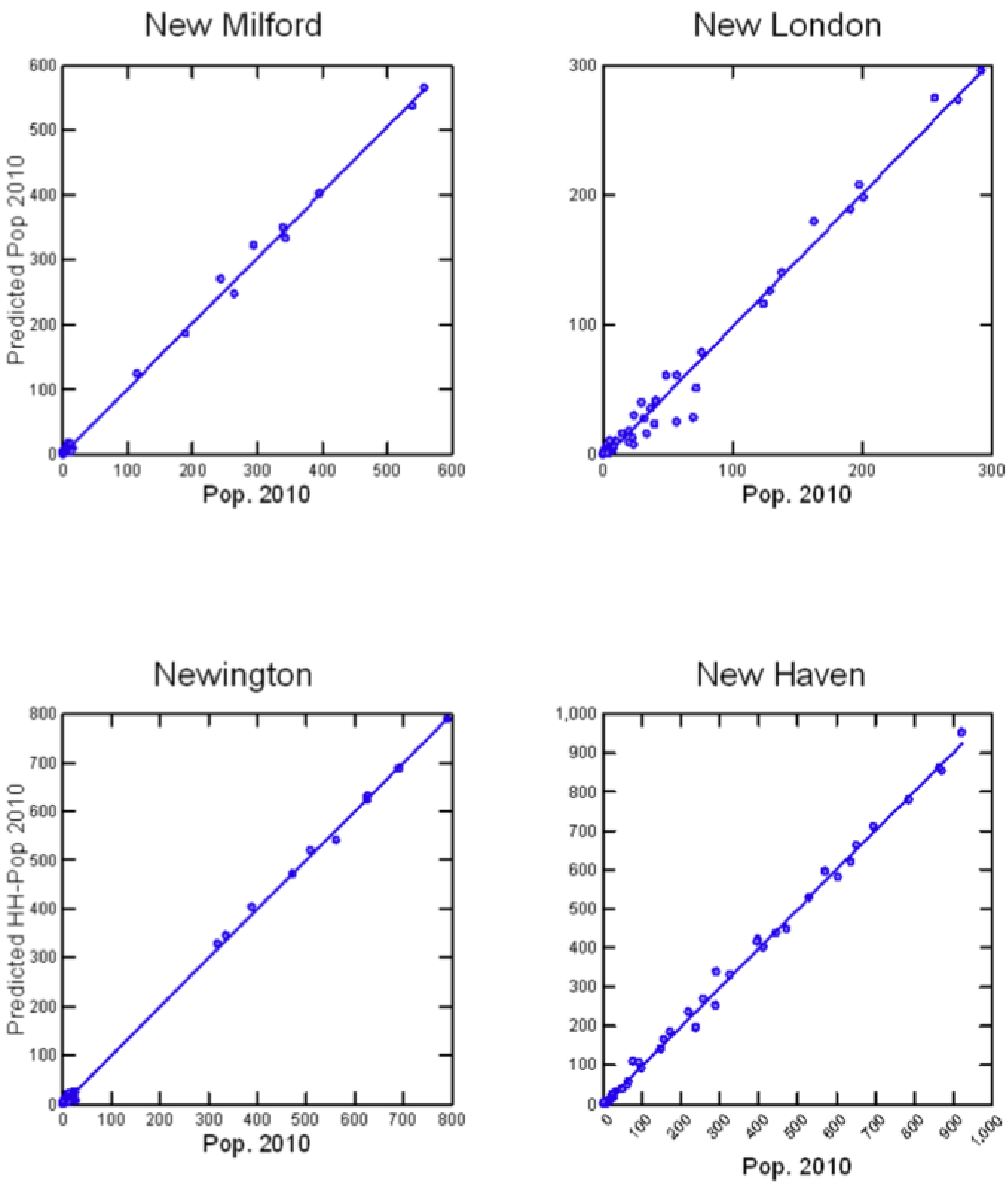
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



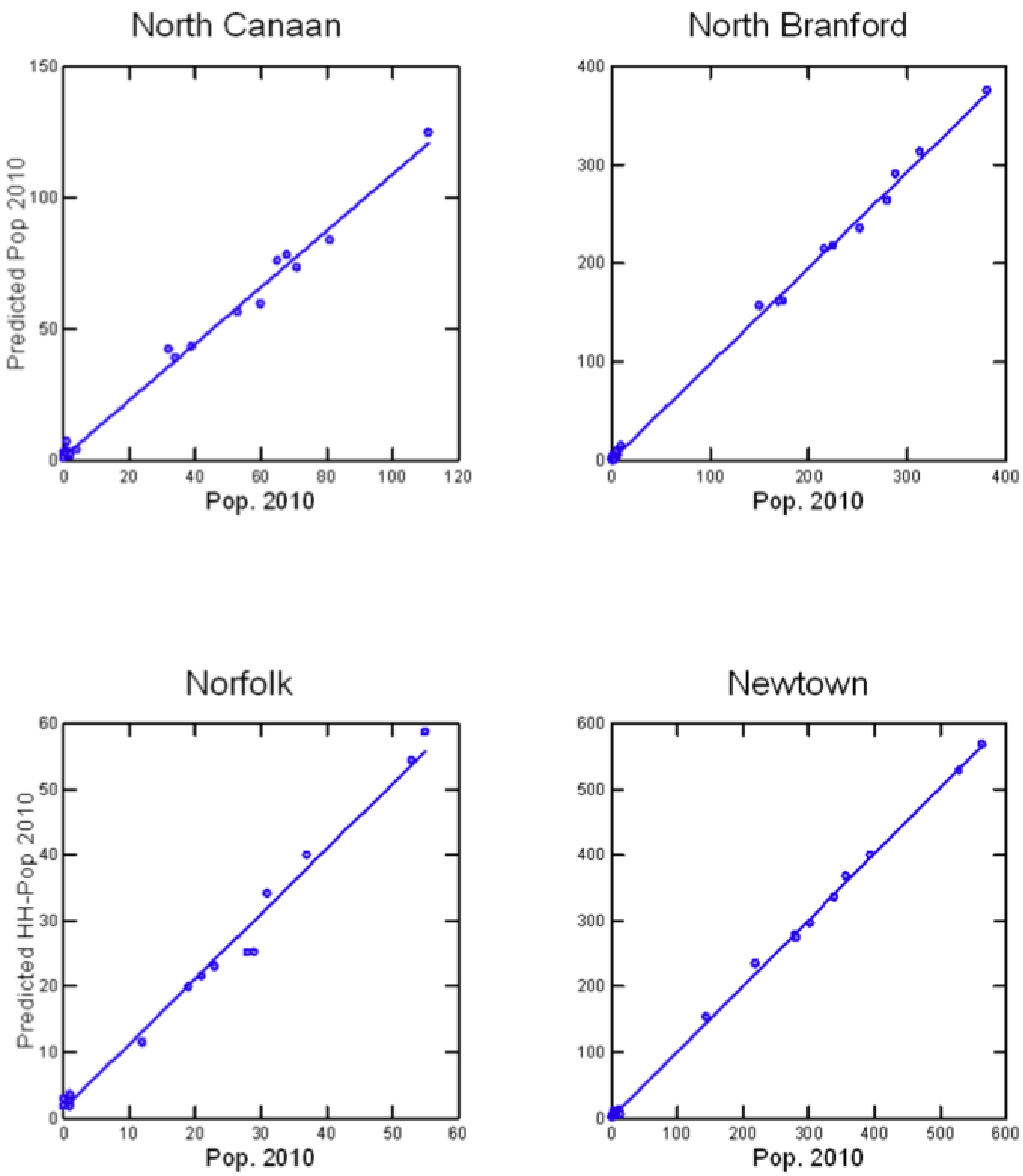
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)


