Leveraging External Data Sources to Enhance Official Statistics and Products

Sallie Keller (Project Lead),
Director Social and Decision Analytics Division
Distinguished Professor in Biocomplexity,
Professor of Public Health Sciences, School of Medicine
https://orcid.org/0000-0001-7303-7267
sak9tr@virginia.edu

Stephanie Shipp https://orcid.org/0000-0002-2142-2136
Mark Orr, https://orcid.org/0000-0001-7950-8752
David Higdon
Gizem Korkmaz 0000-0002-4947-6320
Aaron Schroeder https://orcid.org/0000-0003-4372-2241
Emily Molfino https://orcid.org/0000-0002-6575-030X
Bianica Pires, https://orcid.org/0000-0002-4710-4849
Kathryn Schaefer Ziemer,
Daniel H. Weinberg https://orcid.org/0000-0002-2799-5015

Social and Decision Analytics Division
Biocomplexity Institute & Initiative
University of Virginia

January 5, 2016

Funding: This evaluation was conducted for the U.S. Census Bureau under contract to the MITRE Corporation

Citation for technical report: Keller, S., Shipp, S., Orr, M., Higdon, D., Korkmaz, G., Schroeder A., Molfino, E., Pires, B., Ziemer, K., Weinberg, D. "Leveraging External Data Sources to Enhance Official Statistics and Products." Proceedings of the Biocomplexity Institute, **Technical Report**. **TR# 2021-060**. University of Virginia. https://doi.org/10.18130/fyyx-0t81.

Citation for related wiki: S. Keller, S. Shipp, M. Orr, D. Higdon, G. Korkmaz, A. Schroeder, E. Molfino, B. Pires, K. Ziemer, and D. Weinberg. (2016). Dynamic Report Appendix – Collaborative Wiki, prepared for U.S. Census Bureau for the report Leveraging External Data Sources to Enhance Official Statistics and Products. Proceedings of the Biocomplexity Institute, **Wiki for Technical Report**. **TR# 2021-065.** University of Virginia. https://uva-bi-sdad.github.io/census2016_wiki/ (https://doi.org/10.18130/kn89-xm38)

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE

Abstract

The Census Bureau wants to understand how to leverage external data sources with traditional survey data and understand the effects on statistical data quality and standards of use resulting from incorporating external data. The Census Bureau tasked the Social and Decision Analytics Division, part of the Biocomplexity Institute of the University of Virginia, to address the question: how can we know if external data are useful for federal statistical needs? Today, the broad expansion of data collection across state and local governments, nonprofits, and commercial entities has created many opportunities to leverage external data sources to complement and even improve federal statistics. In the context of this report, external data are those collected by organizations outside of the federal statistical system, including city, county, and state governments, and the commercial sector.

To conduct this study, the team developed an initial data framework that encompasses the theory and methods capable of capturing, repurposing and integrating sources of data. This framework cannot be developed in isolation; rather it must be deeply grounded in real problems. Two specific case studies were chosen to ground the data framework development. The first case study considered measuring housing information directly based on locally available data. The second case study explored the use of state longitudinal education and workforce data. For both case studies, analyses were conducted to determine statistical properties, quality, accuracy, availability, and timeliness for each data source. Alternative estimates of housing and education information were created and recommendations were developed for external data sources, methodologies to produce estimates for the 2009-2013 American Community Survey (ACS) and selected uses of the ACS data.

Acknowledgements

The research team expresses our gratitude to the following individuals who con- tributed their data and their time to help make this project a success.

- **Kate Akers** , Kentucky Center for Education and Workforce Statistics
- **Melissa Beard**, State of Washington Education Research and Data Center
- **Kara Bonneau**, North Carolina Education Research Data Center
- **Chester Carr**, Arlington Department of Community Planning, Housing and Development
- **David Crowe**, National Association for Home Builders
- **Andrew D’huyvetter**, Arlington Department of Community Planning, Housing and Development
- **Joel Franklin**, Arlington Department of Community Planning, Housing and Development
- **Anthony Garrett** Arlington Department of Community Planning, Housing and Development
- **Jill Griffin** Arlington Economic Development
- **Clay Gottschall**, Arlington Department of Community Planning, Housing and Development
- **Elizabeth Hardy**, Arlington Department of Community Planning, Housing and Development
- **Mel Jones,** Virginia Center for Housing Research at Virginia Tech
- **Andrew McCoy**, Virginia Center for Housing Research at Virginia Tech
- **Tod Massa**, State Council of Higher Education for Virginia
- **Richard Millman**, Arlington Department of Community Planning, Housing and Development
- **Andrea Morris**, Arlington Economic Development
- **Troy Palma,** Regional Economist, Arlington Economic Development
- **Vaughn Poller**, James City County Office of Housing and Community Development

We would like to thank **Ron Prevost**, U.S. Census Bureau, **Rick Knowles**, MITRE Corporation, and the Census team for their leadership and guidance of this project.

We would like to give a special thanks for the research support and creative thinking of our 2015 summer student fellows **Stephanie Allen**, State University of New York, Geneseo; **Kyle Angelotti**, University of Virginia; **Berk Norman**, Rice University; and **Adrienne Rogers**, Virginia Tech.

Contents

Executive Summary

Census Challenges and Opportunities

The U.S. Census Bureau collects and provides statistics about the population and economy primarily through surveys. Survey data, especially large federal statistical collections, have long been the foundation of social science research and a vital part of economic planning. Congress, the Executive Branch of the Federal Government, state and local governments, colleges, universities, businesses, and the public use these statistics to evaluate social and economic programs, map economic development, and perform business and strategic planning.

Conducting these surveys and providing statistical products have become increasingly challenging, primarily due to declining response rates and rapidly increasing costs. The challenges are further amplified by the increasing demand for timelier and more geographically detailed data. The Census Bureau seeks opportunities to overcome these challenges through a variety of means. This report addresses one such opportunity, the use of external data sources.

The Census Bureau wants to understand how to leverage external data sources with traditional survey data and to understand the effects on statistical data quality and standards of use resulting from incorporating external data. The Census Bureau tasked the Social and Decision Analytics Laboratory, part of the Biocomplexity Institute of Virginia Tech, to address the question: ***how can we know if external data are useful for federal statistical needs?***

Today, the broad expansion of data collection across state and local governments, nonprofits, and commercial entities has created many opportunities for leveraging external sources of data to complement and even improve federal statistics. External data, in the context of this report, are those data collected by organizations outside of the federal statistical system, including city, county, and state governments, and the commercial sector.

Focus of the Report

The principal goal of this study was the development of an initial data framework that encompasses the theory and methods capable of capturing, repurposing and integrating sources of data. This framework cannot be developed in isolation; rather it must be deeply grounded in real problems. Two specific case studies were chosen to ground the data framework development. The first case study considered measuring housing information directly based on locally available data. The second case study explored the use of state longitudinal education and workforce data. For both case studies, analyses were conducted to determine statistical properties, quality, accuracy, availability, and timeliness for each data source. Alternative estimates of housing and education information were created and recommendations were developed for external data sources, methodologies to produce estimates for the 2009-2013 American Community Survey (ACS), and selected uses of the ACS data.

The deliberate decision to explore multiple case studies cannot be overemphasized. The development of a usable and forward thinking data framework requires leveraging synergies across domains. That effort cannot be performed in a solely theoretical context. A hands-on effort was required to find the sweet spot that yields the optimal benefits from data source evaluation and integration across multiple domains. Combined, the case studies are intended to lay the foundation for a data framework for the collection and use of data sources external those traditionally used to produce official statistics and products.

Research Approach

The research was designed to (1) identify ways to report official statistics using external data without compromising (and perhaps improving) quality, accuracy and standards; and (2) identify opportunities to provide more relevant and timely statistics with greater geographic detail. The initial data framework encapsulated a general approach for repurposing data, from discovery to analysis to inference. That framework was the culmination of a review and deep understanding of the data quality landscape, both across many fields and with respect to external data sources. Among the key challenges in using external data highlighted in the study is the lack of control that the user, the Census Bureau in this context, will have over the collection and processing of the data. This situation presents new challenges for evaluating data quality, as much of the data are idiosyncratic and rife with uncertainties in the context of the repurposing applications.

The data framework development that has emerged from this research contains the following components:

- **Data discovery** - the identification of potential data sources that could be related to the specific topic of interest, e.g. housing and education for this project
- **Data inventory** - the method used to screen and inventory the data sources to determine their value in supporting the research question and if they would be worthwhile to acquire
- **Data acquisition** – the process of negotiating and acquiring the data, and managing legal, privacy, security, and confidentiality practices
- **Data profiling** - a determination of both the quality of the data, provenance, and its utility to the project at hand

- **Data preparation** – the process of cleaning and readying the data for analysis; what is referred to as "wrangling the data"
- **Data linkage** – the process of building links to ensure compatible meaning, schemas, and ontology for data from multiple sources, resulting from the repurposing of the data
- **Data exploration** - the analysis of the datasets by summarizing main characteristics, often with visual methods
- **Modeling and analysis** – for this study, the process includes benchmarking against ACS and exploring new uses of the repurposed data
- **Fitness-for-use assessment** – the characterization of the information content in the results as a function of the analysis model, the data quality needs of the model, and the data coverage (representativeness) needs of the model

The purpose of developing the data framework in the context of specific problems is to identify commonalities through the use of the data framework across application domains. The ultimate goal is to develop a disciplined process of identifying data sources, preparing them for use, and then assessing the value of these sources for the intended use(s).

Implementation of the Data Framework through Case Studies

The research process for each case study was first to discover, inventory, and acquire external data sources to compare to selected ACS housing and education variables. For housing, the data discovery focused on local property data and neighborhood characteristics for two counties in Virginia, Arlington and James City. A total of 61 data sources were discovered. After an initial screening assessment, 23 underwent a full data inventory and 11 were fully acquired. For education, the data discovery focused primarily on the Statewide Longitudinal Data Systems (SLDS). A full inventory across all 50 SLDS systems was conducted, along with the exploration of other commercial data. A total of 73 education data sources were discovered, 33 underwent a full data inventory, and 5 were fully acquired. The 5 data sources included SLDS data from Kentucky, North Carolina, Texas, Virginia, and Washington.

Each of the acquired sources of data were profiled, cleaned, restructured, and linked to other variables to create estimates and then compared to published ACS estimates obtained from the American Factfinder or Public Use Microdata Sample (PUMS) data. The differences were quantified by comparing the estimates based on external data sources to the ACS estimates using the following ratio:

$$Fitness\ Ratio = \frac{ACS\ estimate - External\ estimate}{90\%\ ACS\ margin\ of\ error}$$

Fitness ratios are presented in tables, boxplots, and geographic maps of the area under study to identify areas of comparability to ACS and to explore reasons for differences. The true quantity, ground truth or gold standard is unknown for the quantities being estimated. In the end, the question to be addressed is whether or not the repurposed external data are useful for the intended purposes.

The housing study showed that acquiring and wrangling the county-level property records opens the opportunity to study features of housing at unprecedented levels of geography and time-frequency. This utility was demonstrated through the estimation and characterization of housing unit features and through the development of diversity measures based on home value (Gini and Simpson indicies) and application of hedonic regression techniques. The ACS comparisons highlighted opportunities to use local data in lieu of ACS collection for such variables as "year structure built," "value," and "real estate taxes paid."

The education study leveraged the richness of SLDS longitudinal data. These data provide a comprehensive and longitudinal view of student enrollments, characteristics, curriculum choices, and outcomes. The data were used in the development of logit models at the student and district levels to identify the characteristics of students that do not speak English fluently and high school dropouts that cannot be replicated with ACS data alone. The ACS comparisons highlighted opportunities for SLDS data to improve ACS estimates of "school enrollments" and "limited English proficiency" and to add indicators for "high school dropouts."

Leveraging Data Acquisition and Current Research to Identify Research Directions

Given the insights gained in the two case studies, further research is both needed and likely to be successful. Specifically, the following four research directions are promising.

- The housing and education external data presented in this study have the potential to be used to **replace or impute ACS data.** Promising areas will require evaluating data in other localities. These data include housing tax assessments, sales prices, and year built from property records and student enrollment counts and characteristics from SLDS data. These data need to be examined in more detail and across more localities.
- The housing and education external data provide opportunities to **add new data to the American Community Survey without requiring new survey questions**. Local property data could be a source of new housing diversity and inequality measures as well as provide deeper insights into pre-1940 housing. SLDS data could provide new data about numbers of and characteristics of high school dropouts, curriculum choices, and

other education activities and outcomes. A challenge that must be addressed is creating algorithms to align ACS PUMA to county and school district areas.

- External sources of housing and education data could be used to **update ACS and other federal statistics on a more frequent and comprehensive basis**. Examples include using housing permitting data to capture the rate of change of housing characteristics, housing sales data as a statistical sample, changes in school boundaries on an annual basis, and changes in school enrollments throughout the year.
- External sources of housing and education data provide a more complete picture of a locality and thus the Census Bureau could used these data to **create new publication products**. Examples include creating longitudinal profiles of local areas and states using housing and education data to analyze changes in housing stock and school to work transitions.

Conclusions

A central focus of this study was to determine the feasibility of leveraging external data sources to enhance official federal statistics. A data framework was developed and tested through the lens of case studies on housing and education. The research found that the external sources of housing and education data offer comprehensive and more detailed information and have potential for replacing or supplementing ACS data. The external data examined in this study were primarily administrative records that provide a more complete picture of all housing properties in a locality or all students enrolled in a school district. The longitudinal nature of the data makes them valuable for identifying inconsistencies and facilitates cleaning the data as well as providing for the analysis of changes over time.

Unlike federal statistics, the researcher or user of external sources of data has little to no control over the collection and production of data. The data sources examined in this study are collected for administrative purposes and are designed to meet the needs of the locality or school district where they are collected. To repurpose these data for statistical uses requires a disciplined, yet flexible and adaptable, data framework to assess data quality and fitness-for-use that is dependent on the new intended use.

The work in this study demonstrates that external sources can provide valuable information to support the ACS data. Thus this study is a valuable first step in assessing the value of external sources of data for use in federal statistics.

1. Introduction

The U.S. Census Bureau plays a critical role in the social and economic development of the nation. As the largest statistical agency within the federal statistical system, the Census Bureau is a primary source of statistics about the population and economy of the Nation. Congress, the Executive Branch of the Federal Government, state and local governments, colleges and universities, commercial organizations, and the general public use these statistics to evaluate social and economic programs, for economic planning, and in the development of business and strategic plans.

The foundation of social science research is built upon survey data, especially the large federal statistical collections, and relies heavily on the survey methodology developed in the 1930s primarily by statisticians at the Census Bureau. Conducting these surveys has become increasingly expensive and time consuming, in large part because of declining response rates, despite sophisticated attempts to collect data through adaptive design by multiple means including internet, mail, telephone, and personal visit (see Citro 2014, Groves 2011). Many federal statistical agencies provide national and regional level data but cannot do so at the geographic granularity needed by many users because of limited sample sizes.

Many federal statistical agencies provide national and regional-level data but, because of the limited sample sizes driven by costs, these agencies cannot provide the smaller geographic granularity many users need. The challenges of survey data collection have generated strong pressures for federal statistical agencies to explore the potential benefits and drawbacks of using data that are external to their organizations. These external data sources include federal, state and local government administrative records, non-federal surveys, commercial data sources, and electronic capture of human behaviors, to name some of the most prominent. The exploration of these sources has become an international trend (Holt 2007).

The critical issue in this exploration is to determine exactly the right approach to leveraging the potential of these external data sources. Looking internationally is useful because, for example, Europe has had extensive experience in using data that is external to the national statistical organizations. But, this is not completely applicable to the U.S., as the cultural and governance issues are different. Looking at the recent progress of the U.S. federal statistical system is important but insufficient because that progress has only scratched the surface of possibilities. Addressing the issue at hand will not come from looking in the past, but requires a ***new approach that looks into the future.***

The question at hand is: how can we know if external data are useful for federal statistical needs? This question is what this study was tasked to answer, and based on the work presented in this report, we can now argue that ***a solution for moving forward is continual development of a data framework for the evaluation of repurposed external data.*** This report reflects the beginnings of this effort and provides insight into how to move forward.

The remainder of this chapter provides the background for understanding the development of our approach, the research strategy we used to develop a data framework, the statement of work and the set of research questions that guided this study.

A. External Data and its Sources

The term ***external*** implies being outside of "something" and the "something" in this case is the federal statistical system. This project examines ***sources of data outside of the federal statistical system.*** We have chosen to call these "external sources" because it emphasizes a fundamental principle which we argue underlies the quality and usefulness of data. Data quality is typically discussed in the context of controlling the measurement, data collection processes, and being in control (ownership) of the data. This implies that ***quality is directly proportional to control***. When using external data for statistical or other purposes, the user has no control over measurement, data collection, or the ownership of the data. This fundamental issue is at the heart of characterizing data quality and fitness-for-use for re-purposing external data sources.

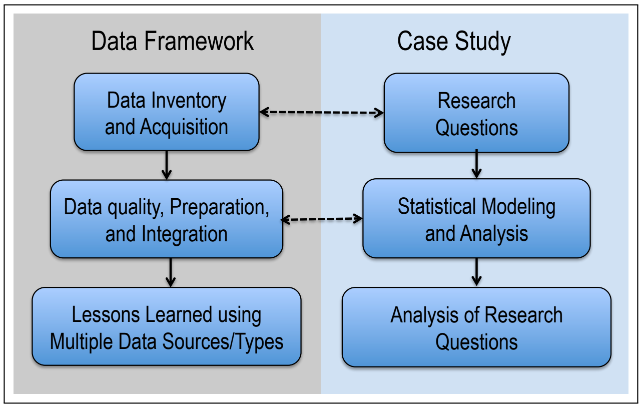
There are several sources of potentially useful data that are external to U.S. federal statistical system. These sources have been classified in similar but distinct ways by different researchers, e.g., designed versus organic (Groves 2011); survey versus big data (UNECE 2014); and, survey versus secondary (UNECE 2015), to name a few. Clearly many valuable sources of data are external to the U.S. federal statistical system, and hence outside of its direct control. The ultimate goal of this study is to develop a sound and generalizable approach for understanding and evaluating this datascape for the purposes of the U.S. federal statistical system.

B. Research Strategy

The principal goal of this study can be described as the development of a data framework that encompasses the theory and methods capable of capturing, repurposing and integrating multiple sources of data. Our research model, depicted in Figure 1.1, is composed of two parallel but interdependent tracks. The first focuses on data framework development including data acquisition, reuse, and integration. The second focuses on domain specific case studies.

Within the data framework track, we explore how to access, assess data quality and usefulness, and integrate multiple sources of data. For the case study track, the goal is to drive the data integration process with specific example problems, and to provide outputs for assessing and documenting the challenges with accessing and repurposing the sources of data. In the

Figure 1.1: Research Model


Data Framework
Case Study
Data Inventory and Acquisition
Research Questions
Data quality, Preparation, and Integration
Statistical Modeling and Analysis
Lessons Learned using Multiple Data Sources/Types
Analysis of Research Questions

Parallel, but interdependent tracks, for developing the theory and methods capable of capturing, repurposing and integrating multiple sources of data.

context of the case studies, this research project assesses and documents: (1) the complexity of accessing the data, (2) the challenges in repurposing the data to support the research questions, (3) the application of statistical methods not traditionally used with these data, and (4) the path forward for developing a data framework for on-going research and application.

We cannot overemphasize that the decision to explore multiple case studies was deliberate. Leveraging synergies across domains requires the development of a stable and forward looking data framework. In short, the data framework needs to continually evolve through application of the research model.

C. Statement of Work

The Census Bureau wants to understand how to leverage external data sources with traditional survey data and the implications on statistical data quality and standards of use with respect to the data sources. Seeking the complementarities and managing the benefits and challenges of the individual data sources requires effort to integrate multiple sources in a coherent way. That effort cannot be performed in a solely theoretical consideration of the data because some of the data parts are too idiosyncratic and too rife with uncertainties. A hands-on effort is required to find the sweet spot that yields the optimal benefits from data source evaluation and integration.

The Census Bureau tasked Virginia Tech to conduct a set of case studies to support the development of initial measures of quality and standards of use of multiple data sources. Each case study includes an inventory of potential data sources, selection of data sources, preparing the data for use, and undertaking statistical studies to integrate the ***external*** sources of data. The case studies are intended to lay the foundation for a data framework for the collection and use of data sources external to the data traditionally used to produce official statistics and products. The selected data sources were chosen to address the overarching issues the Census Bureau confronts of more timely statistics at a finer level of geographic and other detail.

1. Case Studies

Two case studies were selected for this research. The first case study explored options for using alternate data sources to measure housing information. Two geographic regions selected for the study include counties in Virginia, Arlington County and James City County. County, public and commercial data sources were evaluated for quality and usefulness.

The second case study explored the methodological challenges in using state longitudinal administrative education data at the school district and county levels. In particular, we examined the challenges in using state-level data sources and establishing best practices for their use. After a lengthy data inventory process, five states were selected for this case study. These were Kentucky, North Carolina, Texas, Virginia, and Washington.

For both case studies, analyses were conducted to determine statistical properties, quality, accuracy, availability, and timeliness for each data source. Alternative estimates of housing and education information were created and recommendations on the data sources and methodologies to produce estimates for the American Community Survey (ACS) for 2009-2013, and selected uses of the ACS data were developed.

2. Research Questions

The Census Bureau faces increasing challenges relying on surveys for production and reporting of official statistics. The goal for this research was to develop an initial data framework for leveraging ***external*** data sources to enhance official statistics and products. The strategy was to conduct two specific case studies and identify common methodological approaches. The research model is designed to (1) identify economical ways to report official statistics using external data without compromising (or perhaps improving) quality, accuracy and standards; and (2) identify opportunities to provide more relevant and timely statistics with greater geographic details.

The overarching research questions are:

Data Framework

- What features are needed for a data framework that characterizes content, access, timeliness, quality, and potential uses of non-federally collected data?
- For which American Community Survey (ACS) questions and for what subpopulations can non-survey sources of direct estimates be obtained at the unit level? Can

estimates be modeled at the unit level or at some aggregate geographic and/or temporal level?

Case Studies

- How can non-federally collected data sources enhance or complement a representative use of ACS data?
- What is the value of combining data sources, non-federally collected data sources and/or ACS data, to enhance or complement a representative use of ACS data?

To answer these questions, the study used external data from state and local governments and commercial vendors. The external data are benchmarked against the 2009-2013 ACS.

The remainder of this report documents the first steps in the development of a forward looking approach to data discovery, data wrangling, and data repurposing in support of U.S.federal statistics. Chapter 2 provides a short literature review on data quality. Chapter 3 presents the initial data framework that has emerged from this study. Chapter 4-6 present the housing case study and Chapters 7-9 present the education case study. Chapter 10 provides conclusions and recommendations for the next steps in this journey.

2. Data Quality

Data quality concerns are as ubiquitous as the variety of data in the world. Though the concept seems straightforward enough to warrant a simple yes/no, good/bad judgment, the burgeoning literatures on data quality indicate that this notion is deceptive. Data quality involves how subjects are measured, how data are collected (process), and fundamentally the given frame of reference, the the purpose for which the data are collected and analyzed. Tying measurement, collection, and purpose together reflects the notion of fitness-for-use.

Official statistics and data quality have always been closely related, with close examination revealing different but interrelated historical threads. One of these histories reveals that data quality is principally about pin-pointing the sources of measurement error (e.g., sampling and nonsampling error), the culmination of which are the contemporary approaches to total survey error (Biemer et al. 2014). Another history indicates that the early work in official statistics (e.g, Deming (1999)) emphasized very broad notions of data quality that came to maturity not in official statistics but in information systems (Brodie 1980), management and industrial practice (Redman 1992), fitness-for-use (Wang and Strong 1996), and frames of reference (Tayi and Ballou 1998).

These two historical threads intersected in the 1990s, a (re)union that continues to the present day to reflect fitness-for-use concepts from the management and industry practices literature, driven in part by Brackstone's seminal article (Brackstone 1999). He urged that official statistics should expand the notion of data quality from a mean-squared error approach to a more holistic approach, borrowing heavily from the total data quality management (TDQM) process (e.g., Wang et al. (1995), Wang and Strong (1996)). In principle, this holistic approach meant an expansion of both the dimensions on which to judge the quality of data and the processes and institutional structures to provide assurance of data quality. Official statistics is maturing into a comprehensive and modern approach to data quality assurance that incorporates both the total survey error approach and the data quality management approach (ESS 2015, Statistics Canada 2009, UK Office of National Statistics 2013, Australian Bureau of Statistics 2009).

One of the principal questions in this report is: ***To what degree do quality assurance frameworks already adopted by official statistics apply to external data sources?*** This question lays the foundation for the principal goal of this study: *the development of a data framework for evaluation of repurposed external data.*

In the paragraphs below, the current state-of-the-art in data quality in official statistics is described, including a historical perspective describing the development of the information systems, management and industry practices. The extent to which the quality of external data sources has been addressed in official statistics is discussed. Finally, lessons learned from reviewing the data quality literature are linked with the data framework development described in Chapter 3.

