PERFORMANCE WORK STATEMENT

1.1 General

Virginia Polytechnic Institute and State University (Virginia Tech) will conduct research on the feasibility and efficacy of matching public records with the 2013 American Housing Survey (AHS) sample housing units for purposes of disclosure avoidance research.

Virginia Tech will furnish the necessary personnel, materials, equipment, services and facilities (except as otherwise specified) to perform the Performance Work Statement specifications identified in this task order.

1.2 Background

The U.S. Census Bureau is the primary source of statistics about the population and economy of the Nation. These statistics are collected to assist the Congress, the Executive Branch of the Federal Government, state and local governments, colleges and universities, and the general public in the development and evaluation of social and economic programs.

The AHS is a biennial survey of housing units conducted by the Census Bureau on behalf of the Department of Housing and Urban Development (HUD). The AHS sample is typically between 120,000 and 190,000 housing units. The AHS collects the following information about each occupied housing unit:

- Ownership (owner's name)
- Lot characteristics (size, zoning, sewer connection)
- Structure characteristics (type, size, stories)
- Physical characteristics (rooms, bathrooms, size, year built, heating and cooling equipment and kitchen appliances)
- Demographic profile of the housing unit occupants
- Financial characteristics (mortgage, owner-occupied only)
- Property tax and other assessment information

Protecting respondent confidentiality in Federal surveys is a legal requirement as well as a necessity for ensuring respondent participation. For the AHS, protecting respondent confidentiality means protecting both the identity (physical address) of the housing unit and the identity of the household members in that unit. The process of protecting respondent confidentiality is called disclosure avoidance.
A standard product for many Federal household surveys is public use microdata files (PUFs). PUFs typically contain answers to survey questions at the individual, household, and housing unit level. Prior to publication, PUFs and other survey products undergo a disclosure review to determine if, when, and how disclosure avoidance must be implemented.

The AHS adopts three disclosure avoidance techniques: pseudocoding, alteration, and suppression[1]. *Pseudocoding*[2] includes replacing the true value of an indicator with an alternative value with less specificity. Pseudocoding can be applied to continuous and discrete variables. Rounding, topcoding and bottom coding are examples of pseudocoding applied to continuous variables. Aggregation is an example of the pseudocoding technique that can be applied to continuous and discrete variables. The *alteration* technique includes replacing the true value of an indicator with a "nearby" value that can be shown while still protecting respondent confidentiality. One example in the AHS is to replace the true value of an AHS Zone with the value of a nearby AHS Zone. The *suppression* technique includes masking the true value of the indicator. A common masking technique is to set an indicator value to "Not disclosed."

For the AHS, respondent confidentiality must be protected at four different "levels." These include the person-level, household-level, housing unit-level, and physical location. The disclosure avoidance techniques mentioned above are applied to each of the four levels of indicators in the AHS. Table 1 lists some examples of indicators in the AHS, at each level, that require disclosure avoidance techniques.

Table 1. AHS Indicator Levels and Examples

Indicator Level	Indicator Example
Person	Income (very high), age, year migrated to US
Household	Household income, housing cost (rent/mortgage)
Housing Unit	Age of structure, size of structure, size of lot, value of home, purchase price of home
Location	Address, city name, metropolitan area name

Disclosure avoidance is fairly straightforward when indicators have known distributions. For instance, it is well-known that income follows a power law distribution[3]. As such, topcoding the highest 0.5 percent or 1 percent of income values should be sufficient to protect respondent confidentiality. Other housing unit-level variables, such as age, lot size, unit size, building size, or purchase price, have distributions that can be derived from alternative data sources such as tax assessment data.

With respect to location-level indicators, a well-known suppression method for household surveys, including the AHS and the American Community Survey (ACS), is the "100,000 persons" rule. In short, the Census Bureau suppresses any individual geographic indicator or combinations of geographic indicators that would identify an area with fewer than 100,000 people[4]. Analysis conducted by Hawala[5] supported the validity of this rule.

---

[1] It is important to note that these are not the only disclosure avoidance techniques used in surveys. For instance, the swapping technique is used in the American Community Survey.

[2] The terms "recoding" and "blurring" are synonymous with pseudocoding.

[3] Saez, Emmanuel, "Using Elasticities to Derive Optimal Tax Rates," Review of Economic Studies, 2001, 68, 205–229.