North Canaan

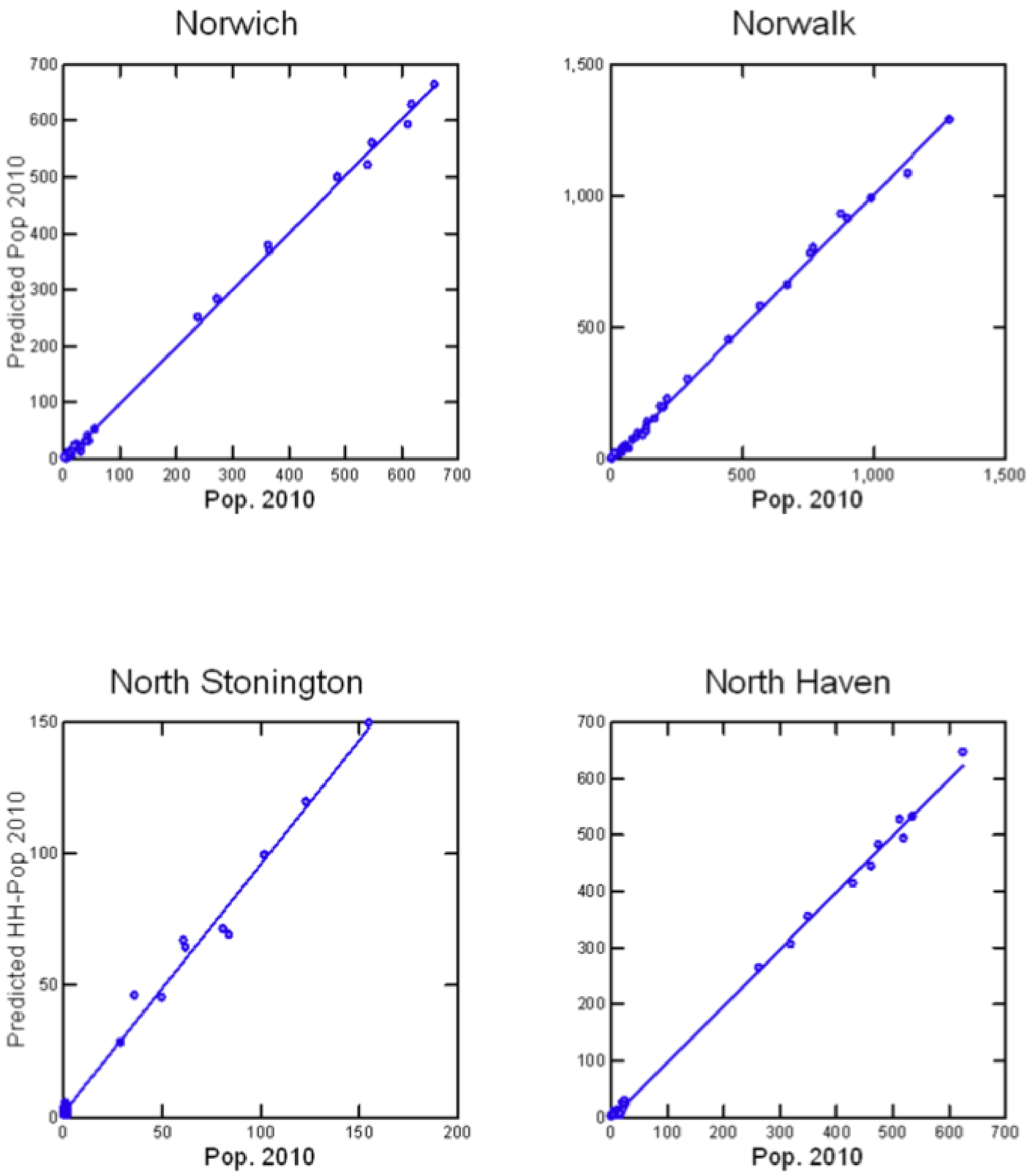
North Branford

Norfolk

Newtown

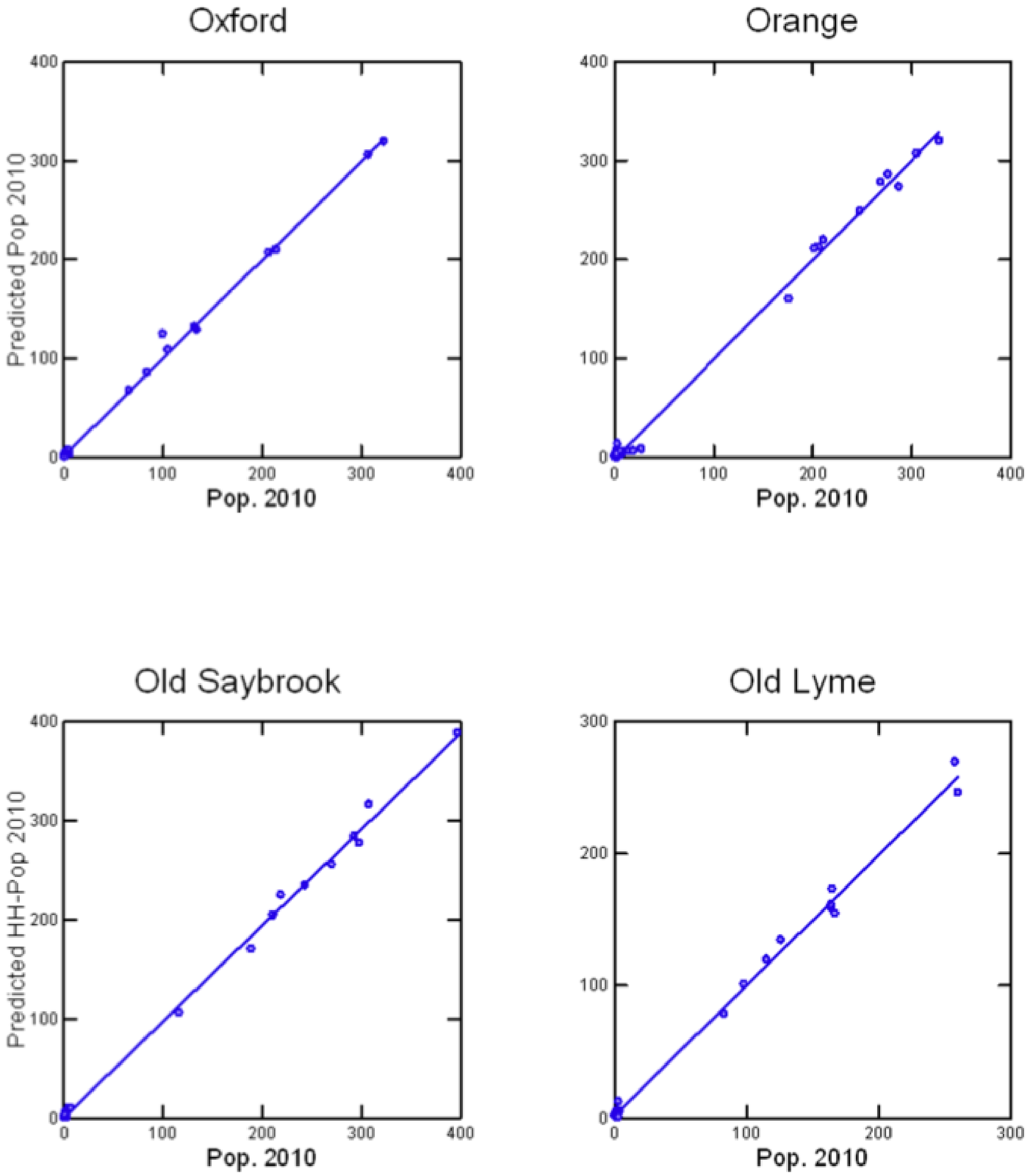
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



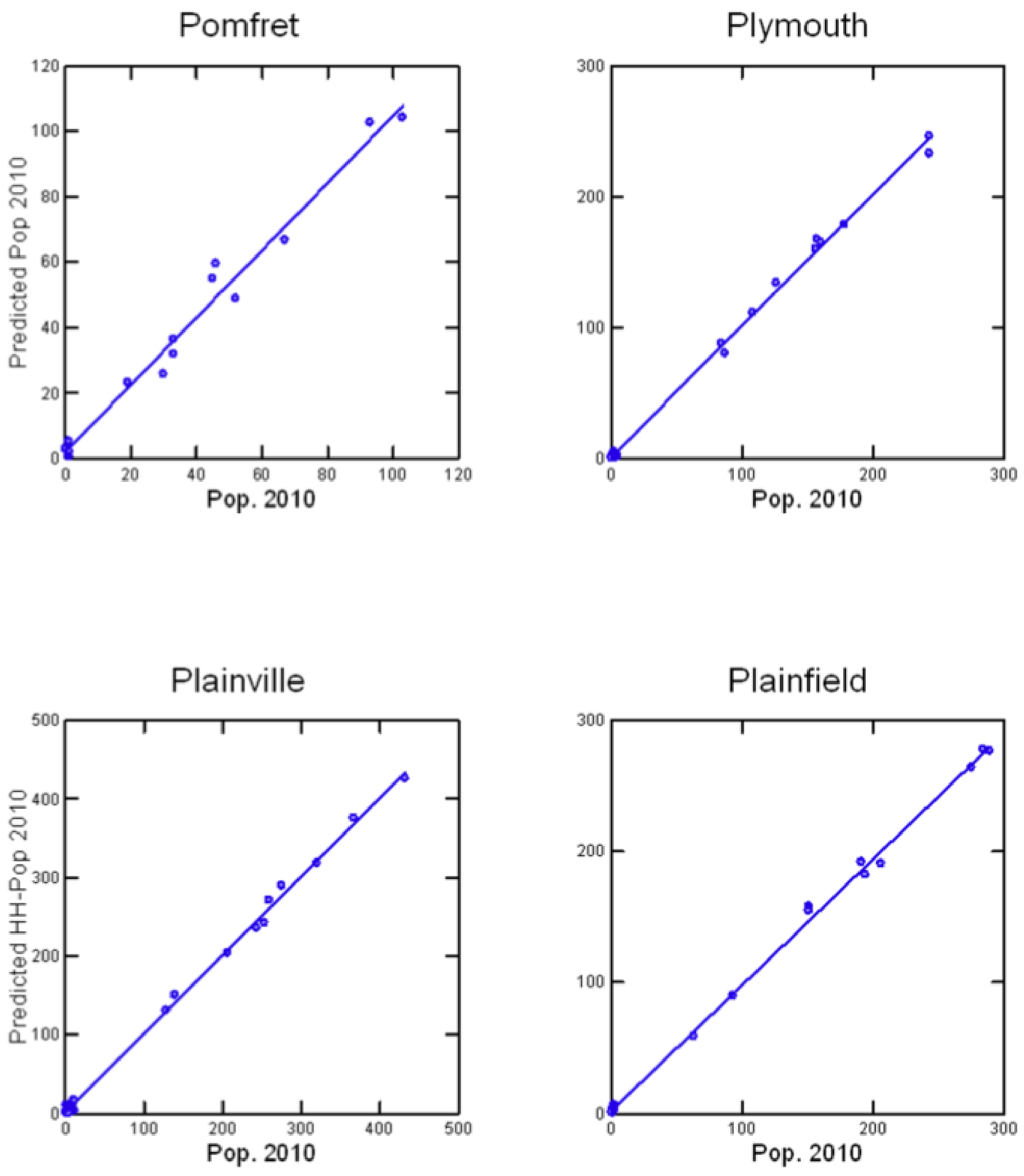
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pomfret