A. Data Quality in Official Statistics

In the 1990s, a sea change in data quality in official statistics took place, largely driven by the recognition that survey error might be amenable to a new approach, one that put more emphasis on the management of data quality. This change was aptly illustrated in a comparison made by Collins and Sykes (1999) between the titles of two conferences: the 1990 International Conference on Measurement Errors in Surveys and the 1995 International Conference on Survey Measurement and Process Quality. The later conference was considered an update to the earlier, however the later had progressed to emphasize processes and quality. Additionally, the 1995 conference proceedings (Lyberg et al. 1997) reflected a growing interest in the concept of fitness-for-use in official statistics (Dippo 1997).

The new focus on the management of quality in official statistics derived its impetus from the TDQM approach that arose from the information sciences and management practices in business and industry. The origins of TDQM is not surprising since business needs were a principal driver of database development in the 1970s and 1980s, and those databases had to overcome significant data quality issues (Wang et al. 1995). The TDQM approach inherited the database focus on highly technical data quality issues, e.g., integrity constraints and schema integration. A parallel situation arose with official statistics in the 1990s.

1. Total Data Quality Management

As TDQM matured in the early 1990s, it provided a general framework for understanding the improvement-through-management approach to data quality. The central idea was simple: to truly understand and change data quality, information systems should be considered analogous to manufacturing systems, with data as the raw material and data products as the output. This analogy afforded a natural springboard for the adoption of principles from Total Quality Management (Juran and Godfrey 1999) into the data space, as evidenced by the TDQM movement adopting the International Organization for Standardizations's ISO9000 (ISO 1992), which focuses on quality of any product writ large.

In effect, TDQM puts equal emphasis on all aspects of quality management from the technical, to administrative operations, to human resources and includes all stages of manufacturing from detecting issues in quality to final customer satisfaction and legal ramifications. The analogy between ISO9000 and data quality has proven useful, so much so that it serves as a basis for the TDQM framework (Wang et al. 1995) ontology that includes the following: management responsibilities, operational and assurance costs, research and development, production, distribution, personnel management and legal function. This comprehensive nature of TDQM is one of its benefits.

A core feature of the TDQM process is its description of the dimensions of data quality which provide a transparent basis for judging the quality of a data source. These dimensions capture multiple aspects of the systems generating the data products in a hierarchical fashion. Table 2.1 presents one of the early examples of this approach (Wang and Strong 1996).

Table 2.1: Hierarchical Data Quality Dimensions

Level I Dimensions	Level II Dimensions
Intrinsic	Believability, Accuracy, Objectivity, Reputation
Contextual	Value-Added, Relevancy, Timeliness, Completeness, Appropriate amount of data
Representational	Interpretability, Ease of Understanding, Representational consistency, Concise representation
Accessibility	Accessibility, Access security

Source: Wang and Strong (1996)

Although there are several dimensional schemes (see Batini et al. (2009) for a comprehensive review), Table 2.1 is presented because of its historical significance for official statistics and its influence on the state-of-the-art today (Hazen et al. 2014).

2. Integration of TDQM and Official Statistics

Brackstone (1999) demonstrated the important influence of the TDQM approach. Brackstone suggested six dimensions of data quality for official statistics: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. Table 2.2 summarizes these definitions. In addition, Brackstone suggested a comprehensive set of institutional mechanisms for managing data quality in federal statistical organizations, and descriptions of the manner in which data quality dimensions might be affected by such mechanisms (e.g., corporate planning, user liaisons, dissemination). In total, his work represents an approach to data quality management that is specific to official statistics and, clearly, influenced by the TDQM approach.

Today, Brackstone's work is strongly reflected in the contemporary official statistics data quality measures both in terms of the dimensions of data quality and a more end-to-end business management model (UNECE 2013, Biemer et al. 2014). Brackstone's influence can be

Table 2.2: Brackstone's Data Quality Dimensions

Dimension	Definition
Relevance	The degree to which data inform issues of importance to the users of data. Although subjective and variable across current and potential data users, NSOs should strive to put in place a program to meet even potentially conflicting user needs. The degree to which this is possible, and the approach for implementation is bounded by resource limitations.
Accuracy	The degree to which data represents the phenomena it was designed to measure. Standard statistical principles of bias, variance, and sources of error (coverage, nonresponse, etc.) are used to measure this dimension.
Timeliness	The temporal difference between time at which the data were collected and when they become available.
Accessibility	From a users' perspective, the ease of accessing the data from an NSO. This incorporates the degree to which it is possible to ascertain the existence of the data, the processes for access, and in some cases, the cost.
Interpretability	The degree to which the supplemental and metadata are useful for interpretation of the data and its uses. This includes conceptual issues, generation and classifications of variables, and data collection methods.
Coherence	The coherence of data represents the extent to which it can be used with other data and over time. This captures conceptual standards, classifications, and methods that go beyond numerical consistency.

Note: NSO is National Statistics Organization.
Source These definitions are paraphrased from Brackstone's original 1999 article.

seen in the core dimensions of data quality per the European Statistical System: relevance, accuracy, timeliness, punctuality, accessibility, clarity/interpretability, coherence/consistency, and comparability over time, space and domains (ESS 2015).

B. External Data, Data Quality and Official Statistics

In the past decade, the use of external data for official statistics has garnered much interest, mostly in reference to administrative data (see UNECE (2015) for definitions). In this context, the dimensions of data quality, as defined above, are deemed useful for gauging quality of administrative data. However, since administrative data are not in the control of a National Statistical Organization, other considerations have emerged (Verschaeren 2012). An influential approach originated from Statistics Netherlands (Statistics Netherlands 2012, Daas et al. 2008, 2009, Ossen et al. 2011) which introduced the following three hyperdimensions of data quality, specifically for administrative data: source, metadata, and data. The source dimension reflects quality related to the data generator such as procedures for access; metadata refers to the existence of and quality of documentation and knowledge provided by the source; data refers to the quality of the data in terms of population coverage, nonresponse, and precision among other more technical factors (see SN-MIAD (2013) for a review). An important research thrust has focused on providing evaluation frameworks for judging the quality of external data, specifically administrative data, at the input stage to serve as a screening mechanism against extensive investment in external data sources of poor quality (SN-MIAD 2013, Iwig et al. 2013, US Census Bureau 2015, Daas et al. 2008, 2009, Ossen et al. 2011).

In the current big data revolution, the situation is somewhat different. Current conditions require a wider set of data quality dimensions, including privacy, security, and complexity (UNECE 2014). In addition, the notion that big data should be viewed in terms of potential trade-offs (timeliness versus representativeness), as increasing efficiency (when combined with non-external data), and as potentially generating new data products (Braaksma and Zeelenberg 2015) casts data quality as relational among all data sources, traditional survey and administrative included.

C. Moving Forward with a Data Framework

The review above was instrumental in guiding the development of the data framework presented in the next chapter. It afforded a deep understanding of the data quality landscape, both in general and with respect to external data sources. More importantly, however, was the realization that, to date, there is scant research into the actual operationalization of the evaluation of repurposed external data for official statistics. The use of tools for screening will not go far enough. The issues are too complex, nuanced, and idiosyncratic for an approach that doesn't delve deeply into the data.

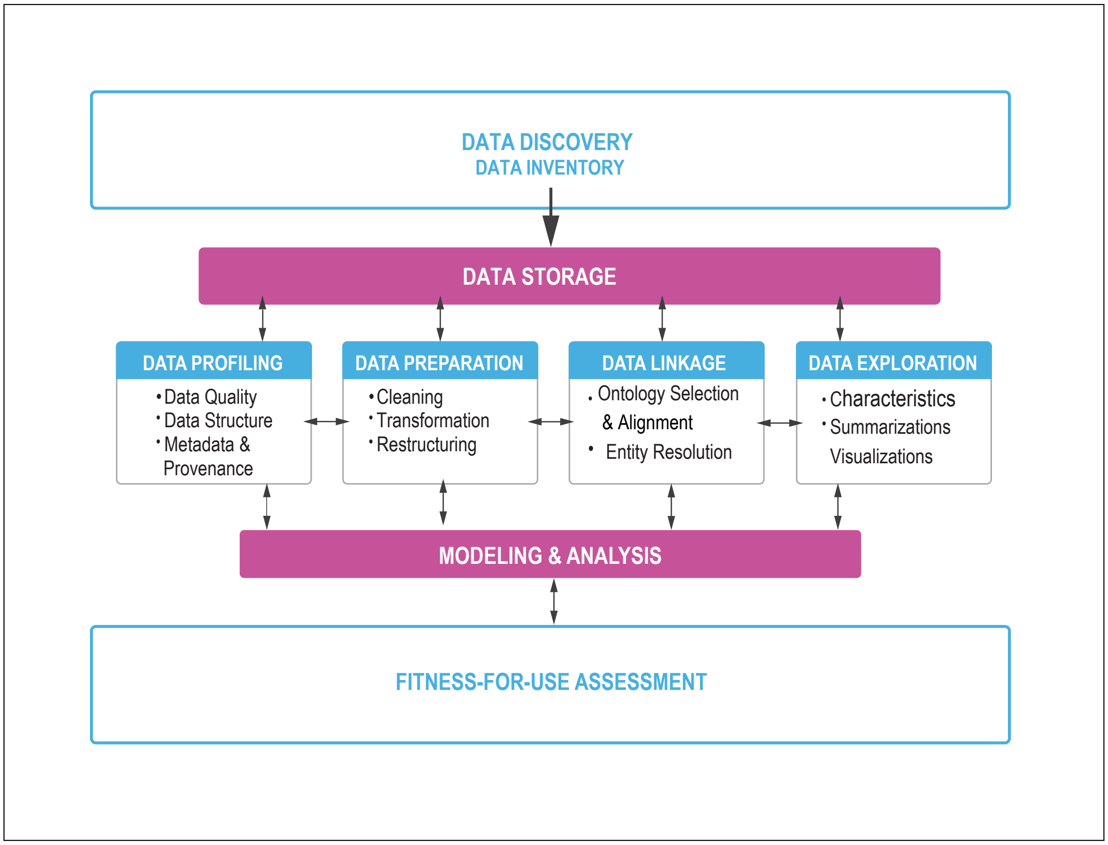
This chapter began with a question: ***To what degree do quality assurance frameworks already adopted by official statistics apply to external data sources?*** A provisional answer to this question is, in theory they seem to apply, but in fact little is known about implementation and operationalization of data quality frameworks for external data sources. The data framework described in the next chapter was driven by this insight.

3. Data Framework

The data framework encapsulates a general approach for re-purposing data, from discovery to analysis to inference. The general components were developed early in the project and informed by the literature review on data quality presented in Chapter 2. The framework structure depicted in Figure 3.1 emerged in the course of this project as a whole and, in the context of the housing and education case studies. The framework is described in the next sections as a set of successive linear steps. However, exercising the data framework is and will always be a highly iterative process.

Examples of the instantiation of the framework for the two case studies is given in Chapters 4-6 for housing and 7-9 for education. Those two sets of chapters do not neatly walk through the framework in the way it is described here. This is because the framework has grown out of intense preparation or wrangling the discovered data for the two case studies.

Figure 3.1: Data framework



This structure has emerged through the data discovery, acquisition, and use studies for the housing and education case studies.

The details of exercising the data framework for the housing and education case studies were captured during the course of the research using a collaborative wiki. The wiki pages that correspond to the data discovery, inventory, profiling, preparation, linkage, exploration, and ACS benchmarking are included with this report as a dynamic appendix. Box 3.1 describes the content of these wiki pages (Keller et al. 2016).

Box. 3.1. Dynamic Report Appendix – Collaborative Wiki

The VT Census Case Studies Wiki is an innovative approach to documenting and sharing the implementation of the Data Framework for the education and housing case studies. The Wiki allowed the VT and Census teams to share knowledge, collaborate, and distribute information. The Wiki was set up to be a working wiki in which outputs from the Data Framework process were shared as work was completed and updated as needed. As part of this report, a static copy of the Wiki was delivered to the Census Bureau. The wiki is referenced in this report as the collaborative wiki and cited as (Keller et al. 2016).
The Data Framework guided the structure of the Wiki. The Wiki categories follow:

- **VT Census Case Studies Home** page describes the purpose for the wiki; the project background, goal, strategy, key research questions, case study overview, and research model.
- **Data Inventory Guidelines** provides a data dictionary of terms used to measure data quality, the screening criteria, the information to include in the short inventory to describe data sources that pass the screening stage, and the categories to include in the full inventory for data sources that meet the short inventory criteria.
- **Data Inventory** lists the commercial, local, state, other, and federal sources of education and housing data identified and described for the housing and education case studies. Each source listed is clickable which takes the reader to the short or full inventory, if conducted.
- **Data Profiling, Preparation Process, and Benchmarking** describes the steps and results from implementing the data framework. This includes issues with benchmarking, areas that need to be addressed, documentation of assumptions, and a list of the ACS education or housing tables and which ones are compared to specific external data sources. Within this section, there is a page per data source per step, e.g. Kentucky Data Profiling; Arlington County ATRACK (rental data) Profiling.

A. Data Discovery and Data Inventory

Data discovery is the identification of potential data sources that could be related to the specific topic of interest, i.e., housing and education in this project. Data inventory is the process by which the sources of data were screened to determine if they might be useful to support the research questions and would be worthwhile to profile.

Our data discovery step started with a brainstorming of possible external data sources, i.e., data sources outside of the federal government, that exist related to housing and education. The goal was to think as broadly and imaginatively as possible to assemble a list of potential sources. Throughout the course of the project, as new ideas and sources of data were discovered, they were added to the lists for consideration.

The first step in the data inventory was to screen the data sources, identifying which sources were worthy of a deeper look and which were worthy of consideration for profiling. The screening includes five questions and a qualitative evaluation of purpose, data collection method, selectivity, accessibility, and description. The specific guidance is given in Box 3.2.

Following this initial screening inventory, a subset of the sources were selected for a full inventory. For the housing case study, 61 sources went through the initial screening and 23 were selected for a full inventory. For the education case study, 73 sources were screened and 33 were selected for a full inventory. The full inventory gathered the additional information presented in Box 3.3.

Box 3.2. Screening Inventory Process

1. Are the data collected opinion-based, (*e.g., people's attitudes, preferences, etc*.)?
2. Are the data collection recurring, (*i.e., must be collected at least annually*)?
3. Are there data available for 2013?
4. Geographic granularity

   For Education

   - Are the data collected at least the school level?
   - Can the data be linked to other education/workforce datasets, (e.g., K-12, higher education, workforce)?
   - If this is a state dataset, how do they define school districts within this state?
   - If applicable, what types of schools does it cover, (*e.g., public, private, charter*)?

   For Housing

   - Are the data collected at the property or housing unit level?

Additional Screening Information

Purpose:

- What is the purpose of the organization collecting the data, (*e.g., the Virginia Department of Education (VDOE) coordinates education for the state and makes policy recommendations*)?
- Why are the data collected and how does the organization use the data, (*e.g., VDOE collects the data for administrative purposes to assess student and school progress and to inform school policies)*?
- Who else uses these data, (*e.g., businesses, policy-makers, citizens, researchers)*?
- Who do they sell the data to, (*e.g., Zillow for individual homeowners, CoreLogic for multiple uses, business for economic development, Chief Economists at trade associations)*?

Method:

- What is the data collection method, (*e.g., paper questionnaire, operator entry, online survey, interview, sensors, algorithms for creating datasets from twitter feeds*)?
- What is the type of data collected, (*e.g., designed collection, intentional observation, administrative data, digital data*)?
- If designed, who created the questions, (*e.g., government, researchers, private business*)?
- What are the raw sources of the collected data prior to any aggregation, (*e.g., self-report, third party*)?

Description:

- What is the general topic of the data, (*e.g., student learning, housing quality*)?
- What are the earliest and latest dates for which data are available, (*e.g., 1995-2005*)?

Timeliness:

- Are the data collected and available periodically, (*e.g, every year or decade*)?
- How soon after a reference period ends can a data source be prepared and provided, (*e.g., one year*)?

Selectivity:

- What is the universe (*e.g., population)* that the data represents (*e.g., students who attended public school in Virginia in 1995*)?

Accessibility:

- How are the data accessed, (*e.g., API, downloaded - csv, txt, etc.*)?
  - ∗ Are they open data?
  - ∗ Any legal, regulatory, or administrative restrictions on accessing the data source?
  - ∗ Cost? Is it one-time or annual or project-based payment?
- Describe any gaps/concerns you see with this dataset

Does this dataset appear to meet for the needs for the Census Bureau study? Yes/No

Box 3.3. Full Inventory Process

Description/Features

- What is the temporal nature of the data: longitudinal, time-series, or one time point?
- Are the data geospatial? If Yes, at what level, (*e.g. census tracts, coordinates*)?

Metadata

- Is there information available to assess the transparency and soundness of the methods to gather the data for our purposes, (*i.e., supplementing the census)*?
- Is there a description of each variable in the source along with their valid values?
- Are there unique IDs for unique elements that can be used for linking data?
- Is there a data dictionary or codebook?

Selectivity

- What unit is represented at the record level of the data source, (*e.g., person, household, family, housing unit, property*)?
- Does this universe match the stated intentions for the data collection? If not, what has been included or excluded and why (*e.g., do the data exclude certain individuals due to the way the data are collected*)?
- What is the sampling technique used (if applicable, *e.g., convenience, snowball, random*)?
- What is the coverage, (*e.g. response rate*)?

Stability/Coherence

- Were there any changes to the universe of data being captured (including geographical areas covered) and if so what were they, (*e.g., changed the geographical boundaries of census tracts*)?
- Were there any changes in the data capture method and if so what were they, (*e.g., revised questions, data collection mode, classification categories, algorithms for social media data*)?
- Were there any changes in the sources of data and if so what were they, (*e.g., data were reported by teachers in 2010 and reported by principals in 2011; used Current Population Survey in 2011 and American Community Survey in 2012)*?

Accuracy

- Are there any known sources of error, (*e.g., missing records, missing values, duplications, erroneous inclusions*)?
- Describe any quality control checks performed by the data's owner, (*e.g., deleted duplicates, checked for recording errors*).

Accessibility

- Are any records or fields collected, but not included in data source, such as for confidentiality reasons, (*e.g., does not include any student files in which there are less the 5 students in a category*)?
- Is there a subset of variables and/or data that must be obtained through a separate process, (*e.g. state level data openly available, but one must apply to get census tract*)?
- If yes, is there a separate legal, regulatory, or administrative restrictions on accessing the data source?
- Cost? Is it a one time, annual, or project-based payment?

Privacy and security

- Was consent given by participant? If so, how was consent given, (*e.g., online form, in-person discussion*)?
- Are there legal limitations or restrictions on the use of the data, (*e.g., Family Educational Rights and Privacy Act -FERPA*)?
- What confidentiality policies are in place, (*e.g., cannot share data outside of requesting institution; does not include personally identifiable information)*?

Research

- What research has been done with this dataset, (*e.g., impact of policies, predictors of student success, housing stock inventory assessment*)?
- Include any links to research if provided.
- List any other data use notes provided by the supplier.

Following the full inventories, a subset of data sources was selected for acquisition, 11 for housing and 5 for education. The data inventory process for each of the data sources was documented on the collaborative wiki (Keller et al. 2016). Some examples are highlighted in Chapters 4 and 7.

B. Data Profiling

Data profiling starts with a determination of both the quality of the data and its utility to the project at hand. An important feature of the data profiling process is that discovered issues are only described and not actually "fixed". The appropriate fix will depend upon the specific needs of the research. If the prescribed "fix" is not appropriate, or even possible there would be no need for any action and attempting a fix at this stage could result in wasted time and effort. For example, it may be inappropriate to painfully re-categorize every missing zoning entry into the 38 zoning classification versus simply normalizing city zoning entries into residential or non-residential housing.

1. Data Quality Measures

Chapter 2 reinforced that a considerable amount of data quality research involves investigating and describing various categories of desirable attributes (or dimensions) of data. A typology emerged from the data work in our housing and education case studies. This typology consists of five data quality areas: completeness, value validity, consistency, uniqueness, and duplication.

a. Completeness

A set of data is ***complete*** with respect to a given purpose if the set contains all the relevant data for that purpose. Data that are missing can be categorized as record fields not containing data, records not containing necessary fields, or datasets not containing the requisite records. A common measure of completeness is the proportion of the data that has values to the proportion of data that "should" have values.

Completeness is a characterization of missing data and is application-specific. It would be incorrect to simply measure the number of missing field values in a record without first considering which of the fields are actually necessary for the task at hand. For example, in our study of Multiple Listing Service (MLS) Real Estate data, a dataset was provided with each record containing 128 fields. Per record, many of these fields were missing data, however, most of these fields were not important to the purpose of the study, (e.g. Listing Agent, Owner Name, Owner Phone). It would not be helpful to categorize the proportion of missing values for those fields. Instead, a decision must first be made as to which fields belong in the analysis for the current purpose.

b. Value Validity

Data elements with proper values have ***value validity.*** The percentage of data elements whose attributes possess values within the range expected for a legitimate entry is a measure of value validity. Checking for value validity generally comes in the form of straight-forward domain constraint rules. For example, the following *comparison-constraint rule* pseudo-code could be used to determine how many entries contain non-valid values for a non-empty text field representing gender.

$$< count\ gender\ where\ gender\ is\ not\ (male, female) >$$

To determine how many entries contain non-valid values for a non-empty integer field representing age, the following *interval-constraint rule* pseudo-code could be used.

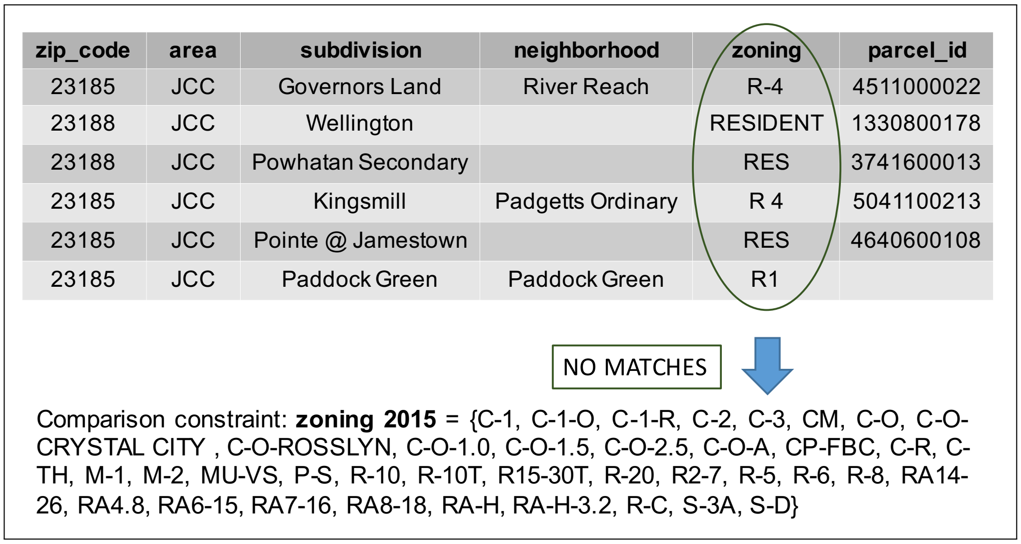
$$< count\ age\ where\ age\ is\ not\ between\ [0, 110] >$$

An example of a value validity problem uncovered in the housing data can be seen in Figure 3.2. While profiling MLS data for a particular locality, it was discovered that the values entered for the field “zoning” were extensively varied and did not align with the official list of zoning categories. The mechanism of input provided for this field was “free text.” In fact, it was found that none of the entries for this field qualified as valid values against the official zone list. However, there may still be usable information within this zoning field, depending on the intended use. For example, if the question at hand simply requires a count of how many properties were “residential”, it may possible to transform the existing entries to adequately represent a true or false with respect to residential property. However, no action was taken at this point to transform this data. The decision to execute transformations was left for the *Data Transformation* stage.

c. Consistency

The degree of logical agreement between record field values in either a single dataset or between two or more datasets is ***consistency***. The rules that specify the logical relationships between the entity values are called *dependency constraints*. A simple example of a dependency constraint violation would be a location disagreement like a zip-code that does not agree with a state code. Another might be the identification of a male who is also pregnant.

Figure 3.2: Value Validity Example



zip_code	area	subdivision	neighborhood	zoning	parcel_id
23185	JCC	Governors Land	River Reach	R-4	4511000022
23188	JCC	Wellington		RESIDENT	1330800178
23188	JCC	Powhatan Secondary		RES	3741600013
23185	JCC	Kingsmill	Padgetts Ordinary	R 4	5041100213
23185	JCC	Pointe @ Jamestown		RES	4640600108
23185	JCC	Paddock Green	Paddock Green	R1	

NO MATCHES

Comparison constraint: **zoning 2015** = {C-1, C-1-O, C-1-R, C-2, C-3, CM, C-O, C-O-CRYSTAL CITY , C-O-ROSSLYN, C-O-1.0, C-O-1.5, C-O-2.5, C-O-A, CP-FBC, C-R, C-TH, M-1, M-2, MU-VS, P-S, R-10, R-10T, R15-30T, R-20, R2-7, R-5, R-6, R-8, RA14-26, RA4.8, RA6-15, RA7-16, RA8-18, RA-H, RA-H-3.2, R-C, S-3A, S-D}

Zoning field in MLS data were allowed to be "free text." This resulted in none of the data being valid when compared to the official zoning categories.