[4] The specific ways in which the 100,000 person rule are applied to the AHS are detailed in the document "American Housing Survey Public Use File Geography: 1985- 2013."

For some AHS indicators, disclosure techniques are applied to the joint distribution of two indicators. For instance, the number of rooms in a housing unit is topcoded and the topcoding value is different for individual metropolitan areas.

Disclosure avoidance becomes more difficult when housing survey questions replicate information available through publicly available data sources, thereby increasing the possibility that a surveyed housing unit or household can be uniquely identified by linking its survey responses to another data set with similar information. As of today, a handful of private sector firms have compiled tax assessment information for huge swaths of the country. One of these firms, CoreLogic, claims to have compiled mortgage information for more than 80% of the homes in the US[6]. Most of the tax assessment and mortgage information is geospatial.

1.3 Objective

Virginia Tech will provide two reports to be utilized by the Census Bureau in the support of future decisions regarding disclosure avoidance strategies for the AHS. The first research report (draft and final versions) should summarize the findings from Task 1 and the second research report should summarize the findings from Task 2.

1.4 Scope of Work

The scope of work includes two tasks. Task 1 includes producing a draft and final research report that summarizes the potential impact on disclosure caused by linking tax assessment, mortgage, and automated valuation model public records to the 2013 or 2015 AHS housing units. Specifically, the impact on disclosure should be assessed for two counties using the pseudo-universe approach described in 1.4.1, or an alternative approach that provides similar information.

Task 2 includes an evaluation of the potential linkages between the AHS and publically available or privately available data set other than tax assessment, mortgage, automated valuation models, and foreclosure data sets. Examples may include, but are not limited to, utility data, credit data, and other transactional data. At a minimum, the research report should describe each data set, how it can be procured, and how it can be linked to the AHS housing unit or household. More specific requirements for Task 2 will be provided upon substantial completion of draft report specified in Task 1.

1.4.1 Pseudo Universe Approach to Disclosure Assessment

As mentioned in 1.2, disclosure avoidance becomes more difficult when housing survey questions replicate information available through publicly available data sources, thereby increasing the possibility that a surveyed housing unit or household can be uniquely identified by linking its survey responses to another data set with similar information.

---

[5] http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00211.pdf
[6] http://www.corelogic.com/about-us/data.aspx

To better understand this issue, consider a housing survey with five questions: number of bedrooms, presence of a garage, number of stories, sale price, and sale date. Now further suppose that this information is publically available for every housing unit in the county through a tax assessment data portal. If the housing survey happened to include the largest home in the county (eight bedrooms) and the tax assessment data showed only one housing unit with eight bedrooms, then the address of the housing unit participating in the housing survey would be determined by linking to the tax assessment data using only one common characteristic (number of bedrooms). That would be a disclosure violation.

The ability to link a surveyed housing unit to tax assessment data is contingent upon the surveyed housing unit being unique within the area being surveyed (e.g., a county). Building on the example above, if there were 50 different housing units with 8 bedrooms, then it would be impossible to determine the address of the surveyed unit. However, suppose that the survey respondent reported their 8-bedroom housing unit sold for $1.1 million in 2014. While the tax assessment data shows 50 housing units with 8 bedrooms, it may only show one housing unit with 8 bedrooms that also sold for $1.1 million in 2014. The point is that each additional housing unit characteristic that is common between a housing survey and tax assessment data (or any similar data source) increases the likelihood that the housing unit is unique, and hence, can be linked to the tax assessment data through a set of common characteristics.

One way to assess the feasibility of finding a unique match between a surveyed housing unit and tax assessment data is to create a "pseudo universe" composed of real housing units populated by simulated households, then evaluate the pseudo universe for unique housing units/households based on the set of common housing units/households characteristics. To make this technique relevant to the AHS, the housing units and households in the pseudo universe should include information similar to what is found in the AHS.

The process of creating a pseudo universe starts with creating real housing units which can be derived from tax assessment data, mortgage data, and automated valuation data, at the housing unit (or parcel) level. Table 1 below includes the AHS variables that are also found in tax assessment, mortgage, and automated valuation data.