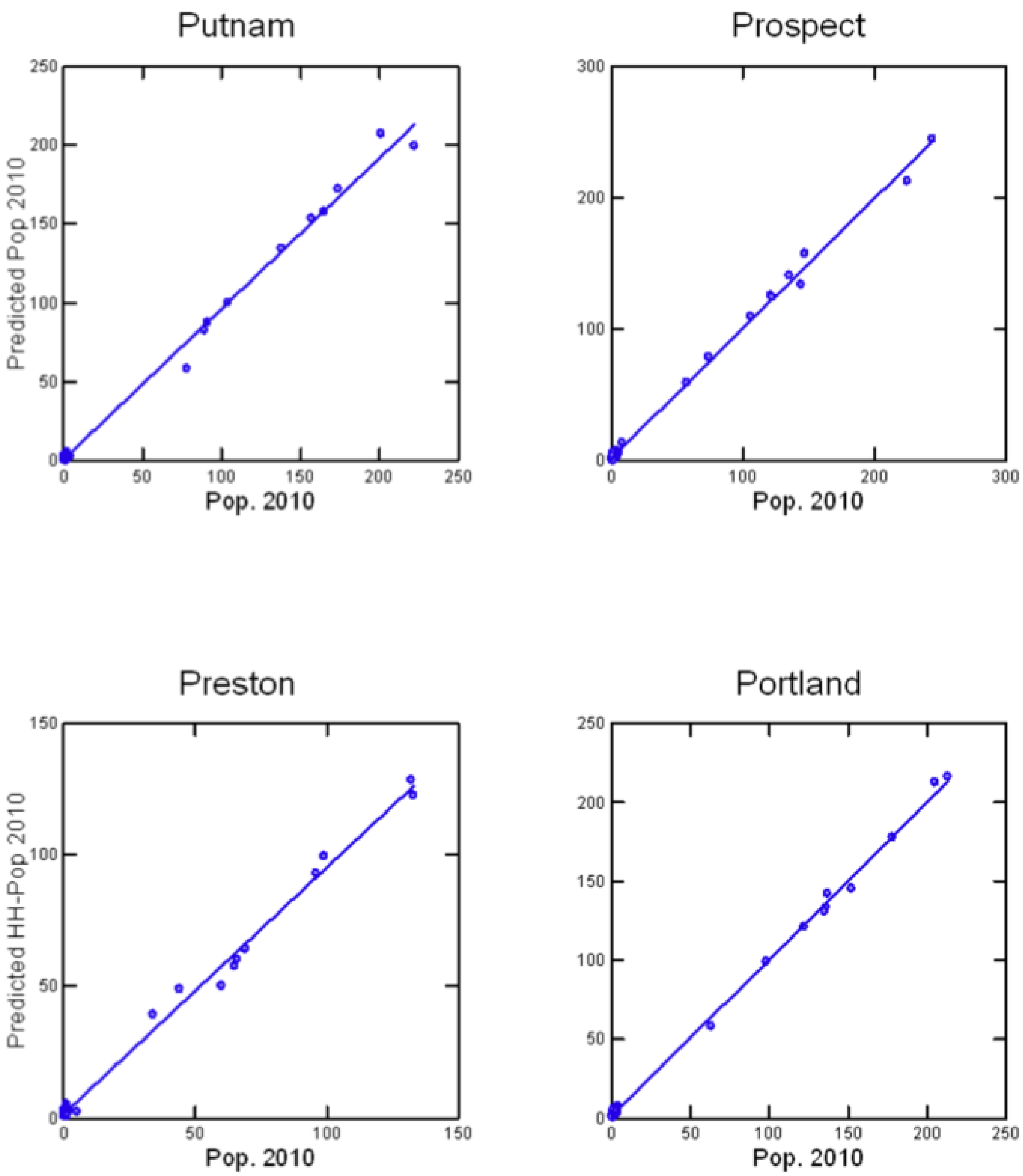
Plymouth

Plainville

Plainfield

Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)