Causes of inconsistency are varied. A common source of inconsistency comes from situations where locally derived information is provided with no associated master list or file. In our study, this occurred with student records such as from the Virginia Department of Education (VDOE). The student demographics occurred in multiple records about the same student recorded in the same year. Gender then had to be derived from the multiple observations. For example, in the VDOE data 16,310 of the 2,346,058 students had more than one value for gender across the multiple observations.

d. Uniqueness

The number of unique valid values that have been entered in a record field, or as a combination of record field values within a dataset, is ***uniqueness***. Uniqueness is not generally associated with data quality. However, for the purposes of answering research questions, the variety and richness of the data is of paramount importance. If a record field has very little value uniqueness, (e.g., entries in the field "State" for an analysis of housing within a single county), then its utility would be quite low and can be thought of as having low quality in terms of the research question at hand.

e. Duplication

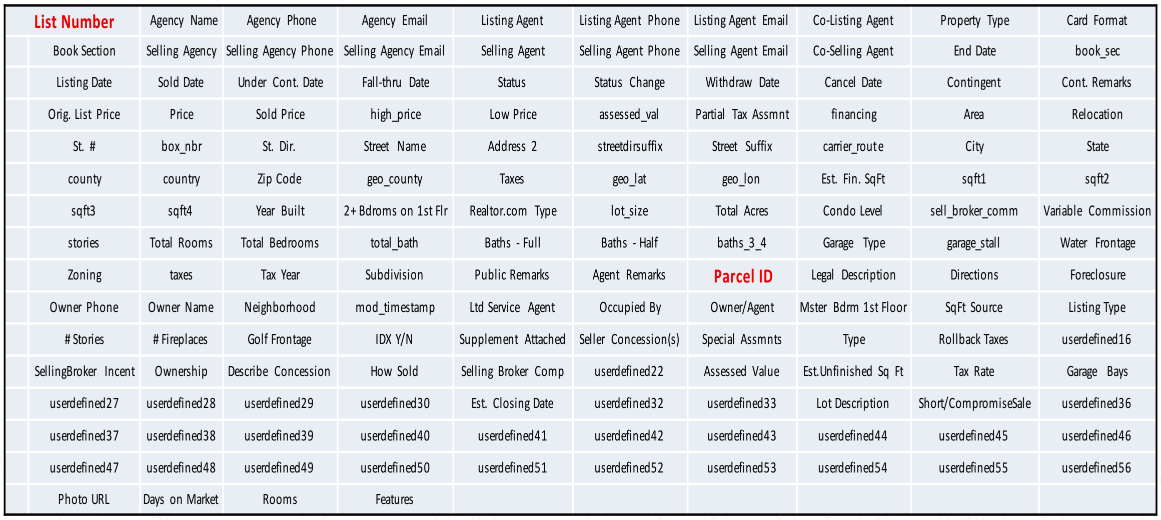
Duplication refers to the degree of replication of distinct observations per observation unit type. For example, in state-level secondary-education registration records, greater than 1 registration per student per official reporting period would represent duplication. While duplication can occur as a result of the accidental entering of the same information multiple times, duplication can occur many times as a direct result of the choice of level of aggregation, e.g., aggregating to a single student registration per academic year when registration information is actually collected multiple times per academic year.

2. Data Structure

Datasets associated with ***external*** data are often created for reasons of program and organization administration and reporting. The data collection and measurement processes associated with these data may not yield a structure that is conducive for purposes of statistical analysis. During the data profiling step, issues about data structure are identified. During the data transformation step decisions on how to restructure data are made and executed.

An example from the housing case study is given in Figure 3.3. The dataset provided was comprised of single records with 128 fields. Each original record was identified by a unique "List Number." However, if a parcel was listed twice it would have two different "List Numbers." As a result, changes in a property or parcel over time could not be tracked from these records because the structure only identified the list number not the parcel number. Changing the structure to include the "Parcel ID" allowed the required historical tracking of changes.

Figure 3.3: Combined Observational Unit Types



List Number	Agency Name	Agency Phone	Agency Email	Listing Agent	Listing Agent Phone	Listing Agent Email	Co-Listing Agent	Property Type	Card Format
Book Section	Selling Agency	Selling Agency Phone	Selling Agency Email	Selling Agent	Selling Agent Phone	Selling Agent Email	Co-Selling Agent	End Date	book_sec
Listing Date	Sold Date	Under Cont. Date	Fall-thru Date	Status	Status Change	Withdraw Date	Cancel Date	Contingent	Cont. Remarks
Orig. List Price	Price	Sold Price	high_price	Low Price	assessed_val	Partial Tax Assmnt	financing	Area	Relocation
St. #	box_nbr	St. Dir.	Street Name	Address 2	streetdirsuffix	Street Suffix	carrier_route	City	State
county	country	Zip Code	geo_county	Taxes	geo_lat	geo_lon	Est. Fin. SqFt	sqft1	sqft2
sqft3	sqft4	Year Built	2+ Bdroms on 1st Flr	Realtor.com Type	lot_size	Total Acres	Condo Level	sell_broker_comm	Variable Commission
stories	Total Rooms	Total Bedrooms	total_bath	Baths - Full	Baths - Half	baths_3_4	Garage Type	garage_stall	Water Frontage
Zoning	taxes	Tax Year	Subdivision	Public Remarks	Agent Remarks	Parcel ID	Legal Description	Directions	Foreclosure
Owner Phone	Owner Name	Neighborhood	mod_timestamp	Ltd Service Agent	Occupied By	Owner/Agent	Mster Bdrm 1st Floor	SqFt Source	Listing Type
# Stories	# Fireplaces	Golf Frontage	IDX Y/N	Supplement Attached	Seller Concession(s)	Special Assmnts	Type	Rollback Taxes	userdefined16
SellingBroker Incent	Ownership	Describe Concession	How Sold	Selling Broker Comp	userdefined22	Assessed Value	Est.Unfinished Sq Ft	Tax Rate	Garage Bays
userdefined27	userdefined28	userdefined29	userdefined30	Est. Closing Date	userdefined32	userdefined33	Lot Description	Short/CompromiseSale	userdefined36
userdefined37	userdefined38	userdefined39	userdefined40	userdefined41	userdefined42	userdefined43	userdefined44	userdefined45	userdefined46
userdefined47	userdefined48	userdefined49	userdefined50	userdefined51	userdefined52	userdefined53	userdefined54	userdefined55	userdefined56
Photo URL	Days on Market	Rooms	Features						

Example of MLS data table that combines "List Number" and "Parcel ID" in a single data table.

Our data framework incorporated several criteria for judging data structure to include the following:

- Missing variable names
- Combined variable issues– more than one variable represented in a column
- Multiple observation directions – columns and rows contain variable names
- Combined observational unit types – multiple types of data per record
- Divided observation unit type – splitting an observational unit type among multiple data sets

As these descriptions imply, the details were quite technical and do not fit within the scope of this report. An example encountered in our education case study is given here.

The example had to do with a dataset containing both individual demographic data and a periodic measurement of weekly attendance where demographic data and weekly attendance are separate observational units in separate datasets. Within administrative data systems, it was also not uncommon to find that a single observation unit type has been split among multiple datasets. This was similar to the consistency discussion of multiple observations with overlapping demographic information. The difference here was that the data recipient may have information split among several datasets.

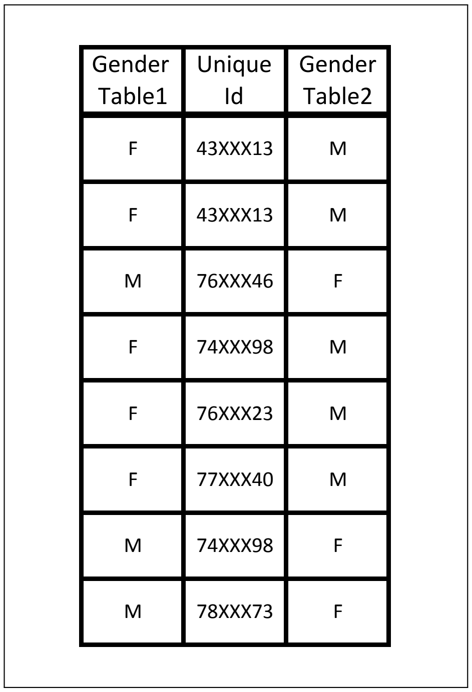
In our example, separate tables duplicate the collection of student demographics leading to mismatches as described with the VDOE data above. Figure 3.4 captures a subset of gender mismatches across two tables from the same education record information system, linked on the "Unique Id" of the student. Decisions on whether and how to transform inconsistent data as a result of a *divided observation* unit type need to factor in the magnitude of the issue as well as the ability to accurately correct the data in a timely enough fashion for the project at hand.

3. Metadata and Provenance

Metadata provides information about the data. In that sense, it is data about the data. The main purpose of metadata is to facilitate the discovery of relevant information pertaining to a particular data element or object. It does this by capturing and recording the following relevant information:

- Observation unit definition
- Observation unit attributes definition
- Semantic confusion
- Multiple attribute names
- Inconsistent attribute formats

Figure 3.4: Divided Observational Units in Multiple Tables



Gender Table1	Unique Id	Gender Table2
F	43XXX13	M
F	43XXX13	M
M	76XXX46	F
F	74XXX98	M
F	76XXX23	M
F	77XXX40	M
M	74XXX98	F
M	78XXX73	F

Example of divided observational units for gender mismatches in multiple tables.

Data process history is ***provenance.*** It refers to where the data came from and what the data are, including its inception, its history of access, transmission, or modification both in terms of what operations were performed and by whom. It provides a context for better understanding, interpretation, and inference.

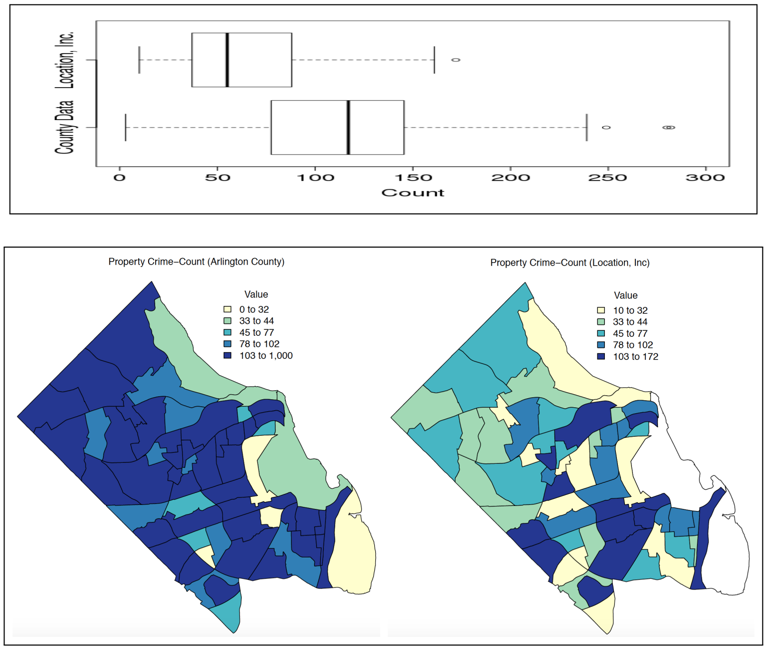
Metadata and provenance are of vital importance to discover if the data sources (datsets and tables), their observation units (records/rows), and their attributes (fields/columns) are consistently named, sufficiently described, and appropriately formatted for analysis and for combination with other project datasets. Two examples from our cases studies are highlighted.

The first example is about semantic confusion or interoperability. The concept of semantic interoperability refers to the ability of data systems to exchange data with other data systems unambiguously. Semantic interoperability is concerned not just with the data syntax, but also with the transmission of the meaning with the data, its semantics. This is generally accomplished by adding metadata to a dataset, thereby defining a controlled, shared vocabulary. Without this shared vocabulary, *semantic confusion* can occur, where names and syntax may agree, but definitions do not. In the education study, while combining two data sets, it was found that two fields had the same name ("Grade") but their definitions are completely different. In this example the attribute "Grade" is referring to both a test or class score and a school year level.

The housing case study provided an example of the importance of knowing the provenance of the data. Some of the datasets were provided by third-party vendors. As part of the value-added of these data products, third-party vendors often performed a set of transformations on the original data to enhance data consistency and quality. Sometimes the transformation processes used were readily available to the client, and the client could validate their application by repeating the transformations and reproducing the results. In other situations, the information may not be available, thus necessitating further investigation and experimentation on the part of the client to ensure that the data provided is, in fact, a true representation of the original source data.

One example from our data was from Location Inc., which provided indicators of neighborhood quality based on patented algorithms. We were unable to reconcile differences found in their crime indexes and data from Arlington County, Virginia Police Incident Tracking system. Figure 3.5 presents the misalignment in these data sources by census tract. The figure shows property crime counts as calculated by Location Inc. and as directly pulled from the Arlington County Police Incident Tracking system. The county data had five census tracts with counts greater than 300. These were not shown in the boxplot to allow a better scaled comparison to Location Inc. Location Inc. did not describe their methods to adjust the counts.

Figure 3.5: Comparison of Property Crime Data Counts



Geographic distributions of 2013 property crime counts by census tract using crime data provided for by the Arlington County and Location, Inc. (**Sources:** Location Inc. 2013, Arlington County Police Incident Tracking System 2013.)

C. Data Preparation

The data profiling operations described above were designed to bring data quality issues to the surface. The next step, as described in this section, focused on making decisions on what to fix and how.

1. Cleaning & Transformation

Data cleaning refers to the process of either fixing or removing data in a data source that is incorrect, incomplete, improperly formatted, or duplicated. Data transformation refers to the mapping of dataset field values from a provided format into an expected or more useful format. In practice, these two activities often occur together and include assessing the following:

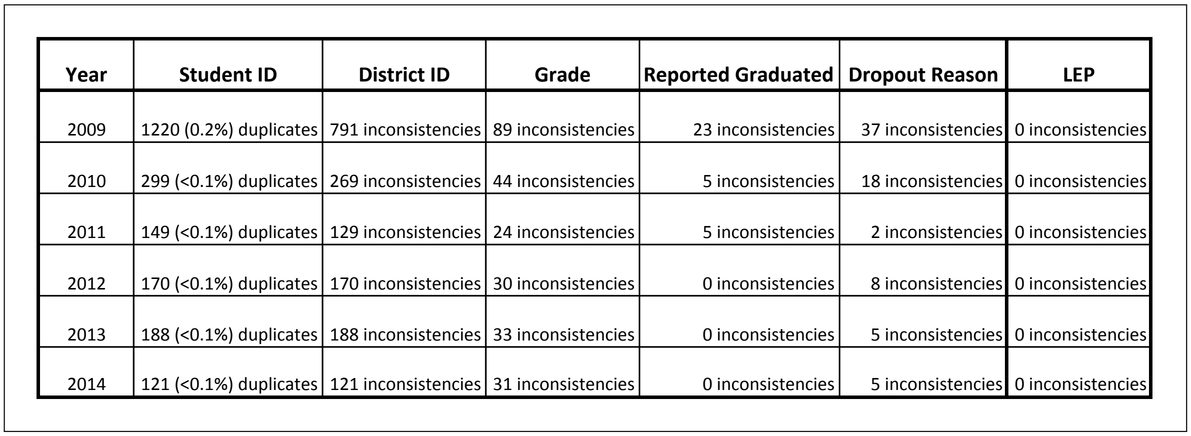
- Missing values
- Date and time formats
- De-duplication
- Outliers
- Normalization
- Feature extraction and construction
- Restructuring

De-duplication serves as an example relevant from our case studies. De-duplication refers to duplicate detection and deletion of all but one unique data record, according to the application of some algorithm for determining whether data contains duplicates. For example, the data in Figure 3.6 for the 2009-2014 Kentucky Longitudinal Data System had 2,147 total duplicate records based on the "Student ID" by "Year." Of these duplicates, 830 (39%) were complete duplicate cases, that is duplicates across all variables of interest, including "School District," "Grade," "reported graduated," "dropout reason," and "Limited English Proficiency." Due to the small number of duplicates and the lack of a consistent pattern, the first entry for a student in the dataset was retained and second entry removed. The only exception was if the second record had a dropout reason, then that record was retained.

2. Restructuring

To address issues of structure discovered during *Data Profiling*, it often necessary to ***restructure*** the data source (dataset) into multiple new datasets that are more easily analyzed. This activity can be thought of as being akin to the process of database normalization, the process of organizing the columns (attributes) and tables (relations) of a relational database to minimize data redundancy.

Figure 3.6: Duplicates and Inconsistencies



Year	Student ID	District ID	Grade	Reported Graduated	Dropout Reason	LEP
2009	1220 (0.2%) duplicates	791 inconsistencies	89 inconsistencies	23 inconsistencies	37 inconsistencies	0 inconsistencies
2010	299 (<0.1%) duplicates	269 inconsistencies	44 inconsistencies	5 inconsistencies	18 inconsistencies	0 inconsistencies
2011	149 (<0.1%) duplicates	129 inconsistencies	24 inconsistencies	5 inconsistencies	2 inconsistencies	0 inconsistencies
2012	170 (<0.1%) duplicates	170 inconsistencies	30 inconsistencies	0 inconsistencies	8 inconsistencies	0 inconsistencies
2013	188 (<0.1%) duplicates	188 inconsistencies	33 inconsistencies	0 inconsistencies	5 inconsistencies	0 inconsistencies
2014	121 (<0.1%) duplicates	121 inconsistencies	31 inconsistencies	0 inconsistencies	5 inconsistencies	0 inconsistencies

Duplicates and inconsistencies in the data. (**Source:** Kentucky Longitudinal Data System.)

An example of restructuring in the education case study was the subsetting of a database into three tables to account for student's race, gender, and disadvantaged status. Each of the three tables were aggregated by "School Year," "Division Number," and "Grade Code" according to a set of rules determining inclusion/exclusion of each variable in the Fall Membership table.

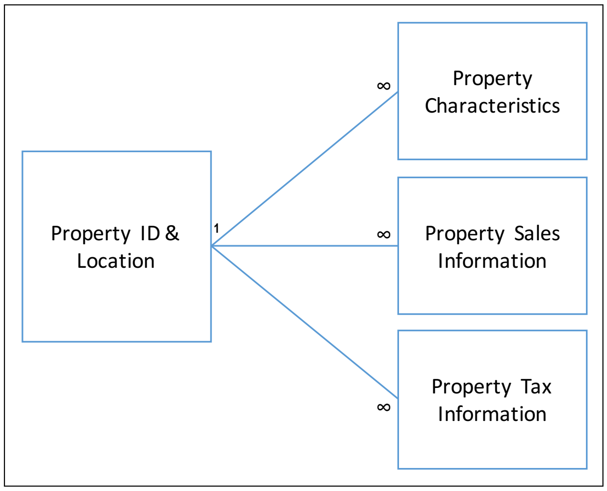
Another example of restructuring the data occurred when dealing with third-party MLS data. It was necessary to divide the dataset into multiple separate datasets,"Property ID & Location," "Property Characteristics," "Property Sales Information," and "Property Tax Information." Each of these new datasets represented a distinct unit of analysis. All of the new datasets were then associated via a new identifier, in this case, "Parcel ID" (see Figure 3.7). However, "Parcel ID" was left blank in over 7% of the entries. Therefore, extra work was required employing the use of a geocoding API (application program interface) to locate a property within county parcel maps that already included a "Parcel ID." Further, an additional complication is the fact that no standardized address format was used in the creation of the MLS record. Therefore, direct interaction and decision making by an analyst was also necessary to finalize the geographic matching.

D. Data Linkage

Two related but symantically different topics encountered in data linkage are ontology mapping and record linkage. We provide an example of each from our study.

Ontology mapping matches semantics between ontologies or schemas that are designed independently of each other. Ontology mapping is done by analyzing syntax, semantics, and structure, in order to deduce alternate semantics that may apply to other ontologies, and therefore create a mapping. An example from the housing study had to do with the multiple represen-

Figure 3.7: Restructuring Data



Restructuring of a single MLS dataset into multiple associated datasets useful for analysis.

tations of condominiums across the datasets. In one dataset, condominiums were represented by the number of units within a single building, as a single record. In another dataset, condominiums were represented as a collection of single owner-occupied units records. To map these two data sets, the geocodes associated with the single unit were used to aggregate all the condominiums associated with a physical location into one record.

Finding records in a dataset that refer to the same entity in other datasets is crucial, if not foundational, to the entire data science process. Putting these records together is record linkage. In the education case study, it was necessary to link the geographies of Texas counties with the geographies of both Texas school districts and Federal Local Education Agency assignments. To accomplish this task, multiple matching and validation steps were applied, including deterministic geocode matching, probabilistic name matching, and analyst review.

E. Data Exploration

Data Exploration refers to the analysis of the datasets by summarizing main characteristics, often with visual methods. Data exploration is used throughout the data framework. Descriptive statistics play a principal role in data profiling, from identifying valid attribute values to checking for semantic consistency. The use of visual techniques like boxplots support iterations between data cleaning and transformation during the data preparation. Distributional characterizations of the data help identify needs and opportunities for data linkage, as was highlighted housing condominium example above. The collaborative wiki (Keller et al. 2016) has numerous examples of data exploration that were used throughout this project.

F. Fitness-for-Use Assessment

The purpose of developing the data framework in the context of specific problems is find for synergies across application domains with respect to the data framework development and use. The ultimate goal is to develop a disciplined process of identifying data sources, preparing them for use, and then assessing the value of these sources for the intended use(s).

Understanding how to approach fitness for use starts with considering the modeling and analyses that will use the data. Modeling depends on the research questions and the intended use of the data to support the research hypotheses. Fitness assessment should be about the fitness of the data for the modeling, from straight forward tabulations to complex analyses. Therefore, fitness is a function of the modeling, data quality needs of the models, and data coverage (representativeness) needs of the models. Finally, fitness should characterize the information content in the results.

As described in Chapter 2, the driving goal of the data framework was to understand the fitness-for-use of these data. The initial fitness assessment in this research was based on benchmarking the tabulations created from the selected data sources against ACS tables. The differences were quantified by comparing estimates based on repurposed external data to ACS estimates using the following ratio:

$$Fitness\ Ratio = FR = \frac{ACS\ estimate - External\ estimate}{90\%\ ACS\ margin\ of\ error}$$

The challenge in the comparison originates from the inability to guarantee that either the *external estimates* or the ACS estimates will be right. There is no gold standard here. In some cases, such as housing tax assessments or year structure built, it is likely that county-level data were more accurate than the ACS. What is most interesting in the benchmarking is to look for patterns in how the differences in the estimates manifested themselves. Several examples are given in the Chapters 5 and 8 the present the comparisons of external housing and education data to ACS data.

A second fitness assessment is presented, with the details in Chapters 6 and 9. This assessment was based on considering how the external data sources enhanced representative research studies that traditionally have used ACS data.

4. Housing Data Framework

This chapter discusses the findings from the data discovery, inventory, acquisition, profiling, preparation, linking and exploration for the housing case study. The material presented in this chapter does not sequentially follow the steps outlined in Chapter 3 because, as described in Chapter 3, the data framework has emerged from the intense look for data in the context of our case studies and documenting what and how we learned about the data sources. The collaborative wiki (Keller et al. 2016) documents the process to acquire information, what was learned, what rules were developed, and what decisions were made.

A. Identifying Housing Data Sources

Data discovery started with the creation of a comprehensive list of potential sources of housing data. Of particular value was a recent review of housing sources in (Weinberg 2014, 2015). The potential data sources identified in the first step are listed in Table 4.1. As described in Chapter 3, each source was first screened to assess if the data met the screening criteria. To benchmark these data against 2009-2013 ACS tabulations, it was important that 2009-2013 data be available from these external sources.

Data sources identified through the initial screening process as potentially relevant for the study underwent a full inventory. The data discovery process identified 61 ***external*** data sources of which 23 passed the screening process and underwent a full inventory. These data sources are denoted with an asterisk in Table 4.1. Based on the full data inventories, the following 11 data sources were identified as worth acquiring:

- Arlington County data: real estate assessments, ATRACK (apartment complexes), Geographic Information System (GIS), crime
- Black Knight Financial Services (real estate assessments)
- CoreLogic (real estate assessments)
- James City County data: Parcel data, GIS
- MRIS: Real estate listing and sales data for Arlington County
- WMLS: Real estate listing and sales data for Williamsburg-James City County
- TransUnion (mortgage data)
- Home Mortgage Disclosure Act (HMDA) data
- U.S. Postal Service vacancy data
- U.S. Department of Agriculture forest inventory data
- Location Inc. neighborhood data

Table 4.1: Non-Federal Housing Data Sources Inventoried.