Table 1. AHS variables common to tax assessment, mortgage, and automated valuation data

AHS Variable Name	Description
AGERES	Age restricted development
BEDRMS	Number of bedrooms
BUILT (NEWC)	Year built
CELLAR	Presence of basement
CONDO	Unit is a condominium or cooperative
FLOORS (CLIMB)	Number of floors in structure
LOT	Lot size
NUNIT2, TYPE	Type of housing structure
NUNITS	Number of units in a building

BATHS and HALFB	Number of bathrooms in unit
PORCH	Unit has porch/deck/balcony/patio
PUBSEW	Unit is connected to public sewer
ROOMS (as well as combination of DINING, DENS, FARMRM, LIVING, OTHFN, and RECRM)	Number of rooms in unit.
ELEV	Working passenger elevator
UNITSF	Square footage of unit
WHNGET	Year unit bought/obtained/received
LPRICE	Purchase price of unit/land
VALUE	Current market value of unit
TAXPMT	Property tax payment
MORTAGE	Presence of a mortgage

The next step in the process is to attach geospatial indicators to each housing unit. Each housing unit created in the first step above should be augmented with the geospatial indicators in Table 3.

Table 3. AHS geospatial indicator

Variable Name	Description	Suggested Alternative Source
FLOODPLN	Unit in a flood plain	FEMA
EWATER	Bodies of water within 1/2 block of unit	National Hydrography Dataset
WFPROP	Unit is waterfront property	National Hydrography Dataset
EGREEN	Open spaces within 1/2 block of unit	Tax assessment information
ETRANS	Railroad/airport/4-lane highway within 1/2 block	Road data sets available from various vendors, including Census TIGER files.
ECOM1	Business/institutions within 1/2 block	Tax assessment information
ECOM2	Factories/other industry within 1/2 block	Tax assessment information
EHEIGHT	Height of apt bldgs within 1/2 blk	Tax assessment information. For some cities, building height may be available as a 3D GIS layer from Google or other entities.
NHDBLDSUD	Nearby single family detached home	Tax assessment information
NHDBLDSUA	Nearby single family townhouses	Tax assessment information
NHDBLDMU	Nearby apartment buildings	Tax assessment information
NHDBLDMH	Nearby mobile homes	Tax assessment information

The final step in creating the pseudo universe is to populate the real housing units with *simulated* persons and households. One way to do this is to make use of synthetic population data sets. One such synthetic population data set, the *2010 U.S. Synthetic Population Version 7*[7]*1* data set, includes geospatial coordinates that can be geospatially linked to the real housing units[8]. This data set was built using information from 5-year ACS data and the 2010 Decennial Census. The advantage to using this data is that it preserves the statistical relationship between location, household size, and income. Indicators derived from the synthetic population data set should include household size, household income, and the following information about the head of the household: age, race, and gender.

Once the pseudo universe data set is created, the data set can be evaluated for uniqueness by binning all the housing units and their respective households into groups based on the unique values of housing units/households indicators. The binning process would produce a count of the number of housing units/households that share a unique set of characteristics.

Once the bin counts are created, the bins containing a unique housing unit/household can be identified – these are bins where disclosure must be considered a problem. A higher threshold of three or five housing units/households per bin could also be considered a potential problem.

To add realism to this exercise, each indicator should receive the disclosure avoidance treatment currently applied in the AHS. For instance, indicators such as square footage, lot size, and value would be top-coded prior to forming the unique bins.

The recommended steps are:

1. Pick two counties for which CoreLogic or similar data (e.g., BlackKnight Financial Services) is available. Consider including at least one county in metropolitan areas included in the AHS National data collection.

2. Create real housing units by extracting the relevant tax assessment, mortgage, and automated valuation model indicators (and any other characteristic data, such as rooms) for each housing unit in the county.

3. Add the geospatial indicators to the real housing units, deriving them from readily available geospatial data (roads and waterways, for instance).

4. Create direct comparison of the administrative records-based housing variables in 1-3 above to the ANDS units in the same metro areas.

5. Populate the real housing units with synthetic households.

---

[7] http://www.epimodels.org/drupal/?q=node/32

[8] There are a number of ways to do accomplish this link, including a simply "nearest neighbor" assignment or more sophisticated linkages based on household size or income.

6. Apply the existing AHS disclosure avoidance techniques to relevant housing unit, geospatial, and household indicators.

7. Form unique bins based on the values of all indicators.

8. Determine if there are any bins with only one housing unit (or whatever threshold is desired).

If this process yields bin counts that are lower than the desired threshold level, further analysis is necessary to determine which of the indicators contribute most to the uniqueness of the bins. A simple way to determine which indicators are contributing most to the uniqueness of the bins is to re-run the above analysis several times, but drop one indicator each time.