Putnam

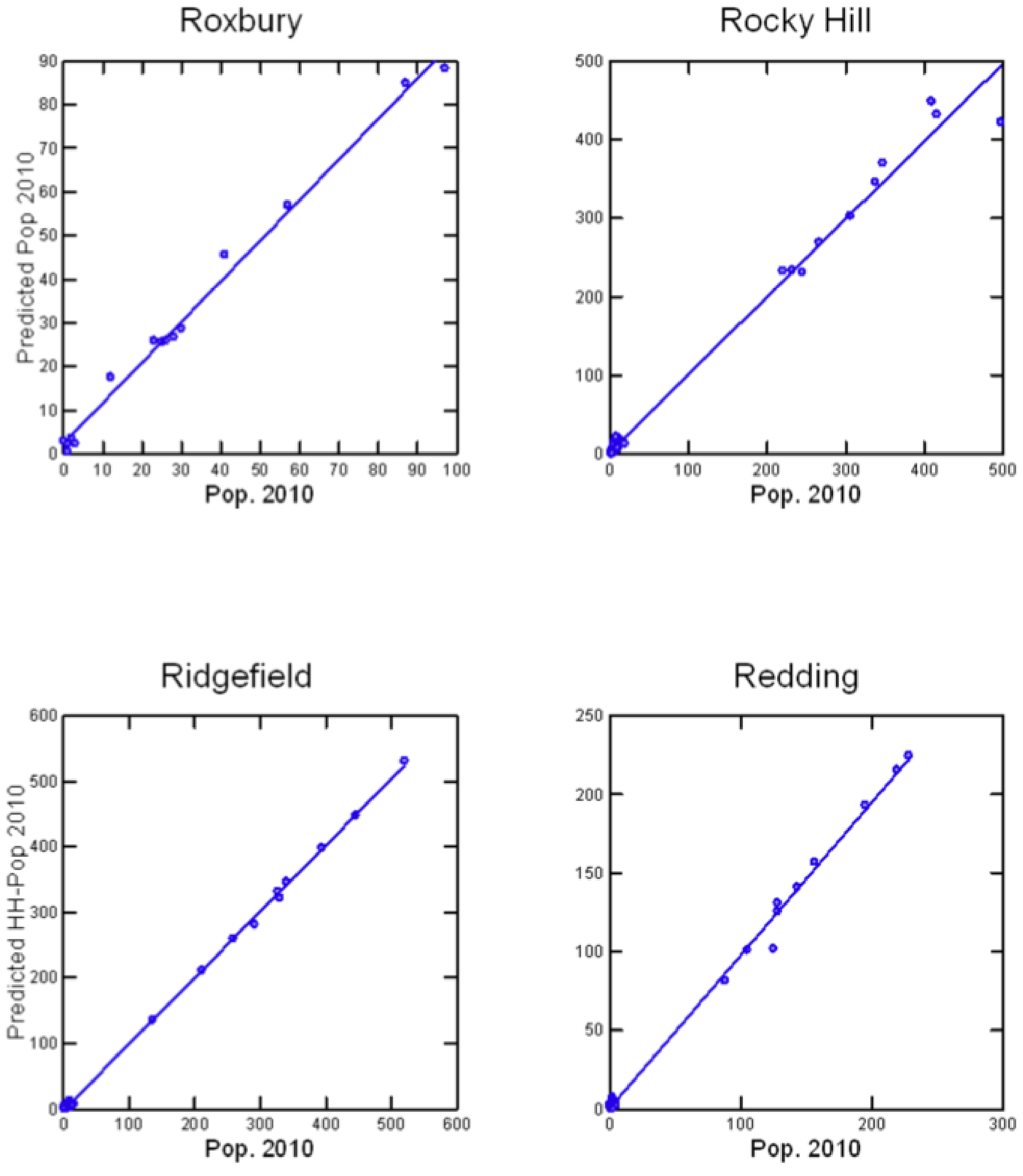
Prospect

Preston

Portland

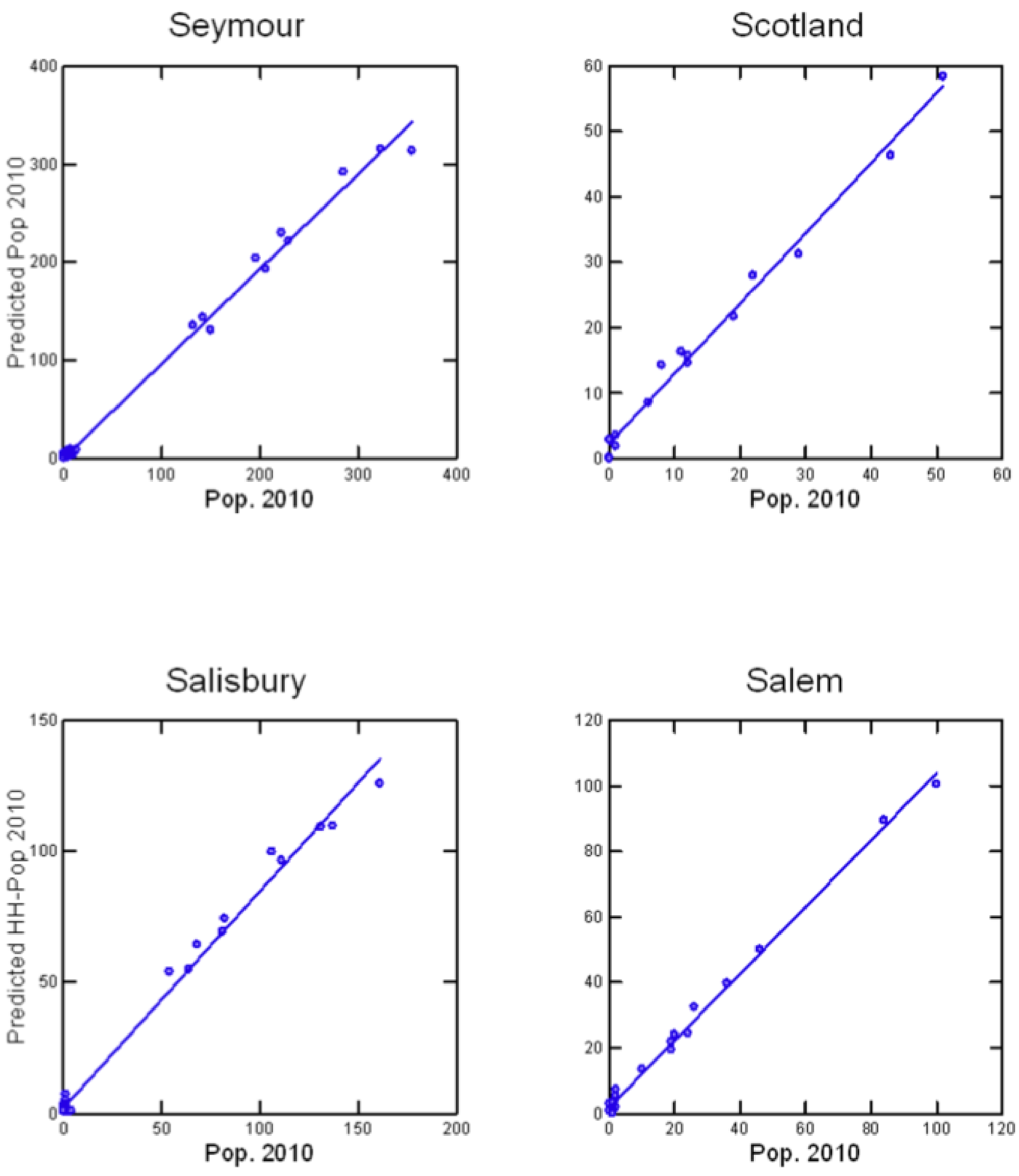
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



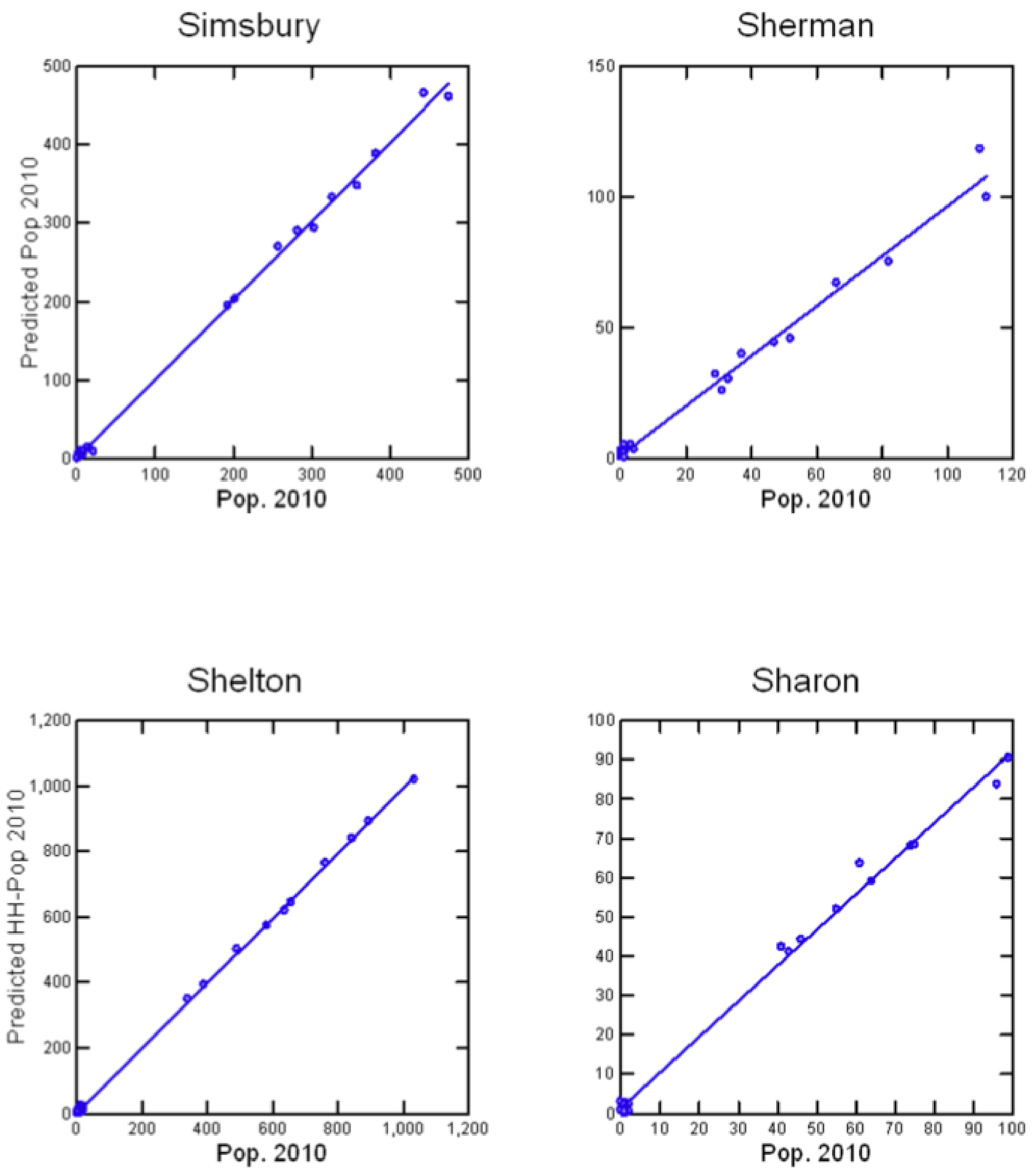
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



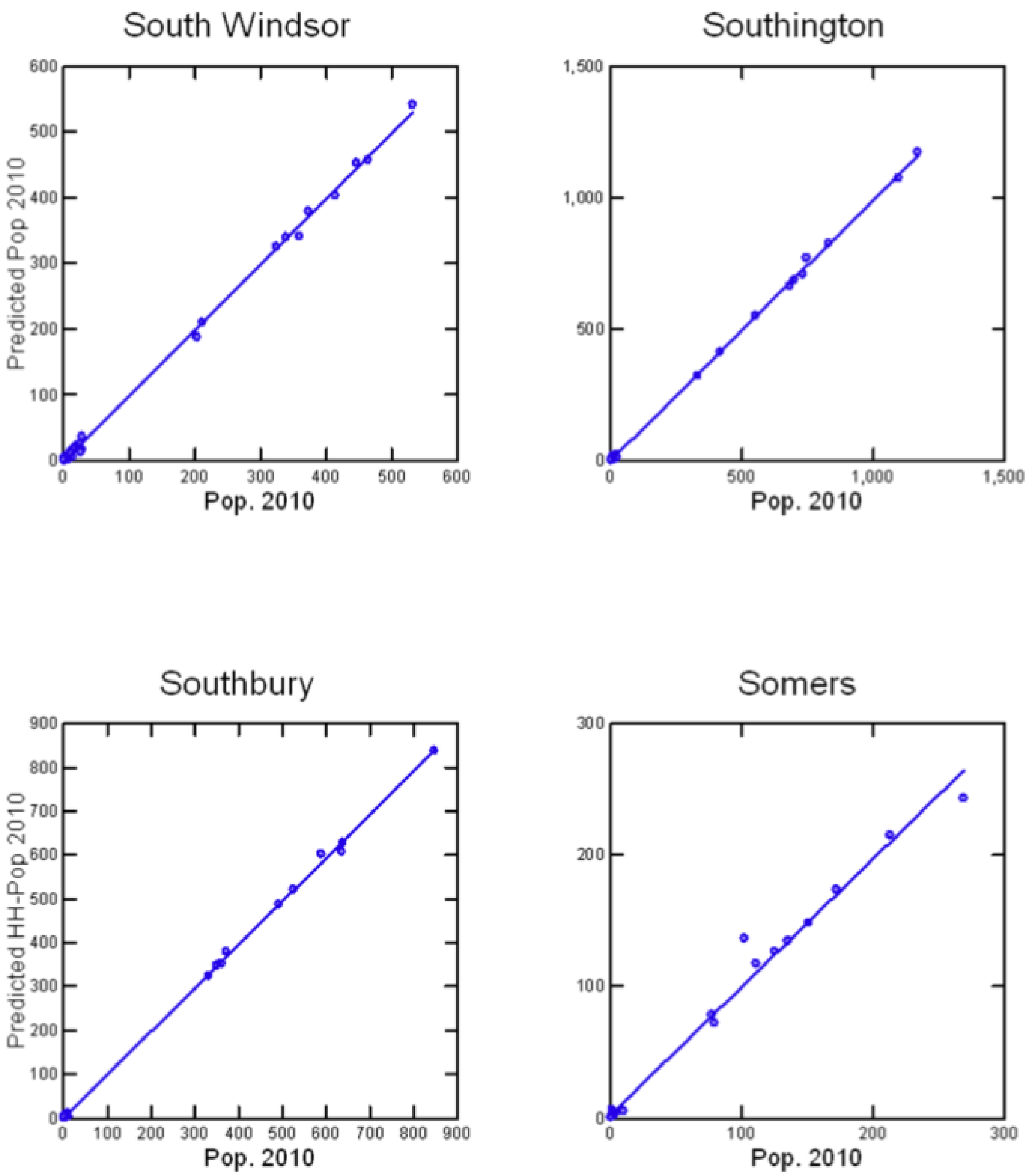
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