Commercial	**State**
Black Knight Financial Services (BKFS)**	Housing Virginia
MPF Research	Northern Virginia Association of Realtors
National Association of REALTORS	VHDA Housing Analysis
Real Capital Analytics	Virginia Housing Coalition
National Association of Home Builders and "Housing Economics" (Data/Forecasting Companion Site to NAHB)	
Mortgage Bankers Association	**Local**
Redfin	Arlington County (AC): CPHD data**
CoreLogic**	AC: Permitting**
Zillow**	AC: Real Estate Assessments**
MLS Data**	Arlington Economic Development**
Equifax Credit Scores	AC: Mapping Center**
RealtyTrac	AC: Crime data**
WegoWise	AC: Building Energy Report Cards
Williamsburg Local MLS**	AC: Bicycle & Pedestrian Counters
Metropolitan Regional Information System (MRIS)**	AC Affordable Housing Study
TransUnion Credit Data**	James City County (JCC): GIS/Mapping**
Experian	JCC: Real Estate Assessments**
Foot Traffic - SentriLock	JCC: Crime Data**
Axiometrics, Inc.	JCC Citizen Survey
Planet Labs	JCC Housing & Community Development
Blackbridge	
CoStar	
Other	
National Change Database (NCDB)	Urban Institute
Community Commons Maps	Panel Study of Income Dynamics
Crime Reports	Yelp
IPUMS-USA	Walk Score**
Google Maps**	RS Metrics
Location Inc. (Neighborhoodscout)**	AirBnB
USDA Forest**	TripAdvisor**
Maponics**	InfoUSA Mailing List
Center for Regional Analysis	ARLnow
Urban Tree Canopy Analysis of Virginia Localities**	National Council on Real Estate Investment and Fiduciaries
Factual	

Note: ** Sources that were selected for a full data inventory at the early stage of the project. All sources subsequently underwent a full data inventory.

During the course of the project, full data inventories were completed on all data sources in Table 4.1 as well as the collection of federal data sources in Table 4.2. The results of the data inventories can be found on the collaborative wiki (Keller et al. 2016).

Public sources of utility cost or use data were not identified. While for most jurisdictions in the country, electricity and utility gas are supplied by regulated commercial entities, (e.g., Dominion Resources and Washington Gas in the Arlington County study area), water and sometimes sewage removal is often provided by municipalities themselves. Inquiries to Dominion Resources and Washington Gas were ignored. Inquiries about obtaining water bills, even at an aggregate level, (e.g., census tract), were refused by Arlington and James City Counties, citing concerns about security based on advice from the U.S. Department of Homeland Security.

Table 4.2: Federal Data Sources Inventoried

American Housing Survey
Assisted Housing (HUD)
Community Reinvestment Act (CRA)
Comprehensive Housing Affordability Strategy
Construction Statistics
Decennial Censuses of Housing
Department of Commerce Economic Indicators
Economic Censuses
Fannie Mae, National Housing Survey (NHS)
Federal Housing Finance Agency
Federal Reserve Economic Data (FRED)
Freddie Mac
Home Mortgage Disclosure Act (HMDA)
Homelessness Data Exchange (HUD)
Housing Vacancy Survey (HVS)
Location Affordability Portal
Longitudinal Employer Household Dynamics
Low-Income Housing Tax Credit (LIHTC) Database
Manufactured Homes Survey (MHS)
National Household Travel Survey
New Residential Construction
Property Owners and Managers Survey (POMS)
Rental Housing Finance Survey (RHFS)
Residential Finance Survey (RFS)
Residential Energy Consumption Survey (RECS)
Survey of Consumer Finances (SCF)
Survey of Market Absorption of Apartments (SOMA)
Survey of Residential Alterations and Repairs (SORAR)
Uniform Crime Reporting Statistics
United Nations Statistics Division- Housing
U.S. Bureau of Labor Statistics, Consumer Price Index
U.S. Postal Service, Vacancies

B. Acquiring Housing Data Sources

1. Local Government Data

Introducing the project and meeting with data owners who work with the data on a daily basis paved the way for obtaining local government data. This project leveraged our established relationship with Arlington County, Virginia. Arlington County has given us access to county data as an agent of the county. This includes access to confidential data, along with the obligation to protect those data commensurate with any county employee. Using this partnership, we contacted the county's demographers, who were instrumental in arranging a meeting of the key data owners throughout the county. With their help, we identified key data sources, made the necessary contacts, and acquired the data. These contacts were crucial for not only obtaining these data but also for answering many questions throughout the data profiling and preparation stages.

At the start of this project, the Virginia Center for Housing Research at Virginia Tech was in the process of establishing a relationship with James City County, a mostly rural county near Williamsburg, Virginia, to conduct a housing stock inventory. Leveraging this opportunity, we were able to use their contacts to access parcel and GIS data from James City County.

2. Commercial Real Estate Data

Four commercial firms that supply some aggregation of real estate property data for a fee were identified. These were:

- Black Knight Financial Services (BKFS)
- CoreLogic

- RealtyTrac
- Zillow

In discussions with the four, it was determined that only BKFS and CoreLogic did independent data collection. Thus, BKFS and CoreLogic were chosen and acquired for comparison to the localities' data and ACS tabulations.

RealtyTrac was eliminated because CoreLogic licenses some of its data to them. This is a result of a Federal Trade Commission complaint that alleged that CoreLogic would unilaterally exercise market power after CoreLogic's acquisition of DataQuick (FTC 2014).

Zillow is an online real estate database company that provides information about homes for sale and rent. It is not a real estate company and thus does not get its data directly from the Multiple Listing Services (MLS). Zillow uses third-party data collection companies to provide the county property data. In addition, owners and sellers can update data on their own properties. Zillow has a patented algorithm, the Zestimate, to create an estimated value for a property.

Obtaining data from BKFS and CoreLogic required going back and forth multiple times between the lawyers at each company and Virginia Tech to agree on and finalize the contracts. Both companies required a Master Services Agreement that governed the general components of data use, and a specific project use or statement of work agreement. Both sets of agreements had restrictions on the commingling of their data with other data sources except for the purposes of the specified project agreement. Both agreements prohibit the direct comparison of their products with other commercial products.

BKFS data for 2009 through 2013 were requested, and a contract was relatively straightforward to negotiate. The cost was $8,250, the data covered the entire state of Virginia, and the agreement ended December 31, 2015. BKFS project agreement was tightly aligned with our project statement of work. This was stated as: *"... for Company's internal business purpose of evaluating and testing the Products to determine whether Company wants to license the Products for use in conjunction with the U.S. Census American Community Survey."* We are required to provide them with feedback in the form of an"Evaluation Report." The data were delivered via FTP in 613 files in August 2015. This was the timeliest delivery and easiest to negotiate of all of the commercial data sources.

CoreLogic data were initially thought to be easy to acquire, as the firm has established a university research portal. After substantial investment in time to negotiate the use of that acquisition vehicle, it was discovered that the portal had only the most current year's data versus the 2009-2013 time frame needed for the study. Thus, negotiations had to be re-opened to acquire the historical data, which is housed in a separate database at CoreLogic. The second set of negotiations was not able to acknowledge or leverage all the legal changes and agreements made during the first set. Therefore, it was necessary to start over again and it took several weeks to finalize the negotiations. Data for the entire state of Virginia were finally acquired for $7000 via FTP in 5 files on November 12, 2015. On the more positive side, the agreement is for 2 years and the use is broadly defined as"Academic Research Purpose"

In addition to the real estate data aggregators, mortgage data from TransUnion were pursued. TransUnion required setting up membership to its Data Exchange Gateway before moving forward. Negotiations with TransUnion involved a successful inspection of our facilities for protecting the confidentiality of their data. It took months for the two legal departments to negotiate the membership. After being approved for membership, the negotiations over the statement of work commenced. As with CoreLogic, none of the earlier agreed-upon conditions carried forward into this new contracting process. This second process dragged on into October and we finally terminated the negotiations. The cost of the data would have been $27,000 for the two Virginia counties.

It was hoped that sufficient mortgage information could be obtained through the Home Mortgage Disclosure Act (HMDA) website, from which data are easily downloaded. Unfortunately, after examination, HMDA did not have useful mortgage data in the context of this project. The ACS collects and tabulates data on the total monthly mortgage payment for owner-occupied units. HMDA has data on total mortgage amount and not monthly mortgage payments. It is important to note that the total mortgage amount data meets HMDA's primary use of examining housing inequities and identifying possible discriminatory patterns.

3. Transaction Data

Data on housing transactions located within multiple listing services (MLS) were identified as a source worth exploring. A potential value of the MLS data is that they contain the most up-to-date information for a subset of housing units. MLS data are continuously updated throughout the U.S. for real estate agents. While there are services that provide the data nationally, MLS data are owned and managed locally. There are over 900 MLS organizations nationally, each has their own process of releasing data. The two MLS files of interest were the Metropolitan Regional Information System (MRIS) that serves the metro DC area and Williamsburg MLS (WMLS) that serves James City County and surrounding area.

Obtaining data from the WMLS required finding the correct contact within the regional group. Once contact was made and the need for the data verbally described, the data were emailed directly to us. In contrast, MRIS is a private company that maintains the database. Similar to the WMLS, the correct contact had to be made. Initially, the inquiry came back stating that a data exchange was not possible. After further inquiry to the right contact, the data were purchased for $500 following the submission of a written statement of work.

4. Other Commercial Data

Additional commercial data sources were considered to support the research questions on the use of external data to enhance the ACS data for specific ACS data use cases. Data from Maponics and Location Inc. were investigated. Maponics provides data at "neighborhood boundaries," which are proprietary areas to delineate areas where people live, work, and socialize. These geographic areas do not align with census boundaries and would have been quite difficult to use.

Location Inc. does include census boundaries. Location Inc.'s original quote for $45,000 for a one-year license would have included lifestyle, neighborhood characteristics,and education data for the two counties. After much discussion, a limited-use six-month agreement that covered a selected number of neighborhood characteristics for $19,000 was negotiated. We acquired the data for Arlington and James City Counties on neighborhood characteristics in October 2015. Part of the time-consuming nature of this negotiation was removal of the clause from the contract giving them the right to advertise Virginia Tech as clients and users.

The use of Yelp and Google Maps were investigated as potential sources to identify additional neighborhood characteristics, such as distance to restaurants or retail establishments. These areas of interest are not covered by Yelp's academic datasets and the terms of service prohibits web scraping or using the application programming interface (API) for research. Web scraping Google's API was attempted but took too much time to be useful within this project's timeline.

C. Housing Data Profiling, Cleaning, and Transformation

As described in Chapter 3, a series of steps were followed to assess these housing data. The description of each of the steps is on the collaborative wiki (Keller et al. 2016). Data profil-ing allowed for examining the quality, consistency, and uniqueness of the data. As described below, this process was an exploratory process that required many interactions with the data, documentation, and data providers. Understanding the unit of analysis for the external data was important for comparisons with ACS tabulations, (e.g., parcels versus housing unit).

Using this knowledge, the data were then cleaned, primarily, by making the ***external*** data conform to the same unit of analysis as the ACS data, i.e., residential housing units. This process required discovering ways to eliminate non-residential properties and vacant land. There were some instances where the data had to be restructured, (e.g., tax payments made into one observation). In addition, errant data needed to be recoded (i.e., year built values in the future or a townhouse with 43 bedrooms). In some cases, values could be cleaned using Arlington

County's online property search (Arlington County 2015). Data were transformed to match the categories used in the ACS tabulations and to put into constant 2013 dollars when appropriate.

None of this was a linear process. Often, we had to go back to profiling and cleaning after examining descriptions of the prepared data. More details about each of these steps is presented below for each of the data sources. In addition, this process provided input to into the data framework itself described in Chapter 3.

1. American Community Survey

The ACS data are not described in this report, except below in some sections where the ACS data collection methodology provides insights into differences between the ACS and the administrative records data, such as timing of data collection.

2. Real Estate Assessments

a. Arlington County Data

Arlington County real estate assessment data were downloaded directly using the county's API. These data came without a codebook, and a codebook had to be created by our project team. The SDAL codebook, available on the collaborative wiki (Keller et al. 2016), contains brief descriptions of seven categories of variables:

- *Property*: The basic information, primary address, owners, etc., about a parcel
- *Improvement-Dwelling*: Information, such as year built and heating type, about each dwelling on a parcel
- *Improvement-Interior*: Information about the interior of a dwelling, such as number of bedrooms or bathrooms, on a parcel
- *Assessment*: Historical assessment information for a parcel
- *Real Estate Assessment Payment History*: Information about the assessment and tax payment history of a parcel
- *Sales History*: Contains one instance for each time the ownership status of a parcel has changed

The selected variables in each files required substantial data cleaning and preparation, which are described on the collaborative wiki (Keller et al. 2016).

Through this process, differences between the county data and the ACS data surfaced. The first difference was for real estate assessment data. The county real estate assessment data included all parcels in the county. ACS data were respondents' answers about residential units. Duplicates, non-residential properties such as parking lots, and vacant parcels in the Arlington

County real estate assessment data were deleted. Table 4.3 provides information on the steps or rules used and the results. This process was not as linear as the table makes it appear. It was an iterative process to create the data descriptions, to discover potential problems, (e.g., total assessment values too high or too low), and to make changes. Discussions with Arlington County's demographers and real estate data staff provided definitions of land-use codes and types of parcels, and reasons for missing data.

Particular problem areas included identifying parking lots, vacant parcels, common areas, and multiple dwelling parcels. Also, a small number of parcels had to be removed that were partially in Arlington County and partially in another jurisdiction. Parking lots were found by pinpointing parcels with no improvement (building) value but with a land value. Removing parcels with no improvement value did not eliminate condominiums that have no land attached because they have land value. Virginia law states that condominiums must have a land value so a portion of the improvement value gets shifted to the land value.

In regards to multiple dwellings on a parcel, the data had to be restructured so that there was one observation per dwelling. It should be noted that the Arlington County real estate assessment data was the *only* data source that included these multi-dwelling units. BKFS did have a field for duplicated parcel IDs, which were suppose to mark multiple dwelling units, but the field was empty. CoreLogic data is set up so that each parcel number is the unique location, thus multi-dwellings cannot be identified in their data either.

Additionally, there were two *year built* variables from two different tables within the Arlington County data: property-year-built and dwelling-year-built. In the property-year-built data, all properties were included and the unit of observation was the unique parcel number. In the dwelling-year-built data, only single-family properties were included and the unit of observation was the dwelling. Thus, a parcel with two dwellings could have two different dwelling-year-built values if the second dwelling was built later than the original dwelling. To create a *year built* variable that had a value for each dwelling and multifamily property, we took the dwelling-year-built data and added the property-year-built data for those parcels not listed in the dwelling-year-built data.

A second key difference between Arlington County real estate assessment data and ACS data was that ACS collects information for each housing unit in its sample, while assessment data pertained to each parcel in the jurisdiction. For example, multi-family buildings are typically one parcel. For the most part, however, the assessment records indicated the number of units on each parcel, allowing the property data to be re-weighted to create comparable statistics to ACS tabulations.

Subsequent to the data profiling and preparation process for the Arlington county real estate assessment data retrieved through their API, historical data CDs were acquired from the county.

Table 4.3: Restructuring Data to Residential Parcels With a Housing Unit

		County	BKFS	CL
	Original N (2013)	**65,433**	**65,443**	**61,591**
Justification	Step Taken			
Residential Only	Use recoded land-use code to keep residential parcels	62,847	62,847	–
	Remove hotels	–	–	61,519
No Buildings/Vacant Land	Remove parcels with $0 improvement value	–	61,174	60,959
	Remove those coded as vacant	62,049	–	–
	Remove common areas	62,049	–	–
Parking Lots	Remove parcels with $0 Land Value	60,988	60,592	60,343
Remaining Vacant Land	Select parcels with land value greater than $15,000	–	60,413	–
Non-Arlington Properties	Remove multi-jurisdiction properties	60,966	–	–
	Final N (2013)	**60,966**	**60,413**	**60,343**

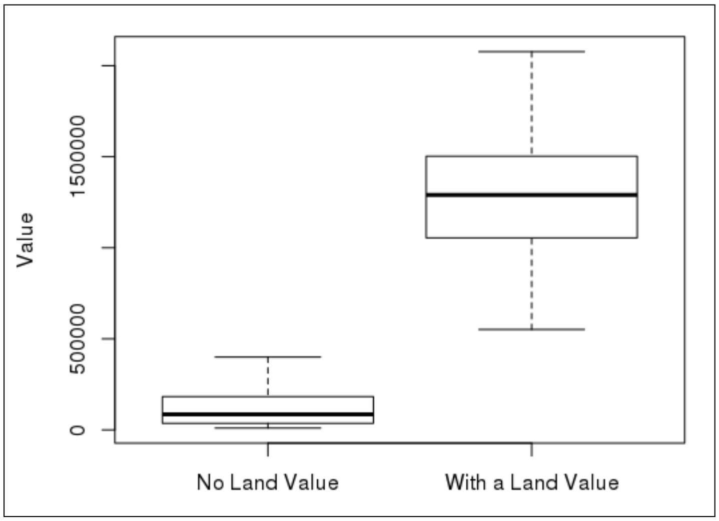
Note: This table shows the steps (rules) taken to restructure the data from all parcels to residential parcels with a housing unit. The "Remaining Vacant Land" step was done because there were no vacant land or common area codes. Some vacant land/common areas do have minimal improvement value.
Sources: Arlington County Real Estate Assessment Data (County) 2013, Black Knight Financial Services (BKFS) 2013, CoreLogic (CL) 2013.

The county makes these historical CDs available upon request. These data CDs did come with a detailed codebook. Differences were found between the data that were available through the API and what was on the CDs. The API version provided interior housing characteristics for each floor of the property (e.g., a two-story residence would have two observations). The data CD had only one observation per parcel.

Assessment data from the API was provided at the assessment event level. A parcel can have multiple assessments per year, each reflected in a new observation. Each event was marked with the reason that assessment was provided (e.g., new construction, annual). Through the profiling and preparation process, it was found that “new construction” assessments prior to 2012 had $0 for land value, which affected the total value for that housing unit (see Figure 4.1). According to county experts, the reason for this was that land value cannot change in the middle of the year, so county officials chose not to post the land value for these mid-year assessments. Thus, we had to replace the $0 land value with the value recorded at the end of the year and create a new total value. Though the data CD had only one assessment per year for each parcel and was more complete, it did not include information on why that assessment occurred.

In order to place the parcels into census geographic boundaries (tracts and block groups), geocodes (latitudes and longitudes) were needed. Knowing that other external data sources did not have geocodes or had inconsistent geocodes, we created a master-address list that contained all unique addresses in the external data and the necessary geographic information. This would ensure that a parcel would have the same coordinates and thus census boundary information across years within a data source and across data sources. Arlington County real estate assessment data formed the base of the master address list. This required making a list of all addresses in the 2009-2013 data and removing duplicate addresses. However, there were 191 properties

Figure 4.1: Total Value of Housing Units for "New Construction" Real Estate Assessments



The boxplot shows the distribution of total value of "new construction" assessments with no land value and with a land value. (**Source:** Arlington County Real Estate Assessment Data 2009-2013.)

which had at least one different address across 2009-2013. To ensure that coordinates and thus census boundaries remained constant across time, these addresses were cleaned using a set of rules. The most frequently used of these rules included:

- **Parcels that had *no street number + a street*, but then later had *street number + same street*.** For these, the street number was added to the street name for that parcel.
- **Parcels that had *no street number + a street*, but then later had *a street number + different street*.** For these, all addresses were changed to the later, more complete address.
- **Parcels that had a minor difference (e.g., one number off, St. instead of Rd.) in their address.** These were changed to match the majority.

The full set of rules used is on the collaborative wiki (Keller et al. 2016).

Geocoding was completed using the Google Maps API based on address. Google Maps Geocoding API is a service provided by Google to convert an address into geographic coordinates that can then be plotted on a map (Google 2015). The latitude and longitude plus additional information about the location is provided such as points of interest listed at the address, administrative areas, and neighborhood names. This service was accessed through the programming lanugage R (R Core Team 2013). Due to the volume of geocodes required, we created a developer's account to obtain an API key and allow for billing. The API is otherwise freely available for use up to 2,500 queries daily. Google Maps API did not produce perfect results with the first run of the initial master address list, which took 19 hours to run more than 61,000 addresses. Several outputs were null, non-Virginia, or approximate results (.7%).

There are three main types of geocoding outputs: rooftop, approximate, interpolated. The goal was to have all outputs be rooftop, as it is the most accurate. Approximate is when Google Maps do not have either the specific address or a close match to the address so will output the centroid of the smallest known administrative area, e.g. center of Arlington County. Interpolated is when Google Maps does not have the specific street number but has close street numbers so it outputs the center between two points, e.g., the ends of a street block. Using the latitudes and longitudes, the parcels are placed within the appropriate census tract and block group. For more information on the rules used to clean addresses and the geocoding process, see the collaborative wiki (Keller et al. 2016).

As a result of implementing the data framework steps, the final number of "active" properties (parcels), increased from 64,827 in 2009 to 65,443 in 2013. Table 4.4 provides more detailed counts of properties and units. Not all properties have information about the dwellings, such as year built and number of bedrooms. There were 206 properties where the unit counts change across the years. The most notable was the incorrect entry of number of units in 2013 for 9 multi-family buildings with differences of 20,000+ units between 2012 and 2013. The largest difference was for one building identified as having 842 units in 2012 but 266,412 units in 2013. Arlington County Real Estate changed systems in 2013, which seems to explain the problem. As this high unit count would bias estimates after weighting, the 2013 numbers were replaced with 2012 numbers.

This choice of a of data transformation may have missed some units completed in 2013 for buildings still under construction. Nonetheless, the difference of unit counts across the years was relatively small. The ACS tabulations group reported number of units into bins, which could absorb much of the difference. Moreover, not all properties have a unit count. Missing unit counts were assumed to be single-family housing units. Yet, it was discovered that this resulted in condominiums being classified as single-family residences (as stated earlier, each condominium had its own individual parcel). Thus, to obtain unit counts for condominiums we had to impute a value. Since each building had its own GIS code, the number of condominium units was imputed by aggregating the units based on the GIS code. Lastly, it was discovered through this process that duplexes also did not have a unit count. We used a specific duplex land use code to assign a unit count to these properties.

Some variables had a substantial number of missing values. These included "heating type" (43% missing). The content of that variable has limited usefulness in any case for identifying heating fuel due to definitional differences between ACS and the local data, (e.g., "forced hot air" vs. "electricity").

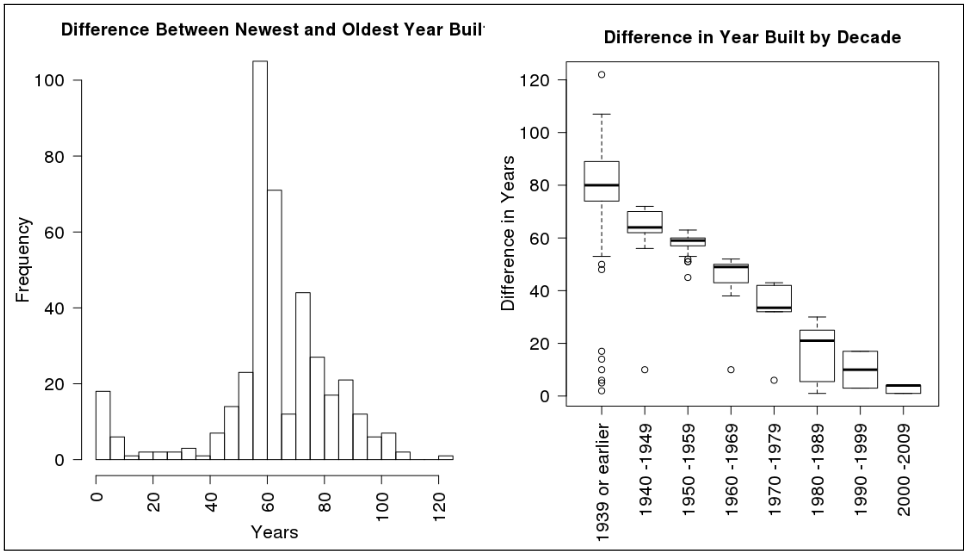
Table 4.4: Parcel and Unit Counts from External Data

	2009	2010	2011	2012	2013
Arlington County					
Property	60,261	60,203	60,465	60,688	60,966
Units (as weighted)	100,991	101,867	102,299	102,511	103,987
BKFS-Arlington County					
Property	–	–	–	–	60,413
Units (as weighted)	–	–	–	–	103,691
BKFS-James City County					
Parcel	24,923	25,387	25,833	26,154	26,606
CoreLogic-Arlington County					
Parcel	59,593	59,959	60,077	60,220	60,343
Units (as weighted)	97,589	98,690	83,640	79,521	79,804
CoreLogic-James City County					
Parcel	24,864	25,317	25,659	26,130	26,591
Units (as weighted)	–	26,013	26,361	26,832	27,294

Note: Weights used to derive the number of housing units are based on the reported number of units on each parcel. The – indicates missing data.
Sources: Arlington County Real Estate Assessment Data 2009-2013, Black Knight Financial Services (BKFS) 2009-2013, CoreLogic 2009-2013.

For "year built," 403 properties had at least one year built that is different across the time period and the range of the difference significantly decreased in later decades, as shown in Figure 4.2. Large differences may have indicated new construction where the original building was demolished.

Figure 4.2: Distributions of the Differences in Year Built for the Same Housing Units



Difference is *(newest year built - oldest year built)* for the same property as recorded over 2009-2013 in Arlington County. (**Source:** Arlington County Real Estate Assessment Data 2009-2013).