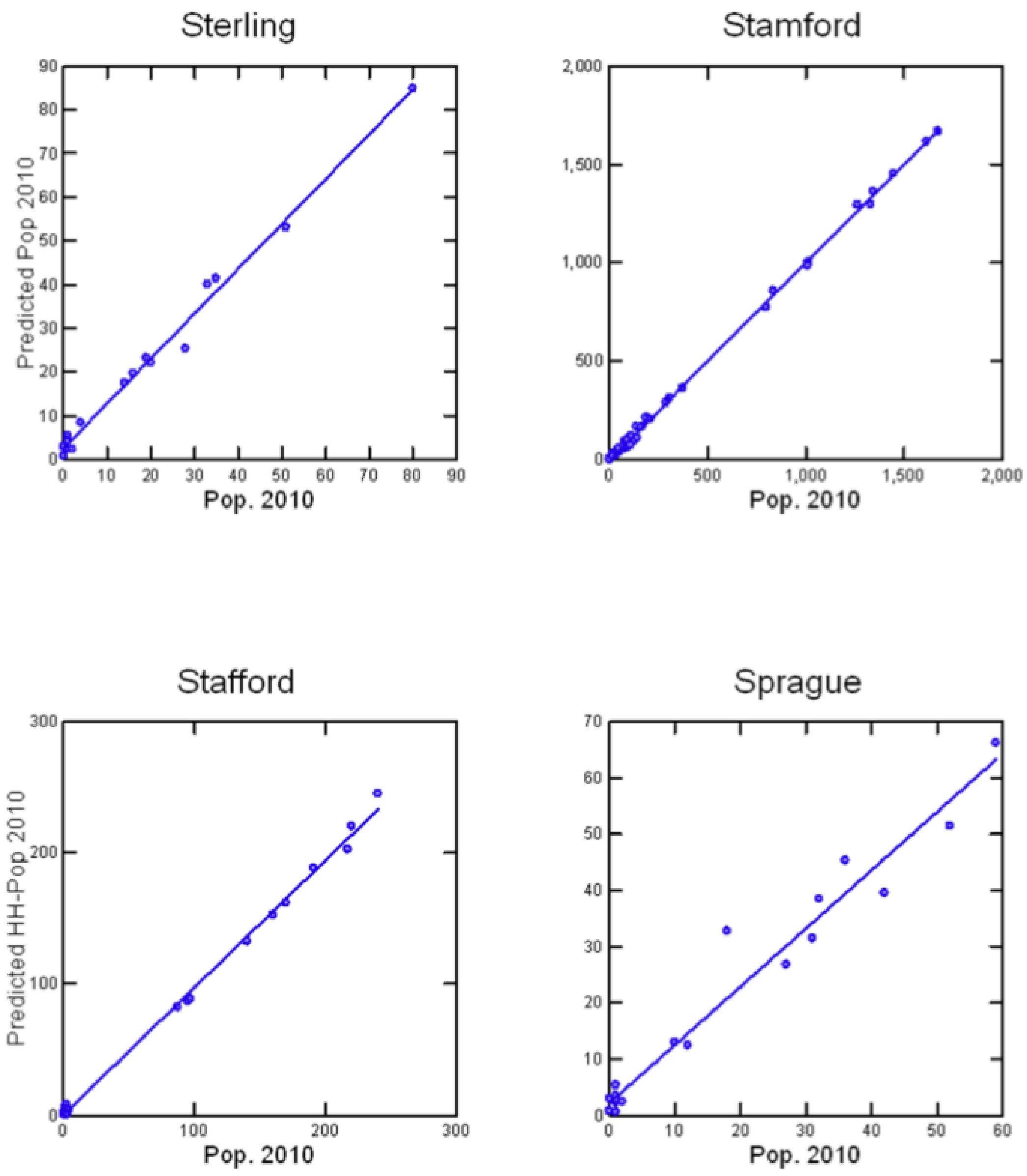
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Sterling

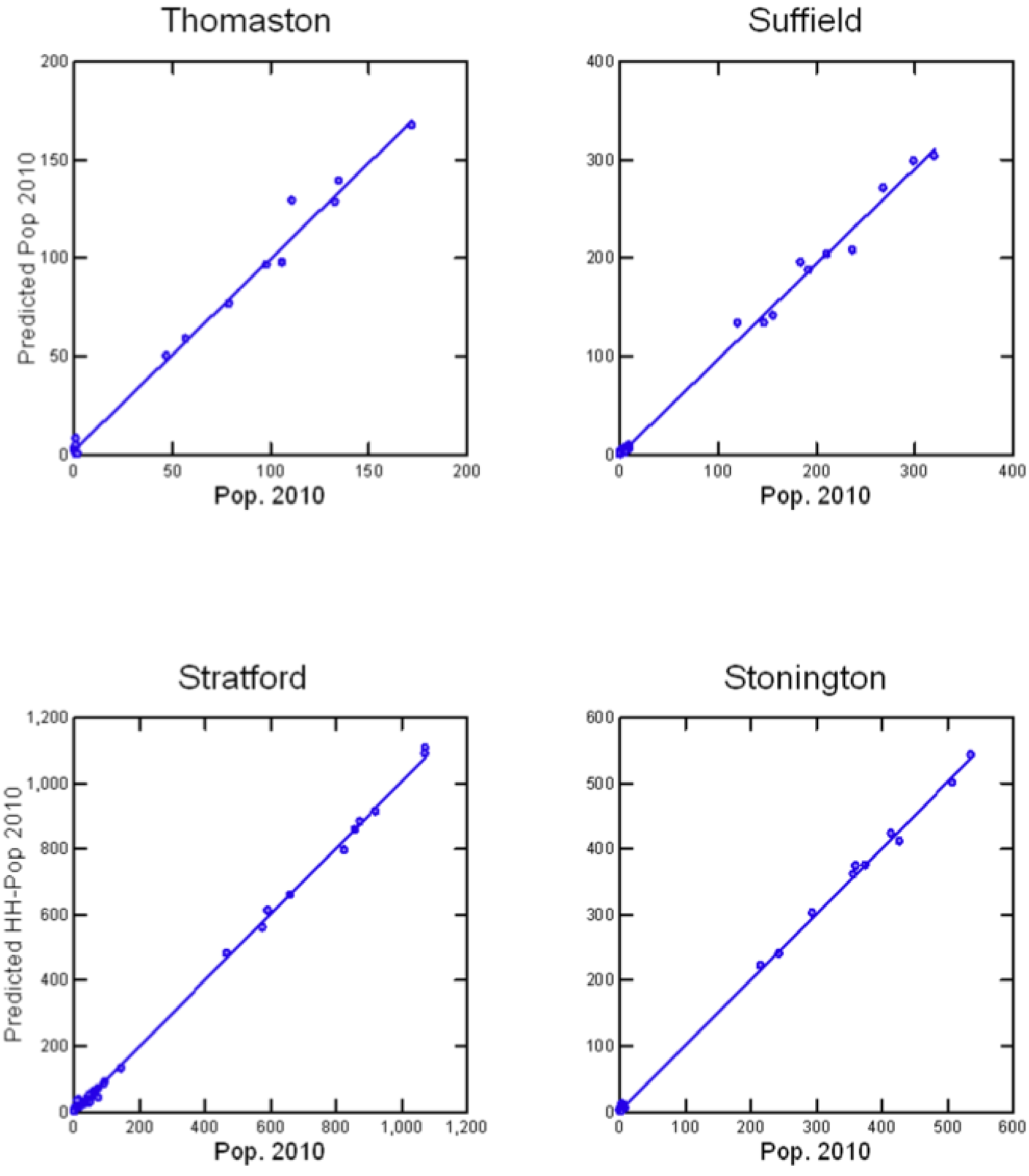
Stamford

Stafford

Sprague

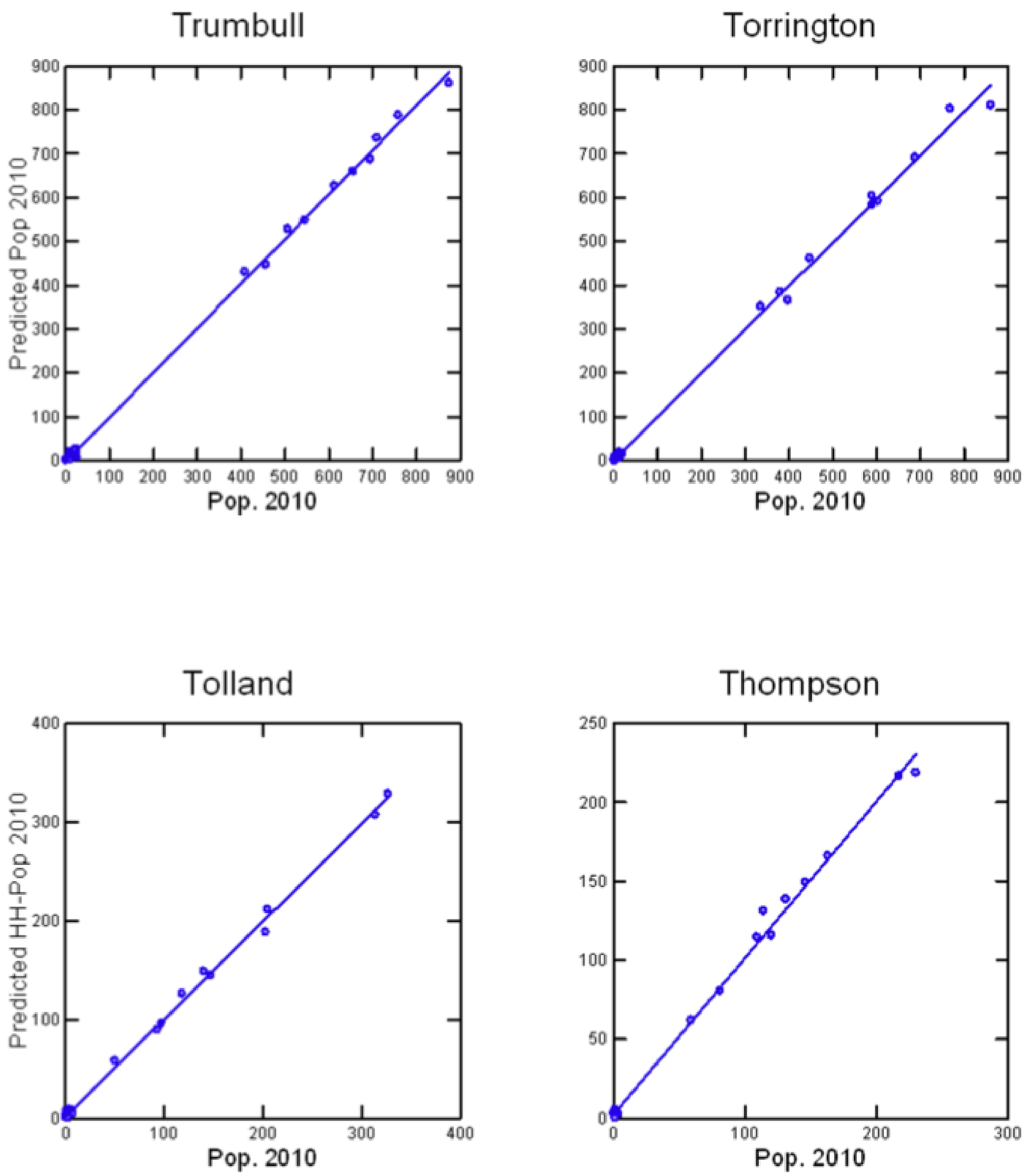
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Trumbull

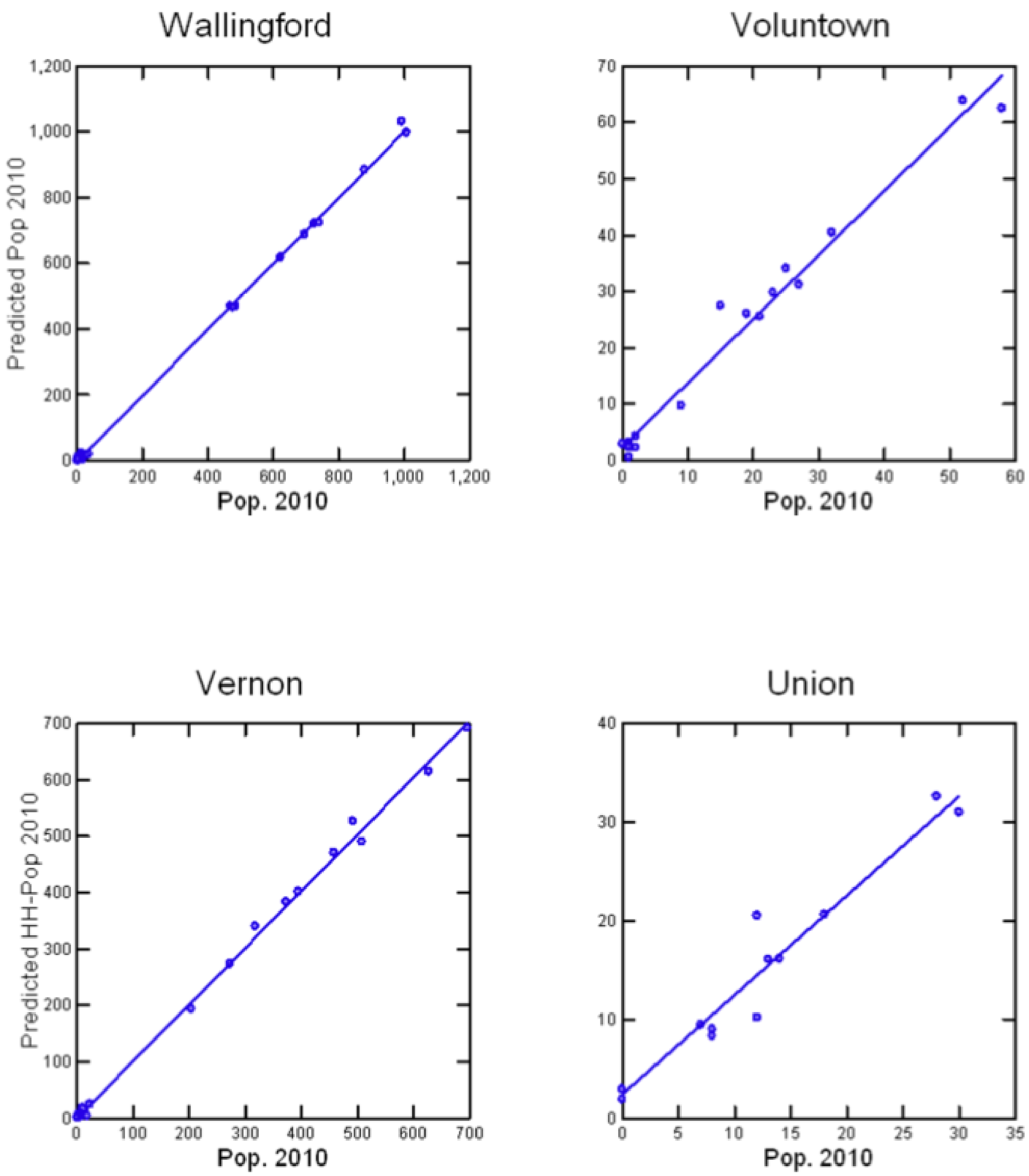
Torrington

Tolland

Thompson

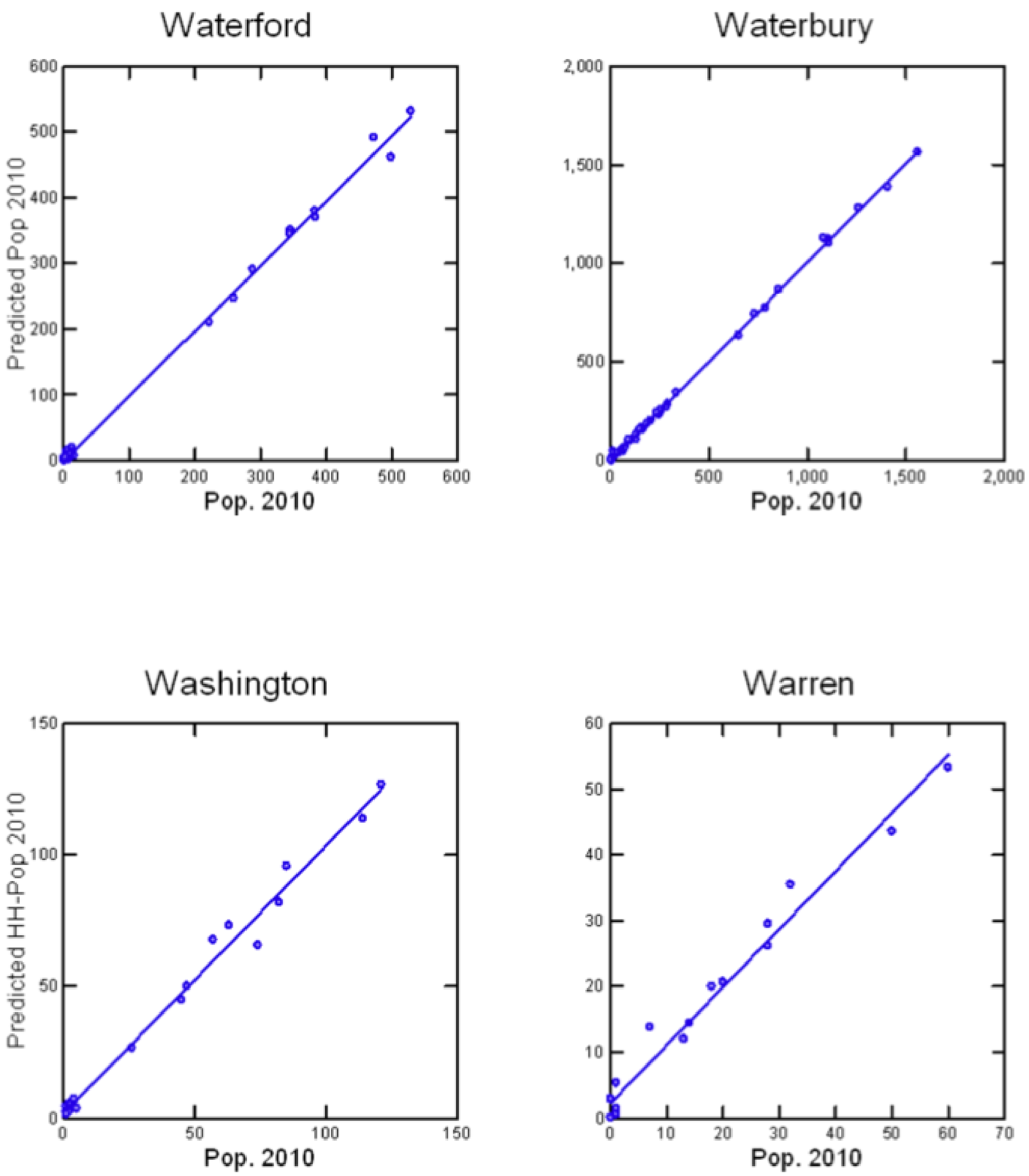
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