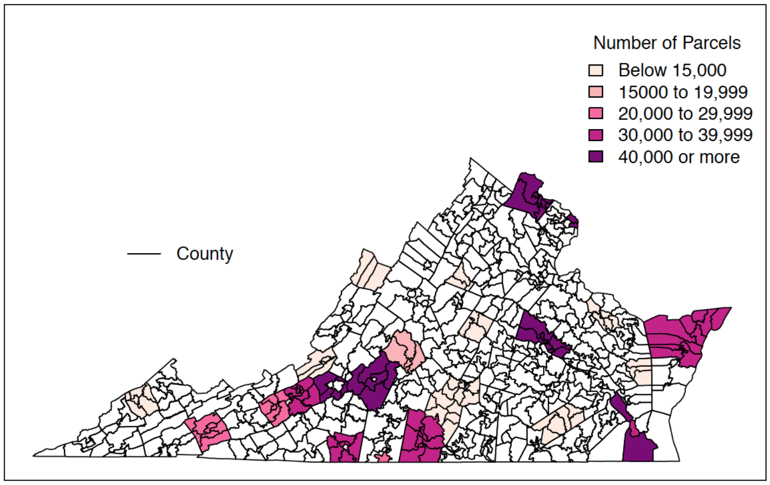
b. James City County Data

James City County data did not have a codebook or data dictionary, so one had to be created. Duplicates, non-residential, and unimproved properties were removed. Unfortunately, historical data were not available, so characteristics were collected only for currently active parcels (as of July 2015). James City County Real Estate Assessment data could only be benchmarked to ACS 2014 1-year estimates. Current and historical assessment data were included, however no dates were provided about when the assessments occurred. The records only had current, previous, and second previous assessments denoted, but not the dates. Historical selling price data and dates were available for the past three sales of a parcel.

3. Black Knight Financial Services Assessment Data (BKFS)

The BKFS data came with an extensive codebook, but the codebook and value descriptions had problems, which are described below. Coverage for Virginia was quite low, ranging from 13 to 28 counties/independent cities over the 2009-2013 period, out of 133 counties/independent cities. Figure 4.3 shows the 2013 coverage. Coverage rate maps tended to follow urban/rural areas with rural areas often missing.

Figure 4.3: Black Knight Financial Services 2013 Parcel Coverage for Virginia



Counties highlighted have some parcels included in the data source. (**Source:** Black Knight Financial Services (BKFS) 2013.)

a. Arlington County

There were no assessment data, (e.g., value), for 2009-2012 in the data that we received. These data were missing as BKFS claimed they did not receive assessment data for these years. BKFS did have physical housing characteristics data during that period.

There was no explicit code in their data to remove vacant properties, parking lots, or common areas after removing non-residential properties. Therefore, we could not use the same process as we did with Arlington County real assessment data to get the BKFS data to the same unit of observation as the ACS. Instead, we used assessment values to remove some of these properties. Yet, as assessment data was not available for 2009-2012, we were only able to transform BKFS 2013 data to the same unit of observation as the Arlington County real estate assessment data. Table 4.3 lists the steps taken to identify the residential properties and Table 4.4 lists the property number counts. BKFS does retain the original land-use code provided by the county, which meant that it was possible to execute the subsequent data transformation steps. Table 4.3 lists the steps taken to identify residential properties for CoreLogic.

In addition to the lack of longitudinal editing, we were limited by BKFS assumptions and restructuring of their original data. BKFS does not transfer over from the original data set entries that are true zeros, rather these are coded Not Available (NA). This means that it was not possible to distinguish between zero bedroom units, (e.g., studio condominiums), versus units with missing data, such as number of bedrooms. Also, just under 50% of the data did not have a year built. After inquiring about this, BKFS said that they do not use the year-built provided in a different condominium table (than the one they used) in the original data. They agreed to provide an updated file with the information from the condominium table. However, we did not receive the updated data at the time of writing this report.

There was also a tendency for units to change address and GPS coordinates without changing physical location in the data. Latitude and longitudes in BKFS are populated using Pitney Bowes (PB) Geostan (Pitney Bowes 2015), an address standardization software. PB updates the database used in the software on a monthly basis. Thus, the latitude and longitude values represents data provided from different periods of time. It is unclear why so much of the data changed over time. Some of these changes resulted in large geographic jumps, large enough to move parcels into different census tracts. Due to the uncertainty of these data, Google Maps API-generated coordinates were created. After examining the addresses from BKFS as compared to Arlington County's, there were approximately 8,000 new addresses to geocode. Many of these reflected minor difference in the street name, (e.g., N Fenwick st vs. Fenwick St). It was also necessary to remove the "#" from the units in the addresses as the Google API does not recognize them and would misplace the address.

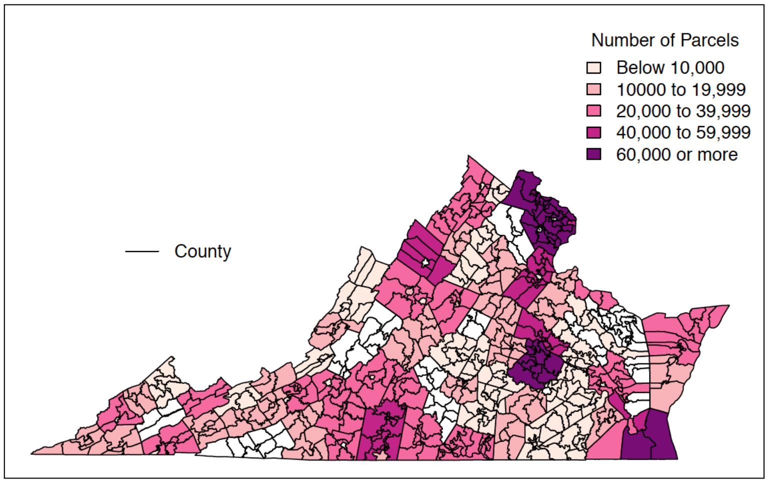
b. James City County

When using BKFS for James City County, we did not have the same insight about their real estate assessment data compared to the insight gained through profiling and preparing Arlington County's original real estate assessment data. Thus, we relied on the same assumptions and rules that were applied to Arlington County data. This can be problematic as each jurisdiction has their own rules and processes with respect to to real estate assessments even within the same state. At the same time, we were dependent on the data BKFS received from their original source. BKFS did not include the county's land use codes and unit counts in the James City County data for any of the years. These data are needed to re-weight the parcels to units. Table 4.4 lists the residential property count only. The standardized land-use codes for BKFS did not provide the necessary detail to identify type of residential property, only that a parcel was in fact residential.

4. CoreLogic Assessment Data

The CoreLogic data came with an extensive data dictionary. Coverage for Virginia ranges from 99 to 112 counties/independent cities, from 2009 and 2014, out of 133 counties/independent cities. Figure 4.4 shows the 2013 coverage. Missing counties were located in rural areas of Virginia.

Figure 4.4: CoreLogic 2013 Parcel Coverage for Virginia



Counties highlighted have some level of parcels included in the data source. (**Source:** CoreLogic 2013.)

a. Arlington County

Initially, we thought the data received from CoreLogic included non-vacant, non-parking lot residential properties. As such, the only step required to get the data to the same level of analysis as the ACS, (i.e., residential properties with housing units) would be to remove a hotel that was listed in the data. Through random parcel checks while cleaning, a parking spot and a common area were found using Arlington Property Search. CoreLogic has standardized land use codes for parking lots, but these codes did not appear in the data received.

While the data for the number of units in a building were complete for the study years, about 30% of the properties changed number of units over 2009-2013 with the range of this difference between 1 to 826 units. There were some single- family-detached housing units which had a unit number other than 0, 1, or NA. For weighting, these were changed to 1 and treated like other single family detached housing units. As seen in Table 4.4, the Arlington County number of housing units was higher than the CoreLogic number across all years.

After weighting and conducting post-cleaning checks, the range of housing value was too large as there were housing units (specifically condominiums) listed as high as $1billion. In addition to their standardized land-use codes, CoreLogic data also included the original county land-use codes but did not include their definitions. Nonetheless, we were able to use our knowledge of the county data to find invalid data.

CoreLogic did include data on geolocation and these data were quite consistent. Between 2009 and 2012, parcels had the same latitude and longitude and 8% had different coordinates in 2013. Yet, the census tract data provided by CoreLogic were unusable due to many invalid entries. As described above, we created tract and block group information using the 2009-2012 latitudes and longitudes.

b. James City County

For James City County, the CoreLogic standardized land codes were used. These provided the necessary detail for the same steps to be taken with these data as with Arlington County data. Just as with Arlington County, the data received only included residential properties. There was no unit count information for 2009. Table 4.4 lists the parcel and unit counts for Core Logic's James City County data. There was one parcel that was classified as a single family home by both the standardized and county land-use, yet the value was over $75million. Looking up this address through the James City County's property search and other online sources, this parcel was found to be a multifamily property and we recoded it as such.

A similar pattern was seen in James City County as was seen in Arlington County regarding latitudes and longitudes. This time though, 100% of the coordinates changed in 2011 and then remained constant for 2012 and 2013. The census tract data provided by CoreLogic was also not reliable for James City County and needed to be recoded.

5. MRIS-MLS Sales Data; Williamsburg MLS Sales Data

The Metropolitan Regional Information Systems (MRIS) data source included Multiple Listing Service (MLS) data for the mid-Atlantic region for 5 years (2009-2013). There was one duplicate observation in the data. Each of the 13,601 observations represented a sale of a housing unit in Arlington County and was identified by a listing ID. Unfortunately, there was not a consistent way to link repeat sales within the data source as some addresses were not consistent across the years.

The Williamsburg area realtors (WMLS) is for James City County and the City of Williamsburg. WMLS has over 10 years of data, from January 2005 through July 2015. This includes 9,717 sales with one duplicate observation.

Codebooks had to be created for both data sources. WMLS data did include parcel numbers, but about 300 parcel IDs (3%) are missing. The lack of parcel number in either data set makes matching to other administrative records databases difficult and time-consuming because matching the address requires formatting to a standard form and fixing errant street suffixes, (e.g., Rd. instead of St.), and prefixes.

6. Location Inc. Data

The Location Inc. data source included neighborhood characteristics by census tract for both Arlington and James City Counties. The data received were: home appreciation, rent, quiet score, walk score, education score, crime indices, and counts and rates for both property and violent crime. Also received were scores for individual schools in the county and the school-census tract crosswalk. The data did come with a minimal codebook, however not much detail was included beyond definition and valid values. These data are patented, patent-pending, exclusive, or proprietary to Location Inc. and the specific process for creating each variable is unknown.

Negligible cleaning/transformation were needed for these data with respect to how they will be used in this project. In the Arlington County data, there were missing values in some of the fields, however these aligned with two tracts with minimal housing units (a tract that surrounds the Arlington Cemetery and the Pentagon and a tract that surrounds Reagan National Airport). Further descriptions of these variables can be found on the collaborative wiki (Keller et al. 2016).

D. Summary

Sixty one external data sources were identified during data discovery and 11 made it through the screening and full inventory process to acquisition. From the 11, the data sources acquired for the study were Arlington County and James City County property data; commercial data from Black Knight Financial Services and CoreLogic (that rely on the local government property data); Multiple Listing Services and associated real estate sales data; and indexes created using proprietary algorithms that describe the livability of neighborhoods from Location Inc. There were many challenges in acquiring these data, primarily in negotiating the terms of the agreements.

Each of the acquired sources of housing data were profiled, cleaned, transformed, restructured, and linked to other variables to prepare the data for benchmarking to ACS tabulations and for analysis (see Chapter 5 and 6). Many challenges were encountered. For example, across the data sources, the profiling and cleaning steps identified the need to delete duplicates and non-residential properties, such as parking lots and vacant parcels. Multiple dwellings on a parcel required restructuring the data so that there was one observation per dwelling. Parcels needed to be transformed to represent unit counts per building. A master address list was created to place the parcels into census geographic boundaries (tracts and block groups) using a consistent set of geocodes from the Google Maps API and rules were created for this purpose. It was obvious that none of the commercial data sources included longitudinal data edits. Overall, the data preparation processes were iterative and non-linear.

5. Housing Comparisons

A. Data Comparisons

A central focus of this study was to determine if ***external*** data, that is, data from outside the Federal Statistical System, can provide estimates comparable to the American Community Survey (ACS) statistics. This chapter presents the benchmarking of the housing data to the 2009-2013 ACS estimates. ACS tabulations used for comparison were at the county, census tract, and block group levels and were accessed through the US Census Bureau's *American FactFinder* (Census Bureau 2013).

It is useful to recognize some fundamental differences between ACS data and the external data sources used in the benchmarking. First, ACS data come from a carefully designed national survey. The data provide estimates and margins of error for several variables of interest to the government and the populace. Methods for weighting and imputation of non-response have been developed and applied to the ACS data to provide official statistical estimates at the national, state, multi-county, county, and sub-county levels.

In contrast, the county-level external data sources acquired for this study were not samples. Rather, these data represent all parcels and housing units known to the jurisdictions. These data could be thought of as the census of the complete population of housing units from which the ACS samples. Therefore, the data do not come with weights for adjustments. For some of the benchmarking, we propose weighting and aggregation schemes to better align the estimates.

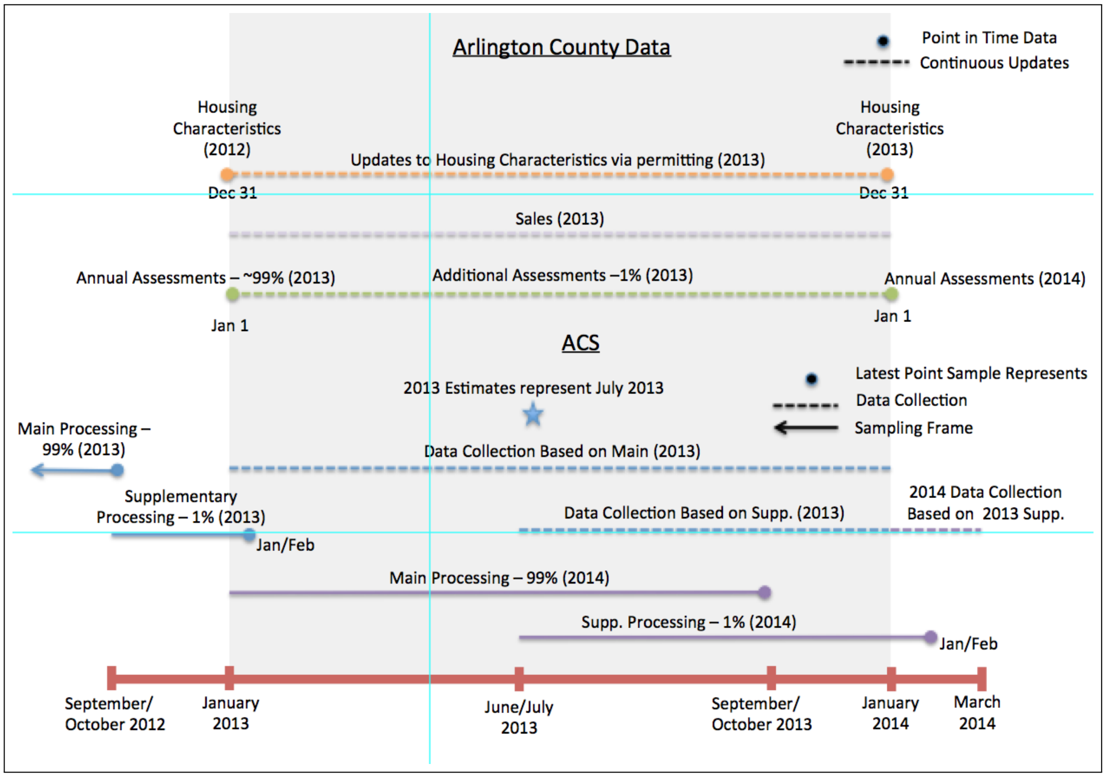
Another difference between the data sources is related to the timing of data collection, which could play a role in the comparisons. The 2013 ACS data and 2013 county data represent different time points within 2013. Figure 5.1 shows the timing for the ACS data collection and Arlington County's tax assessment schedules.

For its main sample processing, the Census selects addresses from its Master Address File (MAF). The MAF is updated twice each year with the Delivery Sequence Files (DSF) provided by the U.S. Postal Service and with updates from various Census Bureau field operations, including the ACS. The MAF that exists in September/October of the previous year accounts for 99% of that year's sample. The MAF is updated a second time in January/February of the sample year and a second/supplementary sample is created. Estimates represent July 1 of that year (Census Bureau 2014).

For the Arlington County real estate assessment data, different data tables represent different time periods. Almost all (99%) of assessment values were timestamped with January 1st of the subsequent year and thus represented the assessed value of the period before that. The remaining

1% of assessments were mid-year assessments and include new construction or reassessments. Housing characteristics were as of December 31 of that year. Changes throughout the year to housing characteristics were not included in the data until the following year, (e.g., a new bedroom was added). Sales data were available throughout the year.

Figure 5.1: Timing of Data Collection


Arlington County Data
Point in Time Data
Continuous Updates
Housing Characteristics (2012)
Housing Characteristics (2013)
Updates to Housing Characteristics via permitting (2013)
Dec 31
Dec 31
Sales (2013)
Annual Assessments – ~99% (2013)
Additional Assessments –1% (2013)
Annual Assessments (2014)
Jan 1
Jan 1
ACS
Latest Point Sample Represents
Data Collection
Sampling Frame
2013 Estimates represent July 2013
Main Processing – 99% (2013)
Data Collection Based on Main (2013)
Supplementary Processing – 1% (2013)
Data Collection Based on Supp. (2013)
2014 Data Collection Based on 2013 Supp.
Jan/Feb
Main Processing – 99% (2014)
Supp. Processing – 1% (2014)
Jan/Feb
September/ October 2012
January 2013
June/July 2013
September/ October 2013
January 2014
March 2014

Comparison of 2013 data sampling frame and data collection of ACS and Arlington County real estate assessment data.

One reason for differences between the timing of the ACS and Arlington County real estate assessment data collections was the rate of change, due to new construction and demolitions, of housing units between the two MAF updates. In Arlington County real estate assessment data, there were between 11 to 1,017 *new* housing units (weighted) each year between 2009 and 2013. Looking at the Metropolitan Regional Information Systems, Inc. (MRIS) data, 62 to 88 new housing units (those identified with a year built the same year they were sold) were sold each year between 2009 and 2013. Thus, the difference in timing between the ACS and Arlington County real estate assessment data appears to have a small effect on the count of housing units.

The following sections provide examples of housing benchmarks, highlighting anomalies, nuances, and features of the data sources. The collaborative wiki (Keller et al. 2016 has an exhaus-tive set of benchmarks for the ACS tables listed in Table 5.1.

Table 5.1: ACS Housing Tables used for Benchmarking Comparisons

		Arlington County			
ACS Table		**BKFS**	**CoreLogic**	**County**	**MRIS / MLS**
B25001 Housing Units	5-year		X	X	
	1-Year	X	X	X	
B25003 Occupancy Status	5-Year		X		
	1-Year	X	X		
B25024 Units in Structure	5-Year		X	X	
	1-Year	X	X	X	
B25034 Year Structure Built	5-Year		X	X	X
	1-Year	X	X	X	X
B25035 Median Year Structure Built	5-Year		X	X	X
	1-Year	X	X	X	X
B25041 Bedrooms	5-Year		X	X	X
	1-Year	X	X	X	X
B25074 Value	5-Year		X	X	X
	1-Year	X	X	X	X
B25077 Median Value (Dollars)	5-Year		X	X	X
	1-Year		X	X	X
B25102 Real Estate Taxes Paid	5-Year		X	X	
	1-Year	X	X	X	
		James City County			
ACS Table		**BKFS**	**CoreLogic**	**County**	**WMLS / MLS**
B25001 Housing Units	5-Year				
	1-Year		X	X	
B25003 Occupancy Status	5-Year				
	1-Year		X		
B25024 Units in Structure	5-Year				
	1-Year		X		
B25034 Year Structure Built	5-Year				
	1-Year		X	X	X
B25035 Median Year Structure Built	5-Year				X
	1-Year		X	X	X
B25041 Bedrooms	5-Year				X
	1-Year		X	X	X
B2507 Value	5-Year		X		X
	1-Year		X		X
B25075 Median Value (Dollars)	5-Year		X		X
	1-Year		X	X	X
B25102 Real Estate Taxes Paid	5-Year		X		
	1-Year		X		

Source: US Census Bureau's *ACS tables are from the American FactFinder* (Census Bureau 2013).

The following ratio was used as a figure of merit or fitness throughout the comparisons:

$$Fitness\ Ratio = \frac{ACS\ estimate - External\ estimate}{90\%\ ACS\ margin\ of\ error}$$

When the *Fitness Ratio* falls outside the ±1 range, the benchmark estimates were not within the 90% ACS margin of error. Falling outside the margin of error does not mean the estimates were bad. The ground truth or the gold standard is unknown for the quantities being estimated. Therefore, estimates that do not match could be equally valid or invalid. The question that needs to be answered is whether they are useful for the intended purposes or not.

B. Housing Units

Arlington County real estate assessment data was used in Chapter 4 as the comparison example to highlight differences in raw counts of housing units. The same comparisons for James City County are presented on the collaborative wiki (Keller et al. 2016). James City County real estate assessment data were only available for 2014. Therefore, we used CoreLogic data for James City County comparisons.

Table 5.2 presents comparisons of the number of housing units in Arlington County based on each dataset by year for the 2009-2013 time period. The number of housing units for the county data, Black Knight Financial Services (BKFS), and CoreLogic were obtained by weighting the parcels by the number of units in each parcel. This was problematic because of the number of properties with missing number of unit counts, which ranged from 206 parcels in 2009 to 463 in 2013 in the Arlington County real estate assessment data. This could result in underestimation of the total housing units if a large fraction of these missing data are associated with multifamily dwellings. Chapter 4 provided a description of the problems with counting number of units. Here, the Arlington County housing unit counts from the county real estate assessment data and CoreLogic were below the 90% confidence intervals for the ACS estimates.

Table 5.2: Estimate of Housing Units by Data Source

	Arlington County					
	2009	2010	2011	2012	2013	2009-2013
ACS	103,813	105,490	106,720	107,734	109,689	106,740
MOE	±872	±619	±417	±537	±504	±191
County data	100,991	101,867	102,299	102,299	103,987	102,331
BKFS	–	–	–	–	110,883	–
CoreLogic	97,589	98,690	83,640	79,521	79,804	87,849

	James City County					
	2009	2010	2011	2012	2013	2009-2013
ACS	–	29,882	30,282	30,623	30,986	30,253
MOE	–	±391	±199	±246	±277	±84
County data	–	–	–	–	–	–
BKFS	–	–	–	–	–	–
CoreLogic	–	26,013	26,361	26,832	27,294	–

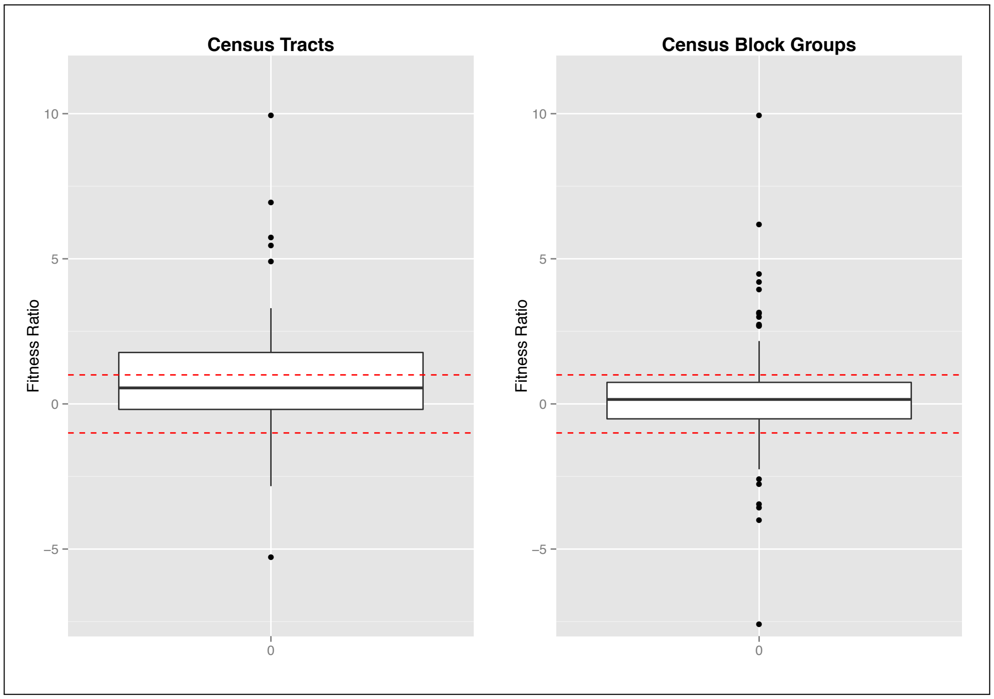
Note: MOE is the 90% ACS margin of error. The – indicates missing data. Housing units for external data were weighted by the number of units. Black Knight Financial Services (BKFS) does not have weighted James City county numbers as there were no unit counts. James City County Real Estate Assessment data only exists for 2014. ACS tabulations for James City County 2009 1-year estimates of housing units were unavailable.