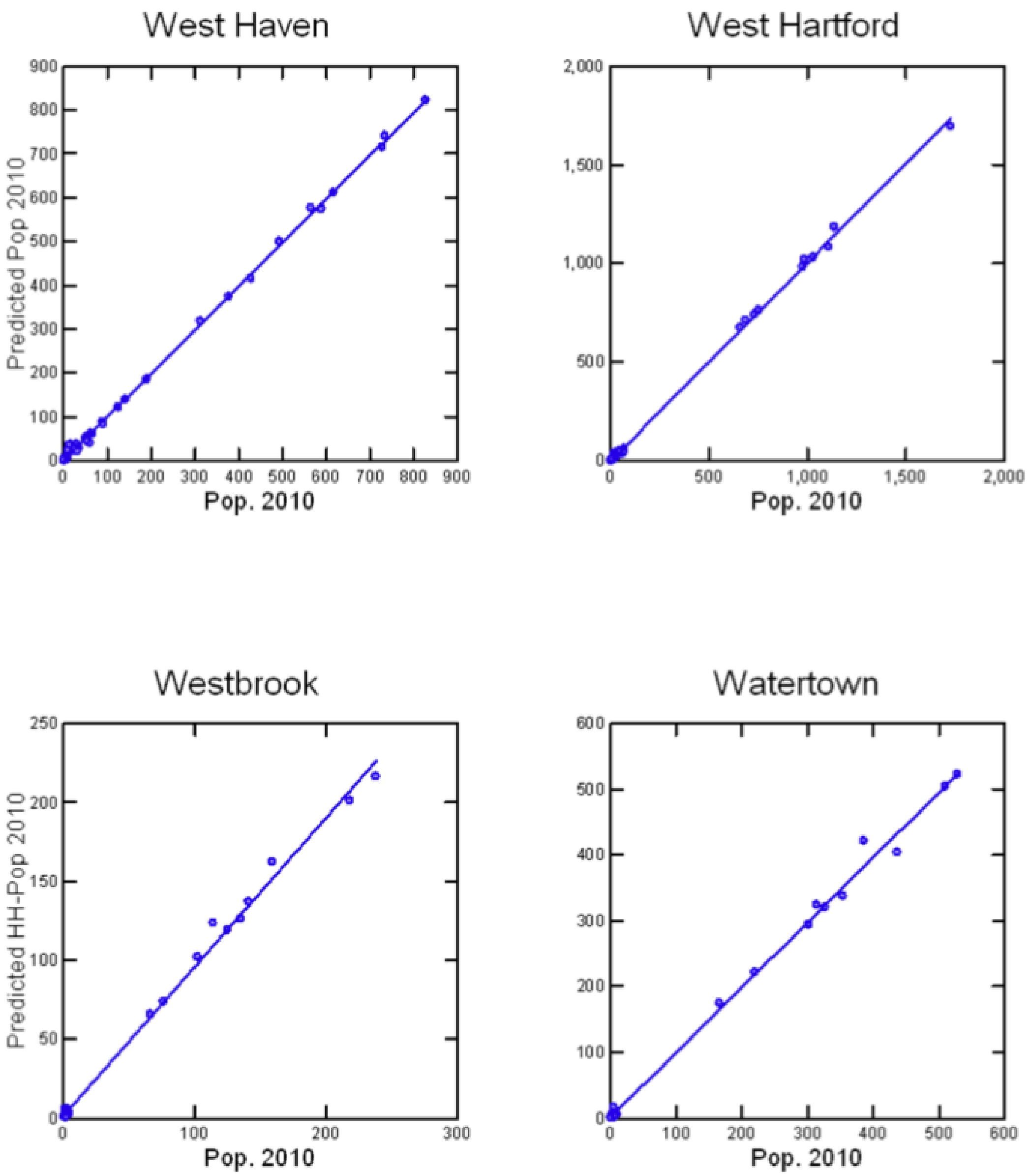
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



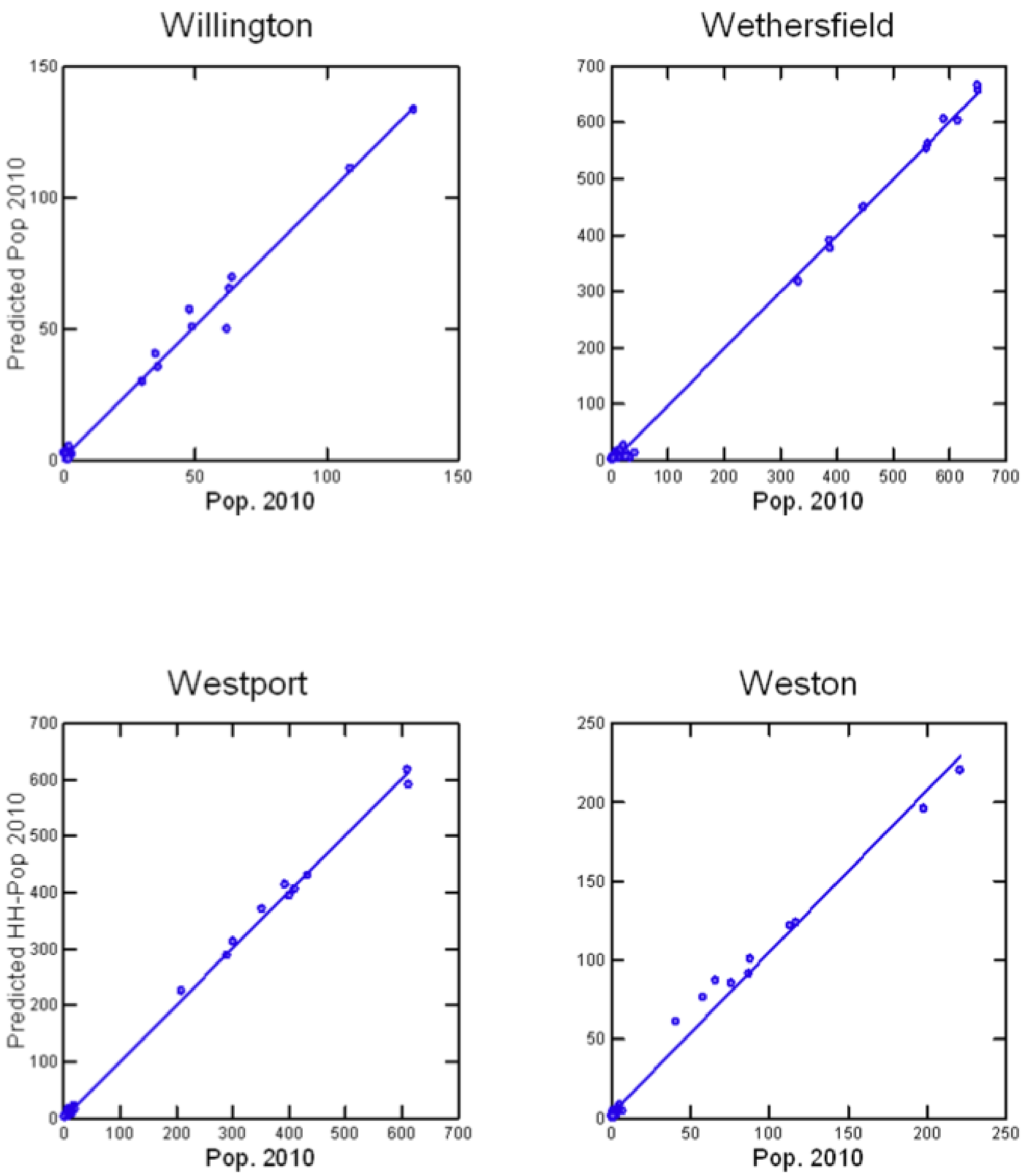
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



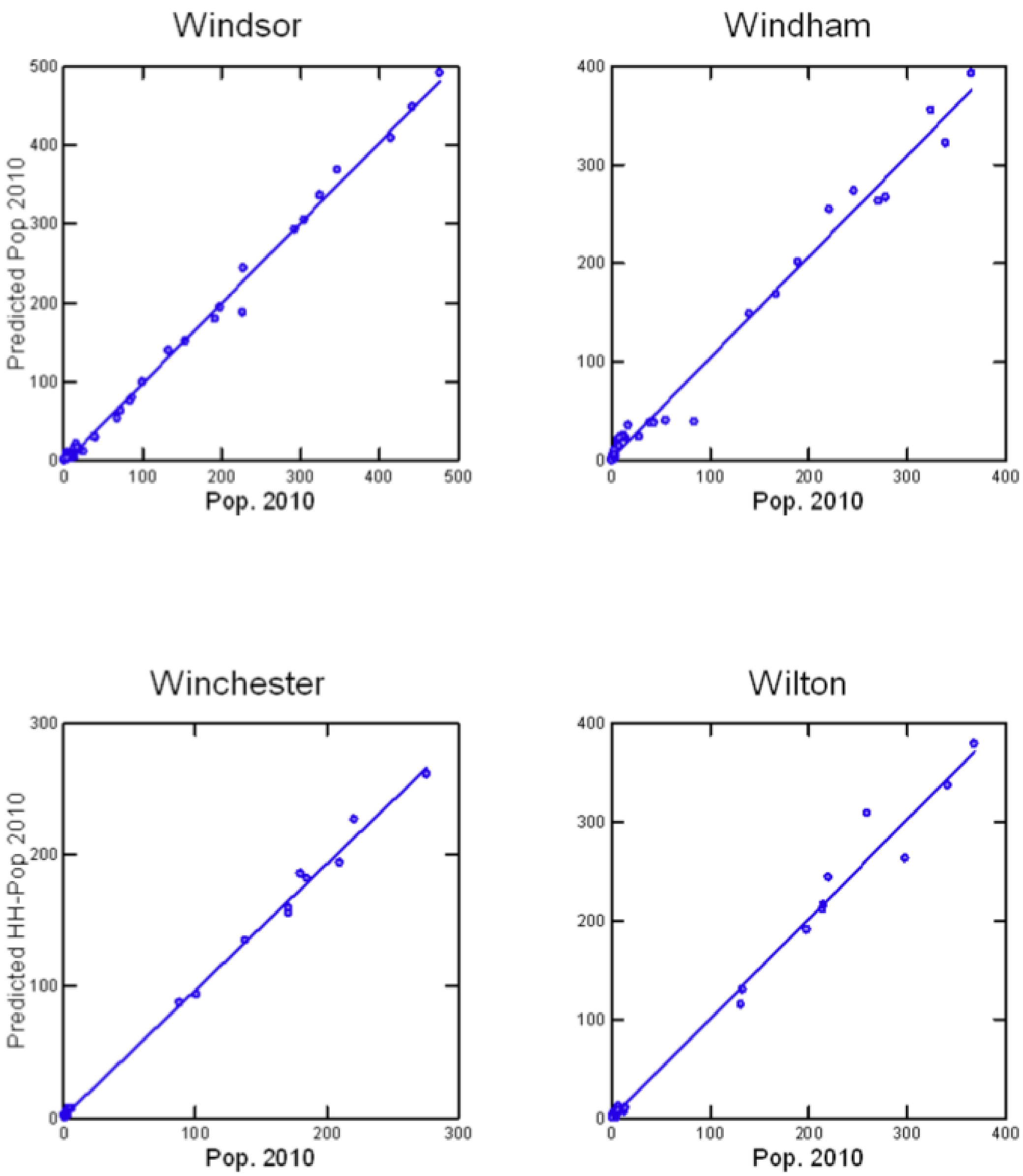
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



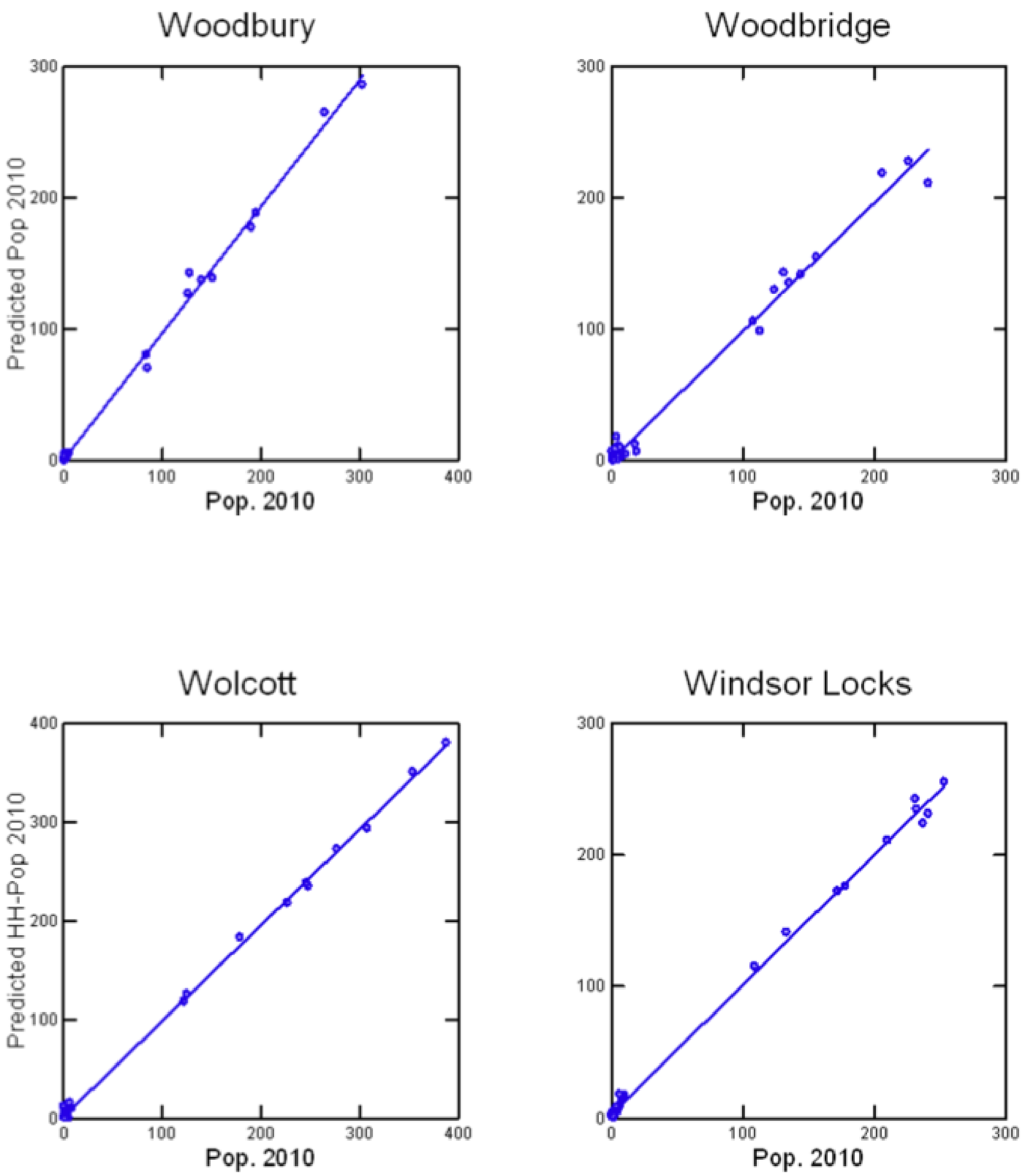
Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Woodbury

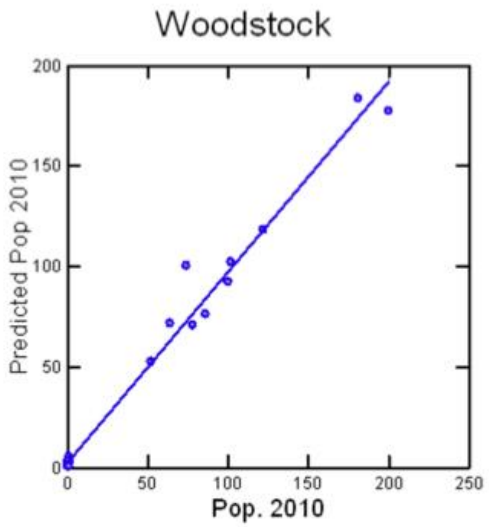
Woodbridge

Wolcott

Windsor Locks

Pop. Estimates Model for Ages 65 and Over - 2010

(Values are plotted where 2010 population > 0)



Appendix C: Two Descriptions of the Annual Population Estimation Process:

Appendix C: Two Descriptions of the Annual Population Estimation process:

Key:

"EST" prefixes identify the model-based population estimates, which are used to estimate the "annual change in population."

"POP2010" refers to the baseline Census-derived Population values for 2010.

"POP2011" refers to the new population estimate for years 2011 (and later), derived from the baseline 2010 population + estimated annual changes in population.

(1) Formula-1: The intuitive description of the estimation process used in the body of this report

e.g. Calculating the Estimated Pop for 2011 ("Pop2011")

Concept:	New Pop estimate in 2011 **=**	Base Pop in 2010	**+**	***Change*** in Estimated Pop. since 2010
Formula-1:	Pop2011 =	Pop2010	+	(Est2011 - Est2010)

(2) Formula-2:   An alternate presentation of the formula-1 (above)

Original Formula:	Pop2011 **=**	Pop2010	**+**	(Est2011 **-** **Est2010**)		
Substituting for the value of "Est2010":	*Since,* ***Est2010*** **=**	*Pop2010*	**-**	*(2010Model-ResIdual)*		
Formula-1 becomes:	Pop2011 **=**	Pop2010	**+**	Est2011	**-** POP2010	**+** (2010Model-ResIdual)
When simplified:	Pop2011 **=**	~~Pop2010~~	**+**	Est2011	~~-POP2010~~	**+** (2010Model-ResIdual)
Formula-2: ==>	Pop2011 **=**			Est2011		**+** (2010Model-ResIdual)

Comments:

Formula-2 makes it clear the fact that the error term from the 2010 models, i.e. the 2010 Model-Residuals", is treated as constant over the prediction period 2011-2014.

This version of the formula also shows that one can also conceptualize the new population estimates for 2011+ as being based on two components: 1) the model estimate for the respective year, and 2) the error term from the baseline 2010 model.

Some readers may prefer one formula over the other.   In either case, we expect this flexibility in conceptualization will further constructive discussion about the estimation approach we adopted.

Appendix D: SAS Programs for Annual Prediction by Model Type

The following files have been archived by the Connecticut Department of Public Health (DPH) for future Population Estimates (2015-2019) using new, annually updated input data.

Model- 0-4 years:

Birth_Bcode3.sas

Birth_score3.sas

Model- 5-9 years:

DOE_Bcode.sas

DOE_score.sas

Model- 10-14 years:

DOE_Bcode.sas

DOE_score.sas

Model- 15-19 years:

SDEDMV_Bcode.sas

SDEDMV_score.sas

Model- 20-64 years:

DMV_Meter_Bcode1.sas

DMV_Meter_score1.sas

Model- 65- 85+years:

Medicare_Bcode3.sas

Medicare_score3.sas