Sources: ACS 2009-2013, Arlington County Real Estate Assessment Data 2009-2013, BKFS 2009-2013, CoreLogic 2009-2013.

Table 5.2 also includes comparisons of the housing units for James City County. Recall that the county real estate assessment data were only available for 2014. BKFS and CoreLogic data span the study period. However, unit counts were completely missing for BKFS. Unit counts associated with multifamily properties are also a problem for the James City data. CoreLogic had 132 multifamily properties with missing, 0 or 1 units associated with them. This lead to underestimating the number of housing units. The unit counts from CoreLogic were all below the 90% confidence interval for the ACS estimates.

The differences in housing unit counts were also examined across census tracts and block groups. There are 59 census tracts containing 181 block groups in Arlington County. Figure 5.2 provides box-plots for the *Fitness Ratios* across census tracts and block groups. While the median of the ratios for census tracts (0.66) falls within the ±1 interval, 22 census tracts had *Fitness Ratios* > 1 for housing unit counts. There were 5 tracts for which the $|$*Fitness Ratio*$| >$ 5.0. Housing unit counts for block groups had more comparable estimates (Keller et al. 2016).

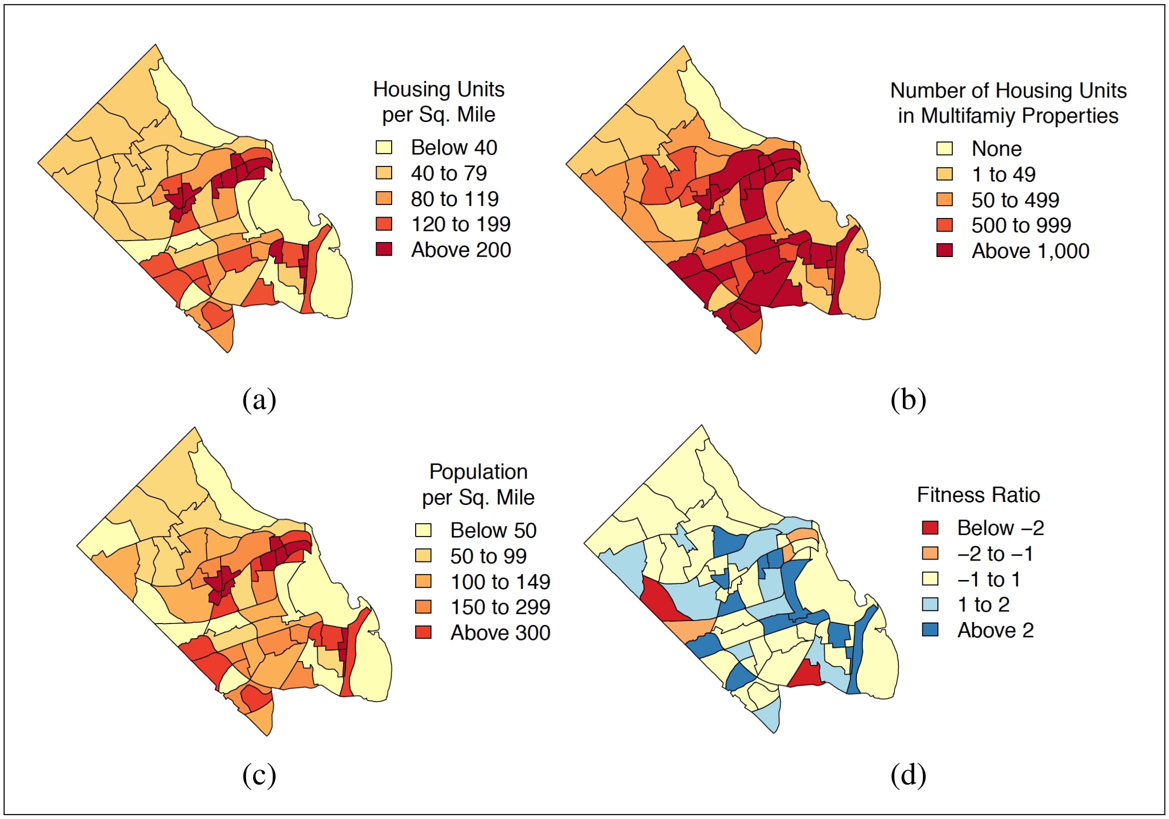
Figure 5.2: Fitness Ratios for Housing Units



Boxplots compare the distribution of the *Fitness Ratios* at the census tract and block group levels and their relation to the 90% ACS margin of errors. Estimates falling outside the red reference lines of ±1 were not within the 90% ACS margins of error. (**Sources:** Arlington County Real Estate Assessments 2009-2013 5-year estimates, ACS 20009-2013 5-year estimates.)

Census tracts with $|$*Fitness Ratio*$| > 1$ were conjectured to be aligned with housing unit density within Arlington County. Arlington County real estate assessment data appears to have underestimated multifamily buildings due to either different or missing units counts for multifamily buildings. To study this conjecture exploratory analysis was conducted to see if unit counts lower than the ACS estimate, (i.e., *Fitness Ratio* > 1), were occurring in areas with high multifamily housing units density. Figure 5.3 visually compares these densities.

Figure 5.3: Comparison of the Fitness Ratio for Housing Units to Densities of Housing Units, Multifamily Buildings, and Population Density.



Panel (a) is the the geographic distribution of housing unit density, (b) is the geographic distribution of the number of multifamily properties, (c) is geographic distribution of the populations density, and (d) is the geographic distribution of housing unit count *Fitness Ratios*. (**Sources:** Arlington County Real Estate Assessment Data 2009-2013 5-year estimates (d), ACS 20009-2013 5-year estimates (b,d), 2010 Census (a,c).)

The reasons for the differences between unit counts in the ACS estimates and the local administrative data were elusive. These differences appear to be in geographic areas with high population and housing unit density. Although these total counts may be off, the actual composition of the housing units was explored further and is discussed in the following sections.

C. Units in Structure

The next comparisons involve the distributions associated with the units in a structure. Table 5.3 presents the housing unit counts by type of the structure in the different data sources for Arlington County. The counts in the cells highlighted in green fall within the 90% ACS margins of error.

Table 5.3: Number of Housing Units in a Structure for Arlington County, 2013

Data	Weighted N	1-Attached	1-Detached	2	3 or 4	5 to 9	10 to 19	20 to 49	50 or more	Unknown
ACS	109,689	10,358	28,436	756	3,068	5,792	9,893	7,010	44,159	0
County	103,987	6,024	27,518	542	12	334	1,220	4,635	63,239	463
BKFS	110,883	32,025	27,544	2	75	962	1,320	3,215	45,356	384
CoreLogic	79,804	12,160	27,439	746	2,585	3,494	3,712	1,820	25,718	2,130

Note: Green highlights estimates that fall within 90% ACS margins of error. The difference in “1-Attached estimates between County and Black Knight Financial Services (BKFS) data are the result of condominiums being coded 1-attached as there was no unit number nor could it be imputed.
Sources: ACS 2013, Arlington County Real Estate Data 2013, BKFS 2013, CoreLogic 2013.

There were some major differences across the sources in the “1-attached” structures, which was likely due to differences in how the ACS and the county code condominiums. The ACS posed the following question to the respondents *“Which best describes this building?”* and listed the categories to be chosen from. For tax purposes, Arlington County created a different parcel ID for each condominium in a multifamily structure without any information about the size of the structure.

The BKFS count for “1-attached” seemed to reflect the number of independent parcel IDs created by the county for condominiums. As described in Chapter 4, adjustments to the county real estate assessment data were made to address this issue. For the county real estate assessment data, we aggregated parcels by GIS codes. Each parcel was assigned a corresponding GIS code, which allows for the data to be merged with a GIS parcel shapefile where each polygon represents a parcel or building. For condominiums, the same polygon was present for all condominiums in that building, (i.e., a many-to-one relationship). Thus, for the county real estate assessment data, we imputed unit counts for condominiums and classified them accordingly.

Table 5.4 presents the distribution of units in structure from the 5-year estimates based on ACS and the Arlington County real estate assessment data. We observe that estimates outside the 90% ACS margins of error correspond to multi-unit buildings. The comparison results were similar for the 1-year estimates (Keller et al. 2016).

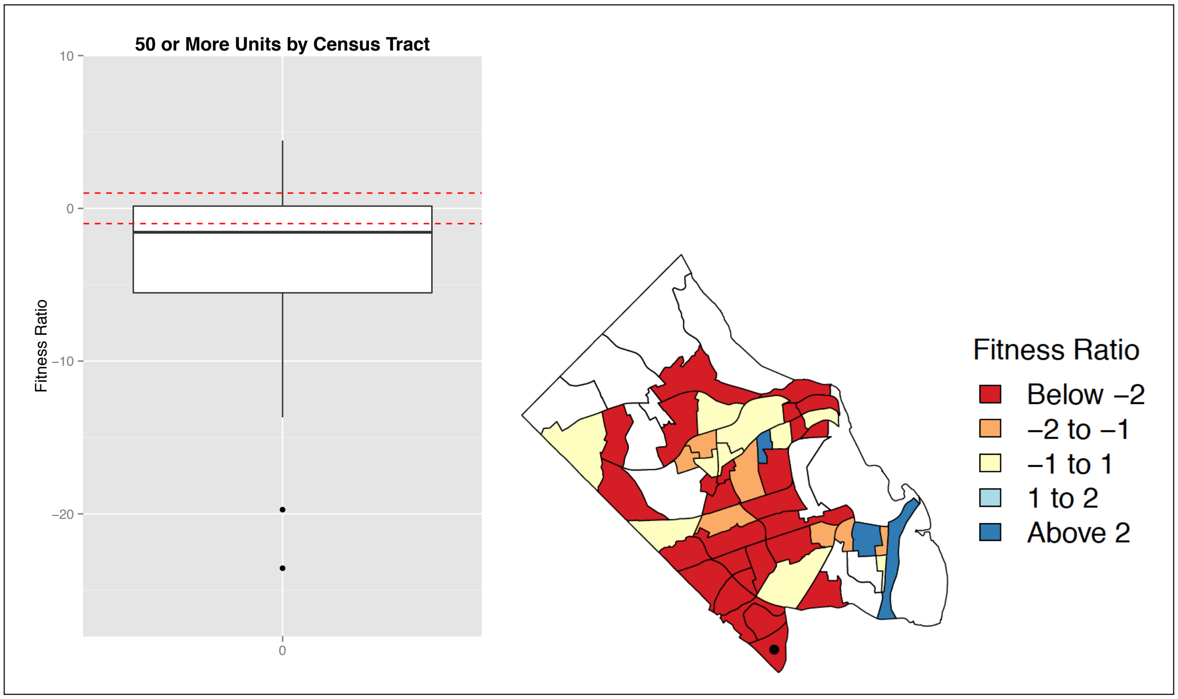
The largest difference occurs for structures with 50 or more housing units. The boxplot for the *Fitness Ratios* and the geographic distribution of the percent of structures with 50 or more units at the census tract level are illustrated in Figure 5.4. Large differences correspond to South Arlington where there were more multifamily properties.

Table 5.4: Comparison of ACS Table B25024 (Units in Structure) with Arlington County Data, 2009-2013

	ACS Benchmark		Direct Estimate from Arlington County		
Units in Structure	Estimate	90%MOE	Estimate	Difference	Within 90% MOE?
1–Attached	9.42%	5.17%	5.78%	3.64%	YES
1–Detached	26.91%	5.47%	26.73%	0.17%	YES
2	0.97%	2.53%	0.53%	0.44%	YES
3 or 4	3.68%	3.09%	0.07%	3.61%	NO
5 to 9	5.46%	4.24%	0.93%	4.54%	NO
10 to 19	7.72%	5.62%	1.83%	5.89%	NO
20 to 49	5.19%	4.59%	5.08%	.10%	YES
50 or more	40.27%	7.31%	58.78%	-18.51%	NO

Note: MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error.
Sources: Arlington County Real Estate Assessment Data 2009-2013 5-year estimates, ACS 2009-2013 5-year estimates.

Figure 5.4: Distributions of Fitness Ratios for Structures of 50 or More Units



Map is drawn by census tract and shows the geographic distribution of the *Fitness Ratios*. White areas in the map correspond to tracts with no structures of 50 or more housing units. Estimates falling outside the reference lines of ±1 were not within the 90% ACS margins of error. One extreme lower outlier (-134.74) was removed from the boxplot. It corresponds to the tract designated with a point at the bottom left. (**Sources:** Arlington County Real Estate Assessment Data 2009-2013 5-year estimates, ACS 2009-2013 5-year estimates)

D. Year Structure Built

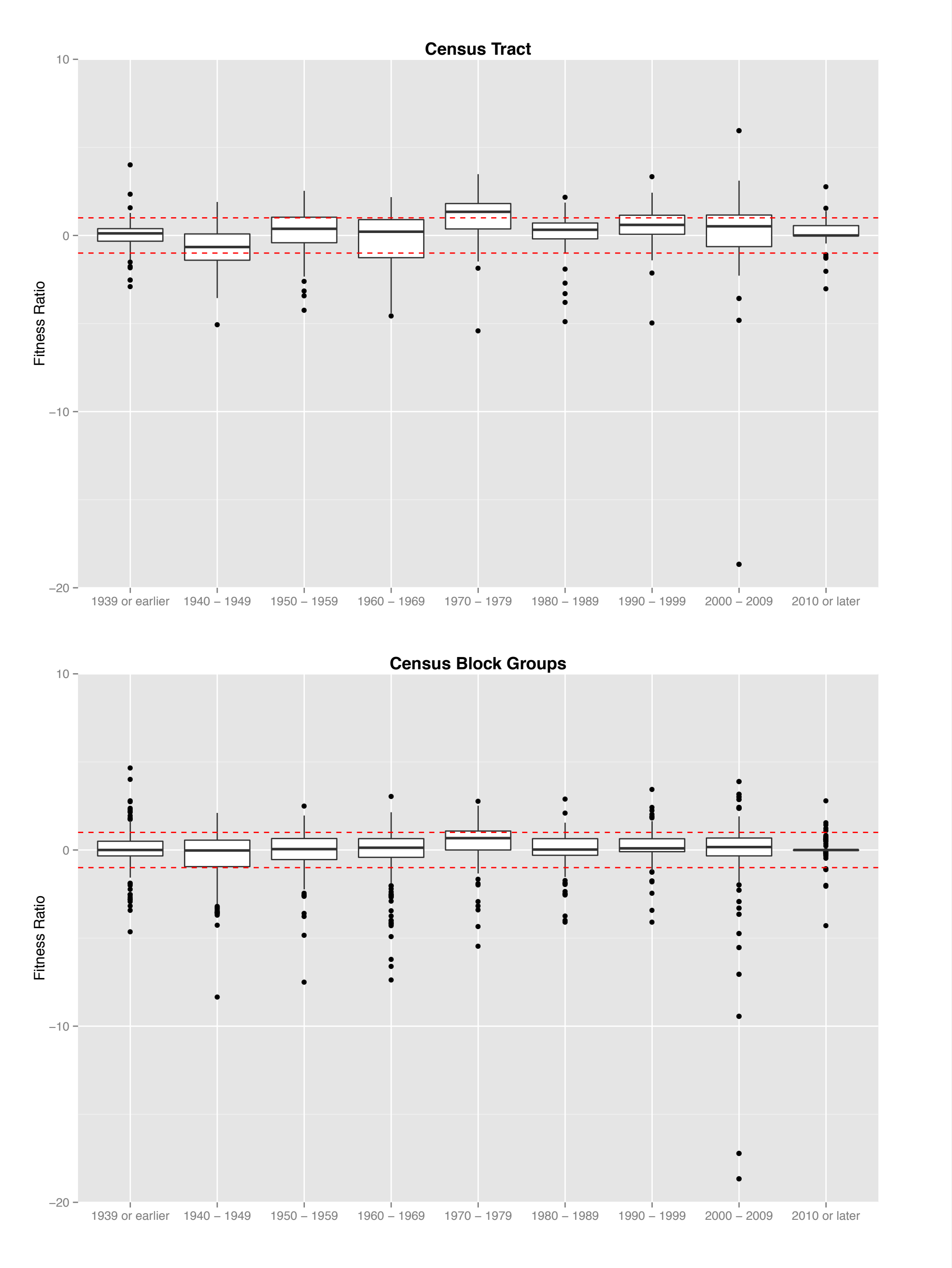
Respondents to the ACS are asked *"About when was this building first built?"* and are given categories of year-groupings to choose among. Yet, many may not know the exact answer to this question, especially those living in apartments. The evidence provided in this section for

Arlington and James City Counties suggests that “Year Structure Built” may be better captured through local data sources than through ACS data collection.

A comparison of the Arlington County real estate assessment data to ACS’s 5-year estimates of housing units by year structure built resulted in some comparable estimates. Figure 5.5 present a more detailed look at the year built data. The figure provides the *Fitness Ratio* distributions for the census tracts and the block groups. Many of the *Fitness Ratios* fall within ±1 range, in both cases.

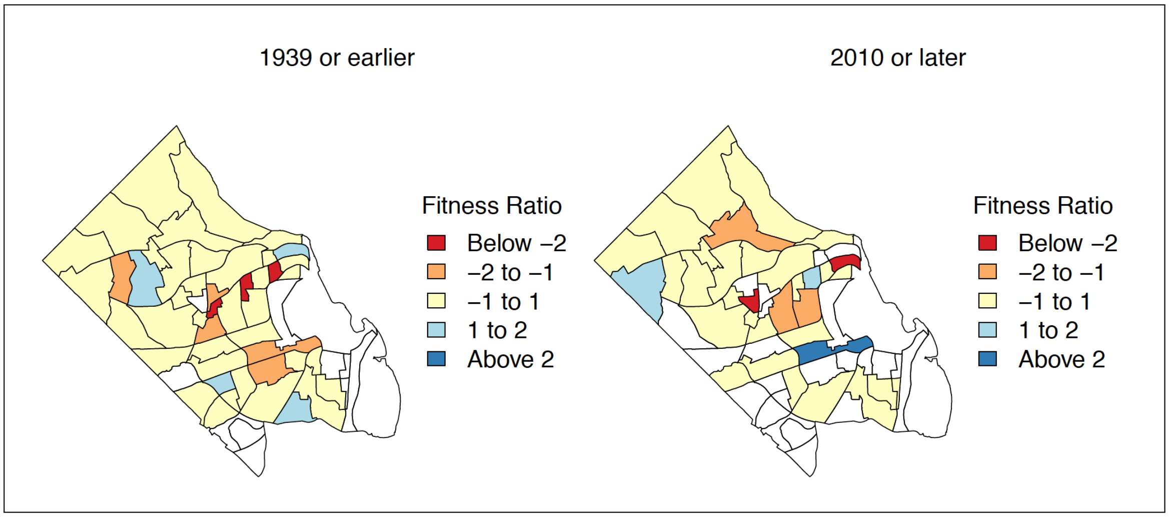
We conjecture that the county data, either directly from the county or through a data aggregator, may provide more accurate values for “year structure built” than the ACS. Figure 5.6 illustrates the magnitudes of the *Fitness Ratios* by census tracts, comparing the oldest (pre-1939) and the youngest (post-2010) structure ages in Arlington County. The ACS data had lower volume of both younger and older housing unit ages in the center of the county when compared to the county real estate assessment data. This is an area with high growth of both residential and commercial properties, has many renter-occupied housing units, and follows the metro/subway line. ACS showed a lower volume of young structures in a region of lower income as compared to the county real estate assessment data. Those census tracts had a high volume of renter-occupied housing units according to the ACS.

Figure 5.5: Comparison of *Fitness Ratios* for Year Built Across Census Tracts and Census Block Groups for Arlington County, 2009-2013



Boxplots compare the distribution of the *Fitness Ratios* at the census tract and block group level. Estimates falling outside the red reference lines of ±1 were not within the 90% ACS margins of error. One extreme lower outlier (-35.78) was removed from the block group boxplot. (**Sources:** Arlington County Real Estate Assessments 2009-2013 5-year estimates, ACS 20009-2013 5-year estimates)

Figure 5.6: Fitness Ratios for Oldest and Youngest Housing Units



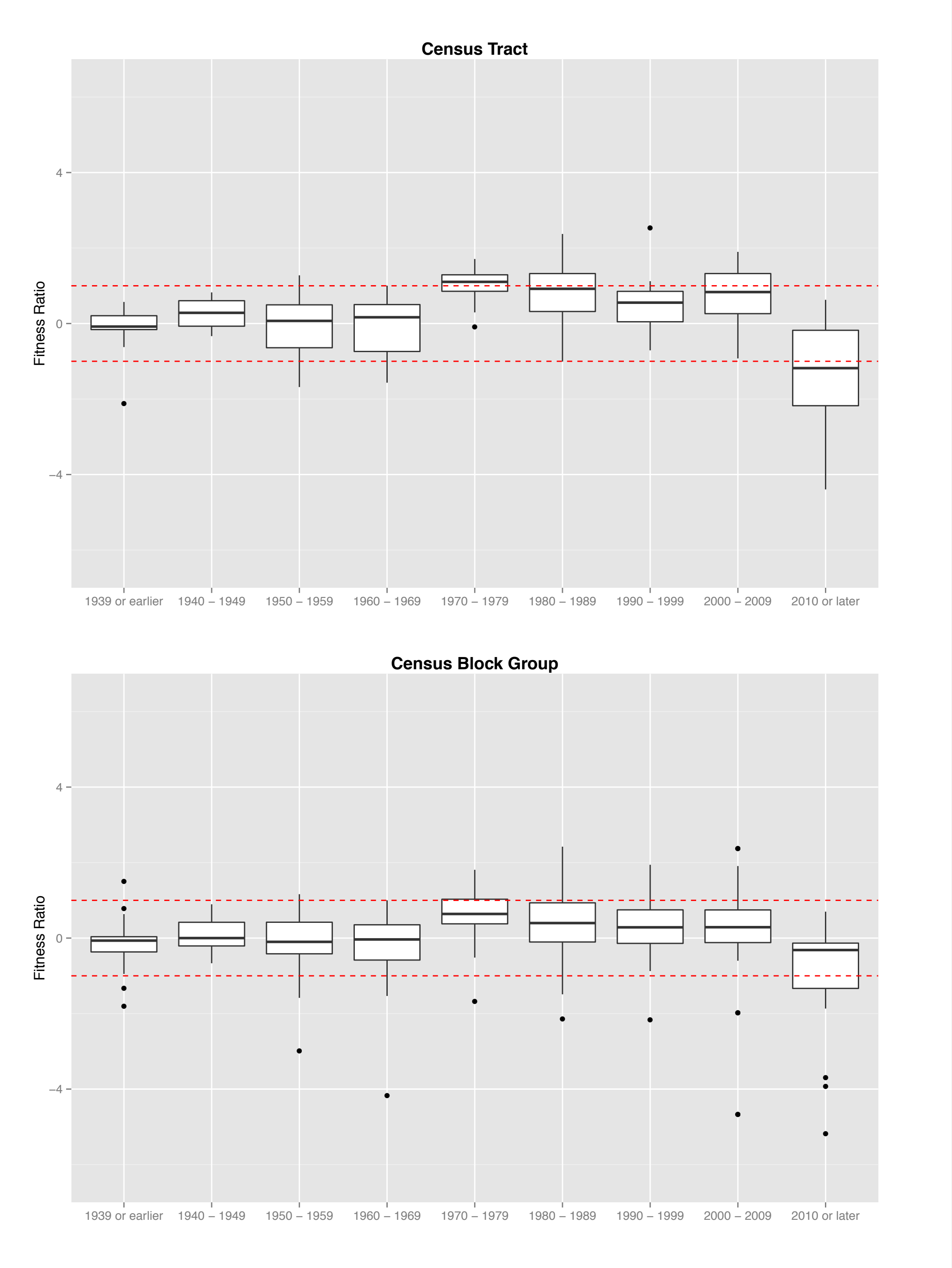
Maps for Youngest and Oldest Housing Units. Maps were drawn by census tract. White areas in the map correspond to tracts with no housing units in that category. (**Sources**: Arlington County Real Estate Assessments 2009-2013 5-year estimates, ACS 2009-2013 5-year estimates)

James City County has similar results for year built. Corelogic data was used for the comparisons because the James City County assessment data was only available for 2014. Figure 5.7 shows the *Fitness Ratios* for the 5-year estimates across 11 census tracts and 27 block groups in James City County. CoreLogic has higher estimates of structures built in 2010 or later as compared to the ACS estimates. James City County has seen a sharp increase in population and housing units during the last decade. According to the 2010 Census, James City County had an increase of 43% in housing units between 2000-2010. ACS data from 2010 and 2013 indicate there has been approximately a 4% growth in housing units in James City County during that period. The local data may be capturing the housing unit age of this growth better that the ACS.

Looking at the "Median Year Built" provided some interesting results. For Arlington County the 5-year estimate for median year built was 1961 based on the county real estate assessment data versus the corresponding ACS value of 1968 (±2). The Arlington County real estate assessment data estimate of the median year built has remained very consistent over the time period. It was 1961 between 2010 and 2013 and 1960 for 2009. However, the ACS single-year estimates between 2009 and 2013 have steadily risen from 1965 (±2) to 1973 (±2).

The "Median Year Built" for James City County is 1994 based on CoreLogic 5-year estimate and is comparable to the ACS estimate of 1993 (±2). Similar to Arlington County data, the yearly estimates of median year built for James City County were very consistent. Based on CoreLogic data it was 1994 for 2009-2011 and 1995 for 2012-2013. The ACS 1-year estimates of steadily increased between 2010-2012 from 1991 (±2) to 1994 (±3), but then dropped in 2013 to 1991 (±2). A corresponding 1-year ACS estimate for 2009 was not available.

Figure 5.7: Comparison of *Fitness Ratios* for Year Built Across Census Tracts and Census Block Groups, James City County, 2009-2013



Boxplots compare the distribution of the *Fitness Ratios* at the census tract and block group level. Estimates falling outside the red reference lines of ±1 were not within the 90% ACS margins of error. (**Sources:** James City County CoreLogic data 2009-2013 5-year estimates, ACS 20009-2013 5-year estimates)

E. Bedrooms

The distribution of the housing units by the number of bedrooms based on the ACS data compared to Arlington County real estate assessment data (5-year estimates) is presented in Table 5.5. In the ACS, the question regarding the bedroom counts is posed as *"How many of these rooms are bedrooms? Count as bedrooms those rooms you would list if this house, apartment, or mobile home were for sale or rent. If this is an efficiency/studio apartment, print '0'."* A set of rules within Virginia State Building Code determines whether a room is classified as a bedroom in the county real estate assessment data (Virginia State 1996). For example, a bedroom must have at least one operable emergency escape and rescue opening. An occupant responding to ACS might not fully know whether the bedroom is actually considered as a bedroom under Virginia State codes.

Table 5.5: Comparison of ACS Table B25041 (Bedrooms) with Arlington County Data, 2009-2013

	ACS Benchmark		Direct Estimate from Arlington County		
Number of Bedrooms	Estimate	90%MOE	Estimate	Difference	Within 90% MOE?
0 bedrooms	4.29%	0.49%	13.17%	-8.88%	NO
1 bedrooms	33.79%	0.91%	18.14%	15.64%	NO
2 bedrooms	29.64%	1.00%	28.52%	1.12%	NO
3 bedrooms	18.14%	0.73%	26.03%	-7.89%	NO
4 bedrooms	9.54%	0.51%	9.62%	-0.08%	YES
5 or more bedrooms	4.61%	0.42%	4.48%	0.13%	YES

Note: MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error.
Sources: Arlington County Real Estate Assessments 2009-2013 5-year estimates, ACS 2009-2013 5-year estimates.

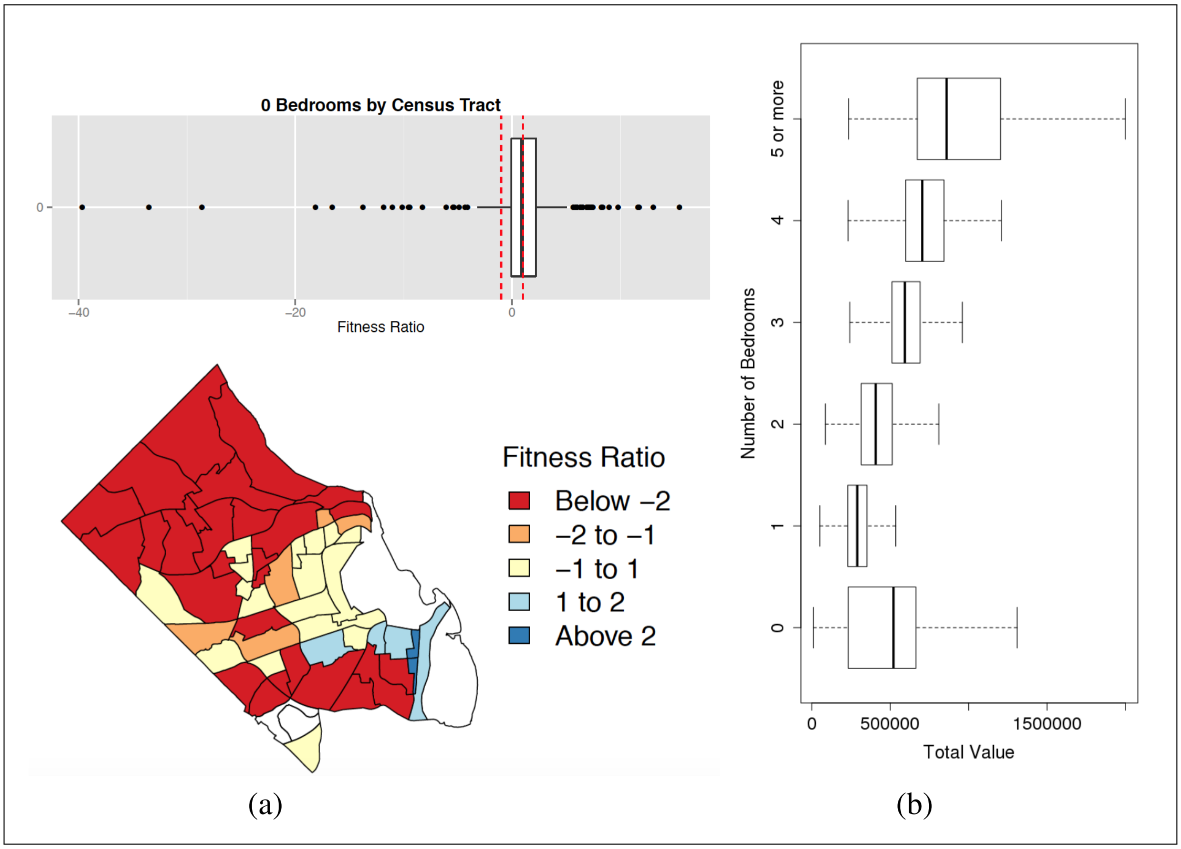
The estimate for 1-bedroom housing units in the Arlington County real estate assessment data was lower than the ACS estimate. This was expected since the housing units in multifamily buildings were underestimated as mentioned earlier. The counts for housing units with 4 and 5+ bedrooms were not affected because there were not as many housing units of these sizes in multifamily buildings.

The county estimate for 0-bedroom units was higher than the ACS estimate. This could be due to the fact that the county real estate assessment data did not include information on individual apartments that were "solely renter-occupied" units. In addition, the definition of a bedroom based on the state code excludes bedrooms that might appear in the ACS data.

Figure 5.8 displays the *Fitness Ratios* for 0-bedroom housing units by census tract. This reveals that most of the discrepancies between ACS and county real estate assessment data corresponds to North Arlington, an area that consists of primarily high-valued single-family detached homes. This was also observed in some tracts in South Arlington, which is a residential neighborhood around the Pentagon.

The “Total Value” boxplots in Figure 5.8 illustrate the distribution of housing values for all categories of the number of bedrooms. We observed that the values for 0-bedroom units compared to the other units were larger than expected, which might explain the discrepancies in the estimates of 1-bedroom housing units in Table 5.5. This implies that some of the 0-bedroom units in the county real estate assessment data may correspond to 1- and 2-bedroom units in the ACS.

Figure 5.8: Distributions of Housing Value by Number of Bedrooms



(a) (b)

Panels provide the *Fitness Ratio* distributions of housing values by number of bedrooms Panel (a) is the boxplot and geographic distribution of the *Fitness Ratios* for “0–bedroom homes” by census tract. White areas in the map correspond to tracts that do not have housing units with zero bedrooms. Estimates falling outside the red reference lines of ±1 were not within the 90% ACS margins of error. County real estate assessment data estimates were larger than the ACS estimates and were concentrated in North Arlington, where there are many single family, high-valued detached homes. Panel (b) is a collection boxplots displaying the housing “Total Value” distributions by number of bedroom in the house. The distribution of the “Total Value” for properties listed as ‘0–bedroom homes” was much wider that expected. (**Sources:** Arlington County Real Estate Assessments 2009-2013 5-year estimates; ACS 2009-2013 5-year estimates)

F. Value and Taxes

The remaining table comparisons presented in this chapter focus on owner-occupied housing units. One difficulty in benchmarking these tables, (i.e., value and taxes paid), was the potential difference in universe between the ACS and the county real estate assessment data. The ACS estimates a housing value for all owner-occupied units. Neither Arlington County nor James City County had indicators to determine if a housing unit was owner-occupied. BKFS and CoreLogic imputed an indicator of owner occupancy. BKFS assigned the indicator by matching the address of the unit with the address of the real estate taxpayer. CoreLogic used proprietary logic to determine if the property owner resides at the property.

1. Housing Value

ACS respondents who own their housing unit are asked, *"About how much do you think this house and lot, apartment, or mobile home (and lot, if owned) would sell for if it were for sale?"* This question could be difficult to answer if the owners bought the housing unit a long time ago.

Table 5.6: Comparison of ACS Table B25075 (Value) with Arlington County Data, 2009-2013

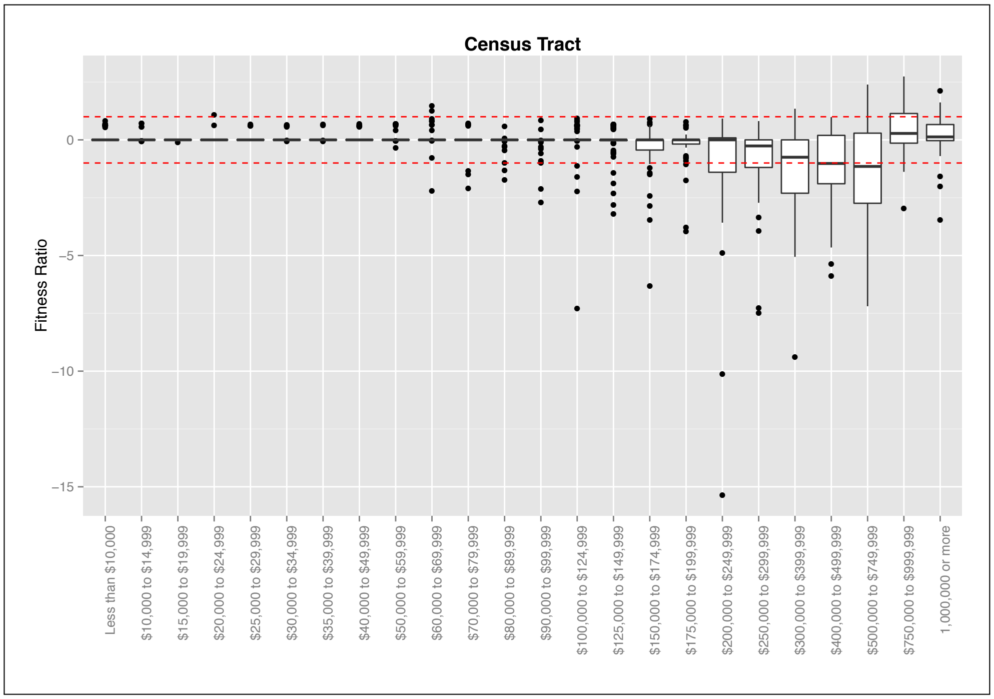
	ACS Benchmark		Direct Estimate from Arlington County		
Value	Estimate	90%MOE	Estimate	Difference	Within 90% MOE?
Less than $10,000	0.25%	0.12%	0.00%	0.25%	NO
$10,000 to $14,999	*0.05%	0.05%	0.00%	0.04%	YES
$15,000 to $19,999	*0.00%	0.07%	0.00%	0.00%	YES
$20,000 to $24,999	0.16%	0.13%	0.00%	0.16%	NO
$25,000 to $29,999	*0.11%	0.11%	0.00%	0.11%	YES
$30,000 to $34,999	*0.11%	0.11%	0.00%	0.11%	NO
$35,000 to $39,999	*0.07%	0.08%	0.00%	0.07%	YES
$40,000 to $49,999	0.16%	0.13%	0.00%	0.16%	NO
$50,000 to $59,999	0.14%	0.08%	0.01%	0.13%	NO
$60,000 to $69,999	0.26%	0.17%	0.08%	0.18%	NO
$70,000 to $79,999	0.19%	0.15%	0.11%	0.08%	YES
$80,000 to $89,999	*0.12%	0.13%	0.23%	-0.10%	YES
$90,000 to $99,999	0.07%	0.06%	0.28%	-0.21%	NO
$100,000 to $124,999	0.66%	0.25%	1.65%	-0.98%	NO
$125,000 to $149,999	0.65%	0.28%	1.52%	-0.88%	NO
$150,000 to $174,999	0.80%	0.25%	1.35%	-0.55%	NO
$175,000 to $199,999	1.11%	0.33%	1.51%	-0.40%	NO
$200,000 to $249,999	3.33%	0.58%	4.94%	-1.61%	NO
$250,000 to $299,999	5.54%	0.91%	6.97%	-1.43%	NO
$300,000 to $399,999	13.30%	1.17%	16.67%	-3.38%	NO
$400,000 to $499,999	11.88%	1.04%	14.05%	-2.17%	NO
$500,000 to $749,999	32.64%	1.69%	35.41%	-2.78%	NO
$750,000 to $999,999	18.45%	1.25%	8.74%	9.70%	NO
$1,000,000 or more	9.96%	0.92%	5.20%	4.76%	NO

Note:MOE is margin of error. The * indicates that the ACS estimate was not significantly different from zero at the 90% confidence level. Green highlights estimates that fall within 90% ACS margins of error.
Sources: Arlington County Real Estate Assessments 2009-2013 5-year estimates, ACS 2009-2013 5-year estimate

Table 5.6 presents the comparison between ACS and Arlington County real estate assessment data for housing value. Given the lack of an indicator for owner-occupancy in the county real estate assessment data, the housing unit counts were expected to be higher than in the ACS data because the county real estate assessment data included both renter- and owner-occupied units. For Arlington County, the difference in the total number of units between the county real estate assessment data and ACS owner-occupied count was 16,832, which corresponded to 28% of the 2009-2013 county real estate assessment data.

We might expect the distributions associated with with high housing values to be similar, but they were not. The county real estate assessment data showed larger counts and percentages of higher-priced homes. This pattern is also present in the census tract *Fitness Ratios* given in Figure 5.9. As this was related to taxes collected by the county, it seemed plausible that the county real estate assessment data may be more accurate than the ACS although this needs more study to understand the differences.

Figure 5.9: Distribution of Fitness Ratios for Housing Value in Arlington County, 2009-2013



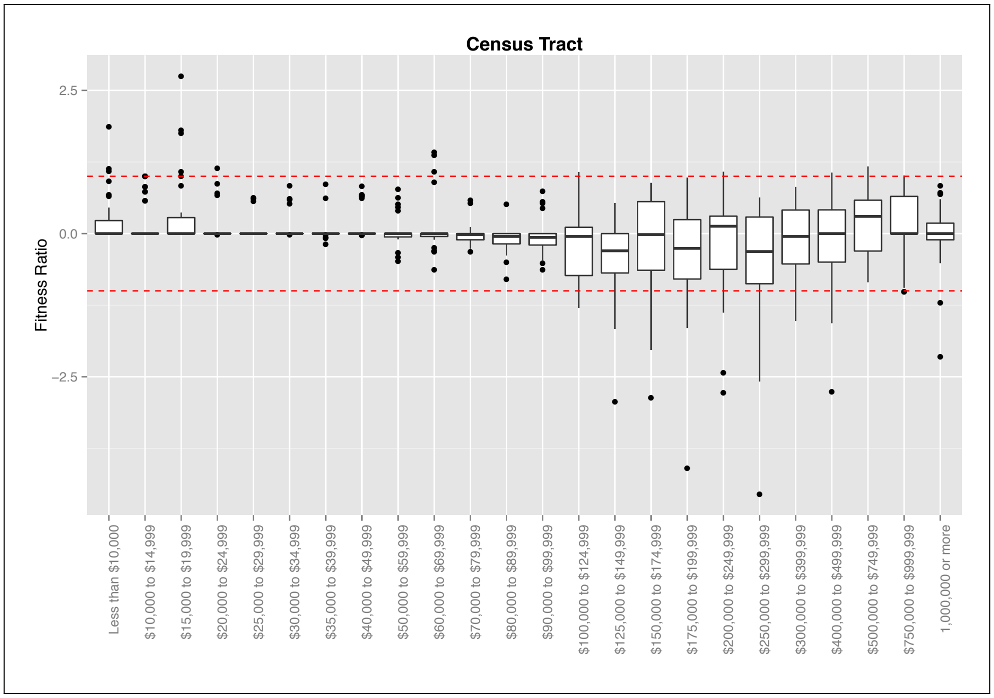
Boxplots display the distributions of the *Fitness Ratios* at the census tract level. Estimates falling outside the red reference lines of ±1 were not within the 90% ACS margins of error. Two extreme lower outliers (-32.19 for $100,000 to $124,999 and -35.12 for $150,000 to $174,999) were removed from the boxplots. (**Sources:** Arlington County Real Estate Assessments 2009-2013 5-year estimates, ACS 20009-2013 5-year estimates)

Turning to James City County presents some interesting findings. While Arlington County data does not differentiate between owner- and renter-occupied units, CoreLogic does calculate an absent-owner flag. Using this, estimates can be created of housing value for just owner-occupied units. CoreLogic has three different value variables: assessed, market,and appraised values. CoreLogic uses these three values to estimate a "calculated total value." However, the three values are not always available. This depends on the source of data Corelogic receives from each local jurisdiction. Arlington County only had assessed values available. James City

County had both assessed values and market values. However, all of the James City County's assessed values were equal to their market values.

Figure 5.10 shows the *Fitness Ratio* for the 5-year estimates across census tracts in James City County based on CoreLogic data. Across all categories, James City's estimates were more comparable to the ACS estimates than was the case for the Arlington County real estate assessment data (see Figure 5.9). ACS had higher estimates of lower valued housing units in James City County. The differences between counties might be the result of how the different jurisdictions assess their properties.

Figure 5.10: Distribution of Fitness Ratios for Housing Value in James City County, 2009-2013



Boxplots display the distributions of the *Fitness Ratios* at the census tract level. Estimates falling outside the red reference lines of ±1 were not within the 90% ACS margins of error. (**Sources:** James City County CoreLogic data 2009-2013 5-year estimates, ACS 20009-2013 5-year estimates)

Housing value could also be estimated using transaction-based Multiple Listing Services (MLS) data. The number of housing units sold in the open market via an MLS listing each year was considerable, about 5% of single-family homes in Arlington County (MRIS) and 3% in James City County (WMLS). Tables 5.7 and 5.8 provide ACS comparisons for the the distributions of sales price data in Arlington and James City Counties, respectively. A comparison of Table 5.7 to Table 5.6 showed that the MRIS has better alignment to the ACS data in the $90,000 to $124,999 range.

MLS data has the potential to act as a sample from which local estimates can be created. Yet, there can be some sample size and selection issues, (e.g., whether the sample of homes sold are actually representative of the housing stock as a whole), that must be addressed. This is recommended for future study.

Table 5.7: Comparison of ACS Table B25075 (Value) with MRIS, Arlington Country, 2009-2013

	ACS Benchmark		Direct Estimate from Arlington County		
Value	Estimate	90%MOE	Estimate	Difference	Within 90% MOE?
Less than $10,000	0.25%	0.12%	0.03%	0.22%	NO
$10,000 to $14,999	0.05%	0.05%	0.00%	0.05%	YES
$15,000 to $19,999	*0.00%	0.07%	0.00%	0.00%	YES
$20,000 to $24,999	0.16%	0.13%	0.00%	0.16%	NO
$25,000 to $29,999	*0.11%	0.11%	0.00%	0.11%	YES
$30,000 to $34,999	*0.11%	0.11%	0.00%	0.11%	YES
$35,000 to $39,999	*0.07%	0.08%	0.00%	0.07%	YES
$40,000 to $49,999	0.16%	0.13%	0.00%	0.16%	NO
$50,000 to $59,999	0.14%	0.08%	0.02%	0.12%	NO
$60,000 to $69,999	0.26%	0.17%	0.06%	0.20%	NO
$70,000 to $79,999	0.19%	0.15%	0.20%	-0.01%	YES
$80,000 to $89,999	*0.12%	0.13%	0.20%	-0.08%	YES
$90,000 to $99,999	0.07%	0.06%	0.19%	-0.12%	NO
$100,000 to $124,999	0.66%	0.25%	0.87%	-0.20%	YES
$125,000 to $149,999	0.65%	0.28%	1.44%	-0.79%	NO
$150,000 to $174,999	0.80%	0.25%	1.32%	-0.51%	NO
$175,000 to $199,999	1.11%	0.33%	1.39%	-0.28%	YES
$200,000 to $249,999	3.33%	0.58%	3.57%	-0.24%	YES
$250,000 to $299,999	5.54%	0.91%	6.76%	-1.22%	NO
$300,000 to $399,999	13.30%	1.17%	17.78%	-4.48%	NO
$400,000 to $499,999	11.88%	1.04%	15.25%	-3.37%	NO
$500,000 to $749,999	32.64%	1.69%	28.60%	4.03%	NO
$750,000 to $999,999	18.45%	1.25%	13.55%	4.90%	NO
$1,000,000 or more	9.96%	0.92%	8.70%	1.26%	NO

Note: MOE is margin of error. The * indicates that the ACS estimate was not significantly different from zero at the 90% confidence level. Green highlights estimates that fall within 90% ACS margins of error. MRIS is the Multiple Listing Services data for Arlington County, Virginia.
Sources: MRIS 2009-2013 5-year estimates, ACS 2009-2013 5-year estimate.

2. Real Estate Taxes Paid

Administrative assessment data are collected by counties or similar jurisdictions with the purpose of assessing and collecting the necessary real estate taxes. The ACS poses the following question only to respondents who own their housing unit: *“What are the annual real estate taxes on this property?”* Therefore, we expected similar problems for the estimates from the county data given the ambiguity on owner-occupied units. However, this issue seemed much less problematic for taxes paid. Using Arlington County real estate assessments the difference in the number of housing units paying taxes between the ACS and the county real estate assessment data was 3,457 or 9% for the 5-year estimates.

Table 5.8: Comparison of ACS Table B25075 (Value) with WMLS, James City County, 2009-2013

	ACS Benchmark		Direct Estimate from James City County		
Value	Estimate	90%MOE	Estimate	Difference	Within 90% MOE?
Less than $10,000	1.21%	0.23%	0.00%	1.21%	NO
$10,000 to $14,999	0.53%	0.15%	0.00%	0.53%	NO
$15,000 to $19,999	0.63%	0.16%	0.00%	0.63%	NO
$20,000 to $24,999	0.49%	0.15%	0.00%	0.49%	NO
$25,000 to $29,999	0.15%	0.08%	0.00%	0.15%	NO
$30,000 to $34,999	0.36%	0.15%	0.00%	0.36%	NO
$35,000 to $39,999	0.17%	0.10%	0.11%	0.06%	YES
$40,000 to $49,999	0.63%	0.19%	0.09%	0.54%	NO
$50,000 to $59,999	0.39%	0.14%	0.14%	0.25%	NO
$60,000 to $69,999	0.30%	0.12%	0.23%	0.07%	YES
$70,000 to $79,999	0.24%	0.11%	0.18%	0.06%	YES
$80,000 to $89,999	0.14%	0.08%	0.23%	-0.09%	NO
$90,000 to $99,999	0.48%	0.16%	0.16%	0.32%	NO
$100,000 to $124,999	2.23%	0.40%	1.44%	0.79%	NO
$125,000 to $149,999	2.30%	0.38%	3.84%	-1.53%	NO
$150,000 to $174,999	4.82%	0.55%	4.52%	0.30%	YES
$175,000 to $199,999	5.01%	0.59%	5.41%	-0.40%	YES
$200,000 to $249,999	11.50%	0.80%	15.99%	-4.49%	NO
$250,000 to $299,999	12.42%	0.83%	16.56%	-4.14%	NO
$300,000 to $399,999	22.32%	0.94%	25.05%	-2.73%	NO
$400,000 to $499,999	12.65%	0.93%	11.92%	0.73%	YES
$500,000 to $749,999	14.95%	0.79%	11.01%	3.95%	NO
$750,000 to $999,999	4.05%	0.44%	2.12%	1.93%	NO
$1,000,000 or more	2.04%	0.38%	0.98%	1.06%	NO

Note:MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error. The James City County data was extracted from WMLS, the Multiple Listing Services data for from Williamsburg, Virginia.
Sources: WMLS 2009-2013 5-year estimates, ACS 2009-2013 5-year estimates.

The structure of the county real estate assessment data presented challenges. Each tax transaction had an individual observation. Thus, a single tax payment typically had two observations: one where the tax was levied and one where the taxes were paid, adjusted, deferred, or relieved. We transformed these into a single observation of taxes levied and paid, and the estimates were created based on the taxes paid. About 2% of taxes levied did not match taxes paid, as the difference was adjusted, deferred, or relieved.

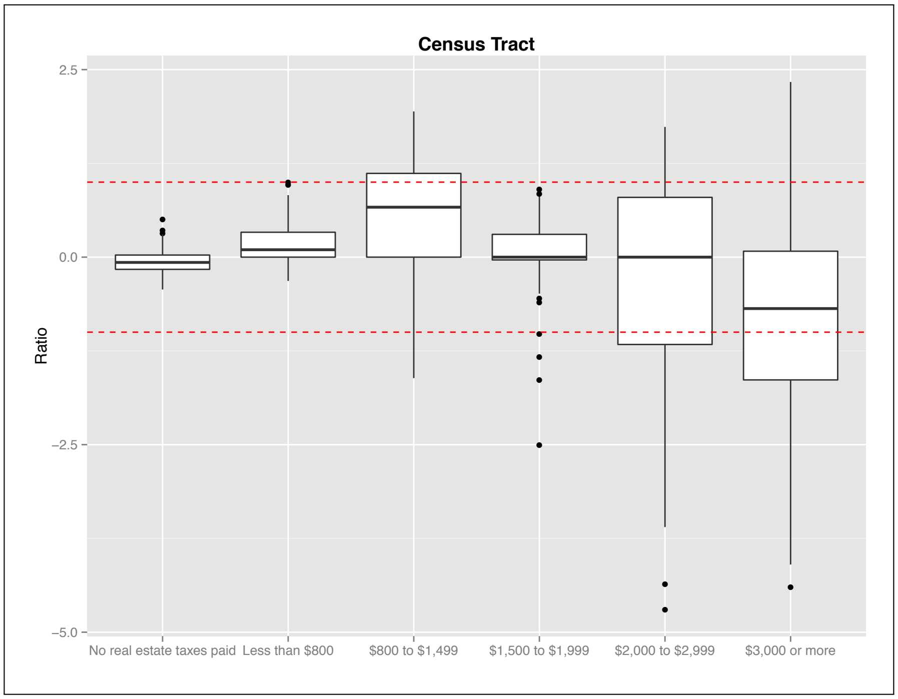
Table 5.9 presents the real estate taxes paid for 2010, comparing the ACS estimates to the county real estate assessment data. The comparison was quite good. Figure 5.11 shows the *Fitness Ratios* for the 5-year estimates across the census tracts for "Real Estate Taxes Paid." The extreme outlier, in the $800 to $1,499 category, with a *Fitness Ratios* of -15 is a census tract in South Arlington and is part of the Douglas Park Community, which is a diverse neighborhood with a mixture of single family homes and condominiums. The overall comparison of the estimates was favorable with the exception that the ACS has lower estimates for the higher taxed properties. This suggests that the local data may be more accurate for this variable.

Table 5.9: Comparison of ACS Table B25102 (Real Estate Taxes Paid) with Arlington County Data, 2010

	ACS Benchmark		Direct Estimate from Arlington County		
Taxes Paid	Estimate	90%MOE	Estimate	Difference	Within 90% MOE?
Less than $800	1,076	775	53	1,023	NO
$800 to $1,499	1,145	793	1,790	-645	YES
$1,500 to $1,999	1,912	1,113	1,777	135	YES
$2,000 to $2,999	4,683	1,463	6,444	-1,761	NO
$3,000 or more	32,113	3,986	33,088	-975	YES
No real estate taxes paid	381	363	655	-274	YES

Note: MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error.
Sources: Arlington County Real Estate Assessments 2010, ACS 2010

Figure 5.11: Distribution of *Fitness Ratio* for Taxes Paid in Arlington County, 2009-2013



Boxplot shows the distribution of the *Fitness Ratios* at the census tract. Estimates falling outside the red reference lines of ±1 were not within the 90% ACS margins of error. One extreme lower outlier (-14.98) was removed from the boxplot. (**Sources:** Arlington County Real Estate Assessments 2009-2013 5-year estimates, ACS 20009-2013 5-year estimates)

G. Summary

The housing case study examined external sources of data for Arlington County and James City County, VA. Acquiring and benchmarking county-level property records opens the opportunity to study features of housing at unprecedented levels of geography and time-frequency. This was demonstrated through the estimation and characterization of housing units using local government property data and data from commercial vendors. Comparing these external sources of data to the ACS identified at least three variables that show promise for use in the ACS. These were tax assessments, sales price, and year built.

6. Representative Housing Use Cases

This chapter explores how the non-federally collected housing data can be used to enhance or even replace the use of American Community Survey (ACS) data in the context of housing research topics. Two specific research use cases were developed. The first was on housing diversity, with a focus on Arlington County, Virginia, and the second was on how housing characteristics relate to housing value applied to both Arlington County and James City County, Virginia.

A. Housing Value Diversity in Arlington County

Diversity can be defined as the inclusion of individuals representing more than one national origin, color, religion, socioeconomic status, sexual orientation, etc. In the context of communities where we live, work, play, and learn, diversity can reflect resilience or other aspects that might make the neighborhood more (or less) desirable to live in, such as housing prices, quality of life, and safety. Policymakers and researchers are interested in developing indicators to identify diversity of neighborhoods. Characterizing spatial diversity of housing within a region provides information for addressing several policy questions such as provision of affordable housing and neighborhood quality.

Diversity measures in the literature are classified broadly as either socio-economic diversity, i.e., racial and ethnic diversity, variation in income, education, age, etc., or diversity in housing stock such as variability in housing and lot size, the age of structures, the mix or single family and multiple family. Weinberg (2011) studied neighborhood income inequality for 2005 through 2009 measured by a Gini index of household income inequality. The author estimated a regression model describing the relationship between the inequality index and census tract characteristics such as the distributions of race, ethnicity, and employment. Narwold and Sandy (2010) explore the roles of housing stock diversity and socioeconomic diversity on housing prices. The authors used the Simpson index of diversity (Simpson 1949) and showed that the value of residential homes increased with higher diversity in the size of homes, but tends to decrease with higher levels of diversity in the age of homes. Other studies that focused on race and ethnicity measure diversity as the share of a metro area's population in its largest racial or ethnic group where the smaller the share of the largest group, the more diverse was the the neighborhood (Humes et al. 2011).

The housing data used in many studies, including the ones mentioned above, are based on ACS data and, thus, neighborhoods were defined as census tracts. For Arlington County, Virgina, the diversity scores for one year can only be calculated in aggregate for the two Public Use Mirodata Areas (PUMA) in the county using ACS Public Use Microdata Sample (PUMS) data. This was because of the lack of geographic detail in the ACS data.

Housing value is an indicator of wealth. Iacovielle (2011) has shown that the total aggregate value of residential real estate (ignoring debt) is about 50% of aggregate net worth. Depending on the question addressed by researchers and/or policy makers, the area of interest could be a city, or something smaller, such as a census tract, census block group, or even as small as a residential block. It is important to characterize distribution of housing values in the area of interest to identify patterns and to conduct analyses that will inform policies targeted at these areas.

Housing value is the metric for diversity used in this analysis. In addition to using the ACS data, we used the Arlington County real estate assessment data. The exact locations (latitude-longitude) of residential homes in the county real estate data provides the opportunity to study the spatial diversity ***within*** a census tract. The county real estate assessment data allow for a full representation of the distribution of housing values within a census tract. They provide a more refined characterization of diversity than ACS alone.

We focused on information the ACS data cannot provide, and addressed the following questions: (i) Where are the high-value (and the low-value) single-family homes in the county? Do these cluster and cross census tract boundaries? Which census tracts show diversity in value? Which census tracts show homogeneity in value? (ii) Which census tracts show diversity in other housing characteristics, (e.g., where are the new units, the large units, the multifamily units)?

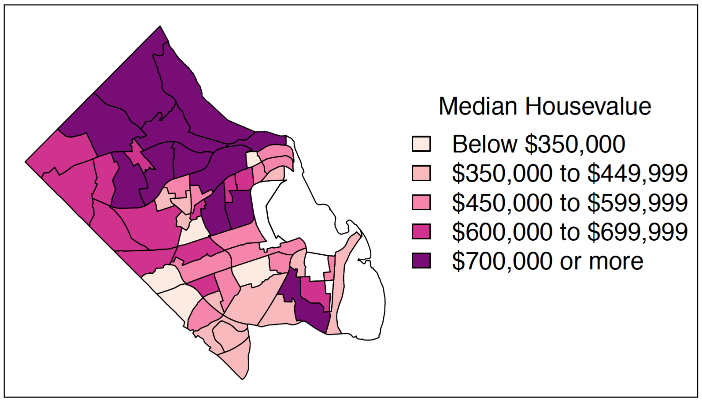
To answer these questions, Simpson indices of diversity were estimated for housing data using value, year built, property type, and number of bedrooms for each census tract for 2013 (following Narwold and Sandy (2010)). The indices associated with house value were the most interesting and are reported here. We also computed Gini indices of house value for single-family housing in each census tract for 2013 (following Weinberg (2011)). For both sets of index calculations 2013 Arlington County real estate assessment data and ACS PUMS data were used.

The Arlington County real estate assessment data included 59,289 housing units with assessed values and geocoded addresses placing them within the Arlington County census tracts. These data do not differentiate between owner-occupied and renter-occupied housing units. For the ACS, PUMS data only owner-occupied housing units within Arlington County were used. Owner-occupied units are single-family, detached units, single-family attached units, or condominiums. There were 498 owner-occupied housing units in the 2013 ACS PUMS. We excluded the following 45 owner-occupied units from the ACS PUMS observations in Arlington County, reducing the estimation sample to 453 housing units:

- Eliminated 2 mobile homes; 1 boat, recreational vehicle, or van; and 2 vacant-for-sale units
- Eliminated 4 units whose value was reported at $1,000 or less and 1 unit whose value was reported as $11,000
- Eliminated 35 units for which the property value was imputed by the Census Bureau

For background, the median housing value in Arlington County for the owner-occupied units is $598,200 (±$22,618) based on the ACS 2013 1-year estimates. Figure 6.1 shows the geographic distribution of the median values at the census tract level using the ACS 2009-2013 5-year estimates. This illustrates that owner-occupied single family housing units with higher values were concentrated in the north of Arlington as compared to the south.

Figure 6.1: Median Value of Housing Units by Census Tract, Arlington County, 2009-2013



North Arlington County has higher home values than South Arlington. White census tracts are areas will too few observations for an ACS estimate and margin of error calculation. (**Source:** ACS 2009-2013 5-Year Estimates.)

a. Simpson Index

To quantify the geographical distribution of the house values within the county, the Simpson index of diversity was calculated (Simpson (1949)). The specific formulation used is

$$1 - \sum_{i=1}^{R} p_i^2,$$

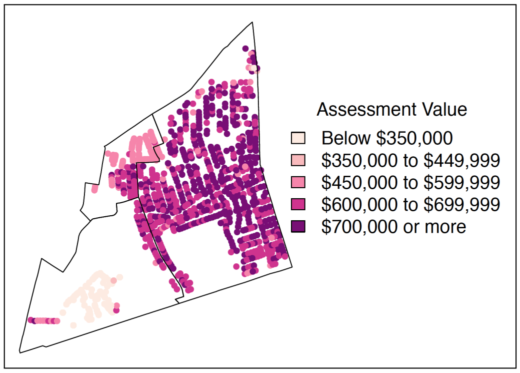
where R is the number of housing value categories. This diversity score equals the probability that two entities taken at random from the dataset of interest are different on the characteristic of interest. This implies that the higher the score, the more diverse is the region.

The ACS PUMS estimates can only be calculated for the 2 PUMAs in county (North and South Arlington). Using ACS PUMS data, the Simpson diversity index for housing values at the county level was 0.78, and for North and South Arlington were 0.75 and 0.79, respectively. These indices do not imply much difference in the diversity of home value of the two areas.

Based on the Arlington County real estate assessment data, the Simpson diversity index for the county was 0.81 and the indices for the North and South Arlington PUMAs were 0.75 and 0.84, respectively. These indices imply that the housing values in the southern part of the county were more diverse as compared to the northern part of the county. This could be a more accurate representation of the diversity in the county than provided by the ACS estimates.

The Arlington County real estate assessment data for 2013 allows for more geographical detail for single-year estimates. As an example, Figure 6.2 shows the assessment values of the housing units in three census tracts in Arlington County. The higher level of granularity of the data allowed observation of the heterogeneity within a census tract as well as the similar values around tract borders. This suggests that the policies, that are related to housing values, targeting census tracts (as the unit of analysis) could be improved with the use of ***external*** data sources.

Figure 6.2: Home Values in Three Arlington County Census Tracts, 2013

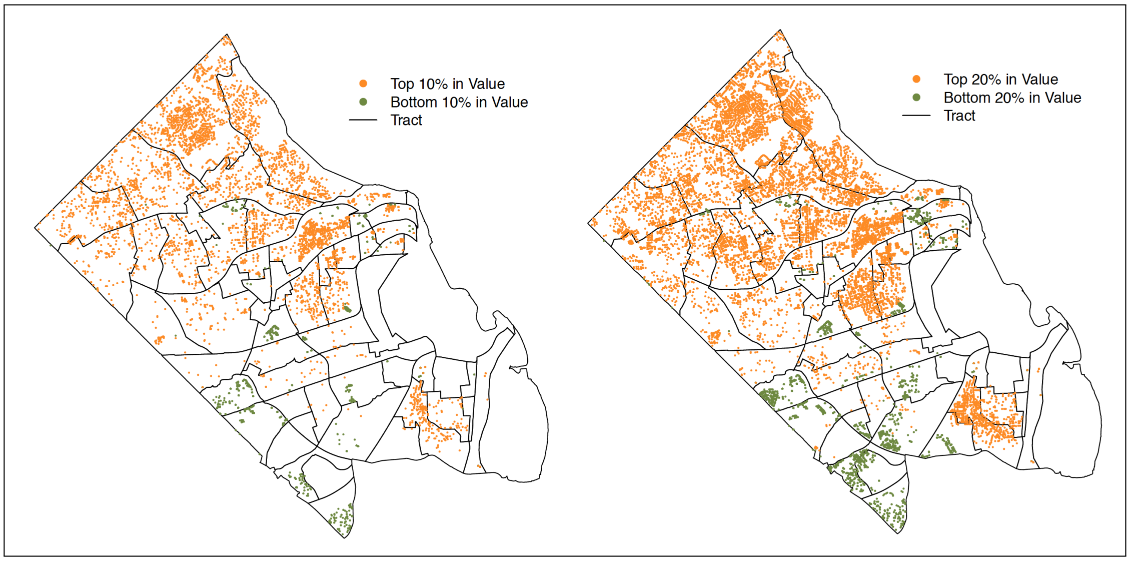


Census tract boundaries are the solid lines. (**Source:** Arlington County Real Estate Assessment Data 2013.)

Figure 6.3 illustrates the geographic locations of the lowest and highest valued units in Arlington County at a high spatial resolution. Such analyses could inform policies targeting the areas with lowest/highest valued housing units.

Arlington county real estate assessment data can be used to characterize diversity at the census tract and block group levels using Simpson indices. Figure 6.4 presents these indices based on house value at both geographic levels. Note the rich spread in the indices across the county, ranging for .03 to .88, as compared to aggregate PUMA values of .75 and .84. Homogeneity, defined as similar diversity levels, was observed in census tracts that were geographically closer. The census tracts located in the midwest part of Arlington appear to have more homogeneous

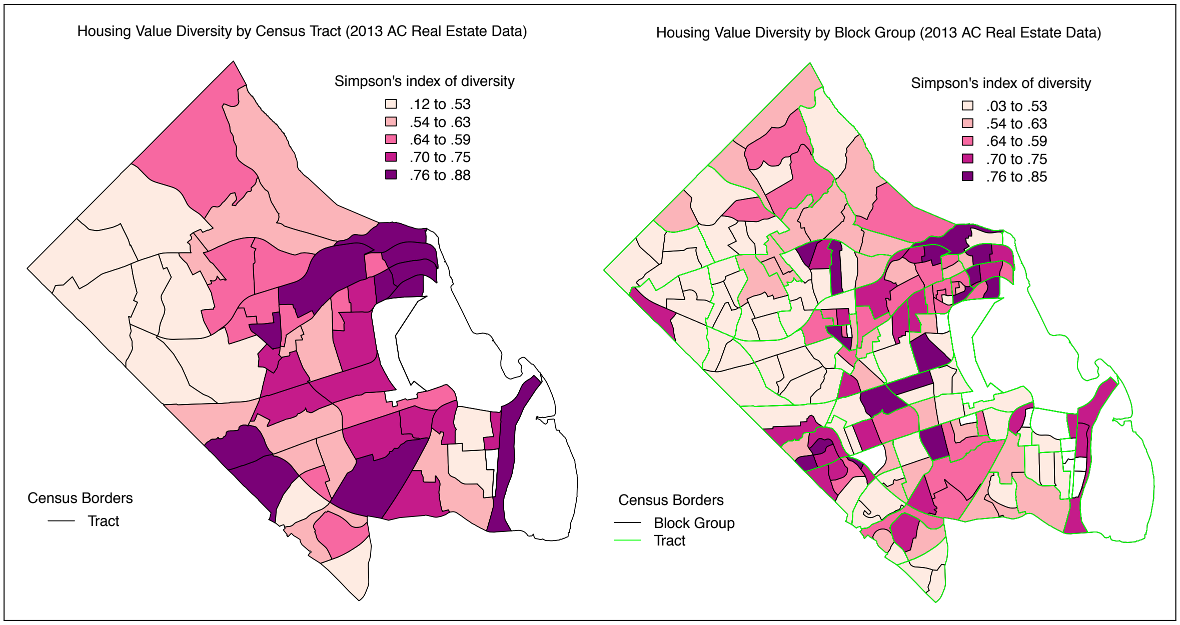
Figure 6.3: Locations of Low and High Value Homes in Arlington County, 2013



Locations of the top and bottom 10% (left) and 20% (right) of assessed values for single family homes in Arlington County in 2013. (**Source:** Arlington County Real Estate Assessment Data 2013.)

levels of housing values compared to the census tracts in the north and south regions. Comparing the census block group distribution on the right to the census tract distribution illustrates that there was heterogeneity within census tracts.

Figure 6.4: Simpson Indices of Housing Value Across Arlington County, 2013


Housing Value Diversity by Census Tract (2013 AC Real Estate Data)
Simpson's index of diversity
.12 to .53
.54 to .63
.64 to .59
.70 to .75
.76 to .88
Census Borders
Tract
Housing Value Diversity by Block Group (2013 AC Real Estate Data)
Simpson's index of diversity
.03 to .53
.54 to .63
.64 to .59
.70 to .75
.76 to .85
Census Borders
Block Group
Tract

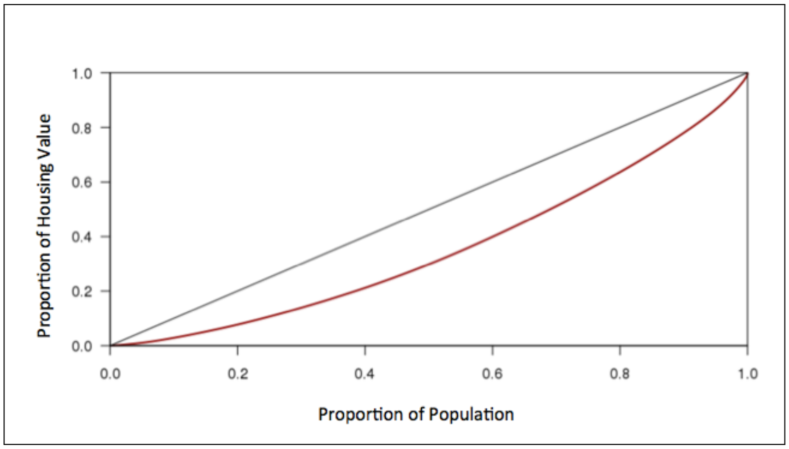
Maps of Simpson indices of housing value by census tract (left) and by block group (right). Simpson indices for census tracts or block groups with fewer than 25 observations were excluded (white regions) (**Source:** Arlington County Real Estate Assessment Data 2013.)

b. Gini Index

The Gini index was computed as an alternative measure to quantify housing value diversity for Arlington County. The Gini index measures the inequality among values of a frequency distribution, (e.g., levels of income). A Gini index of zero expresses perfect equality, where all values are the same such as everyone having the same income. A Gini index of one (or 100%) expresses perfect inequality among values such as one person having all the income and all others have none. The Gini index is mathematically based on the Lorenz curve, which plots the proportion of the total housing value of the population (y axis) that is cumulatively accounted for by the bottom x% of the population. The Gini index can then be thought of as the ratio of the area that lies between the 45-degree line of perfect equality and the Lorenz curve over the total area under the line of equality.

The Lorenz curve for Arlington County based on the 2013 real estate assessment data is shown in Figure 6.5. The figure is similar for the ACS PUMS data. The overall Gini indices of housing value for Arlington County computed individually based on ACS PUMS and on the Arlington County real estate assessment data were identical up to three significant digits, 0.312. Using just the ACS PUMS data, we obtained a Gini index of 0.292 for the North Arlington PUMA and 0.287 for the South Arlington PUMA. Based on Arlington County real estate assessment data, the indices were 0.268 and 0.265, respectively for north and south. We would expect the county-wide Gini index to be larger and to capture more heterogeneity in housing values across the county. For North and South Arlington, the Gini indices do not indicate differences between the two PUMAs.

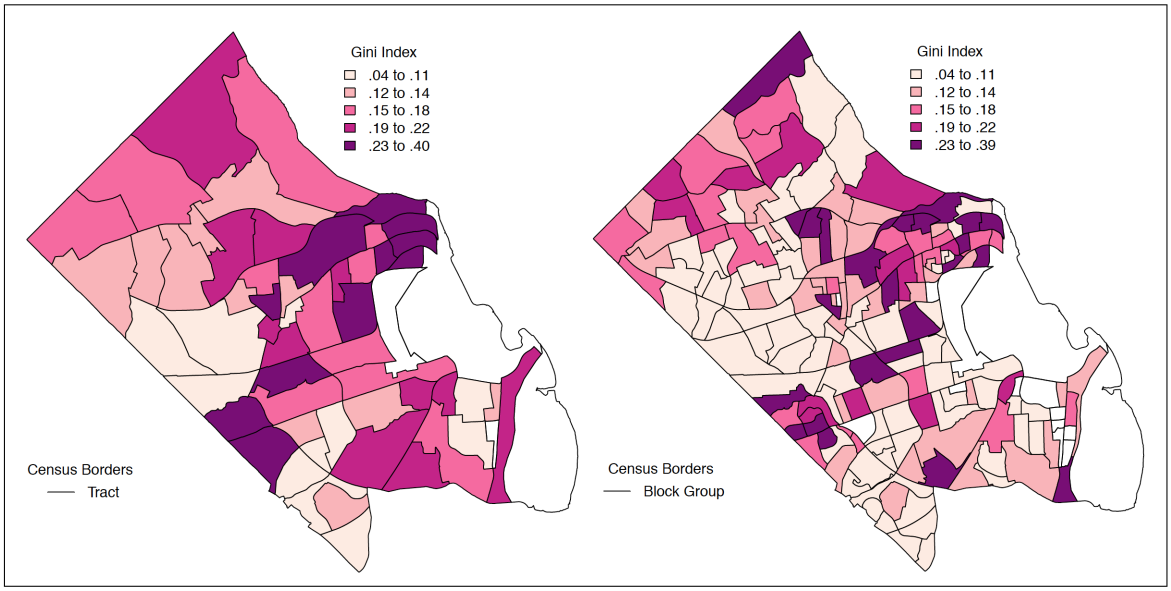
Figure 6.5: Lorenz Curve for Housing Units with Assessed Values in Arlington County, 2013



Lorenz curve calculated for housing units with assessed values in Arlington County. (**Source:** Arlington County Real Estate Assessment Data 2013.)

Using the Arlington County real estate assessment data, Gini indices were also computed at the census tract and block group levels. Figure 6.6 displays the distributions of the housing value Gini indices across census tracts and block groups. No estimates were made if a census tract or block group had fewer than 25 owner-occupied housing units. We observe that the distributions obtained using the Gini indices are similar to the ones based on the Simpson index in Figure 6.4.

Figure 6.6: Gini Indices of Housing Value Across Arlington County, 2013



Gini indices of housing value by census tracts (left) and by block groups (right). Gini indices for census tracts or block groups with fewer than 25 observations were excluded (white regions). (**Source:** Arlington County Real Estate Assessment Data 2013.)

B. Hedonic Regression of House Value

One approach to understanding the dynamics of the housing market involves investigating the determinants of house value. A standard approach has become the use of hedonic indices (Zabel 2015). In the standard hedonic specification, the value of a home, V_{ijt} for house i, in city j, in year t is specified as a function of the characteristics of the house and its' neighborhood:

$$ln(V_{ijt}) = \beta_{0t} + H_{ijt}\beta_{1t} + N_{ijt}\beta_{2jt} + L_{jt}\beta_{3jt} + u_j + e_{ijt}, \tag{6.1}$$

where H_{ijt} is a vector of house characteristics, N_{ijt} is a vector of neighborhood characteristics, L_{jt} is a vector of local public goods, u_j is a geographic fixed effect, and e_{ijt} is the random error term. This specification can also use rent as the dependent variable when studying rental units.

This hedonic housing application is a specific application of a more general formulation, which relates the price or cost of a good to the specific combination of attributes it comprises.

Examples include automobiles, computers, and housing. An often-cited early work on hedonic equations is Court (1939), on automobiles, with more modern applications stemming from Griliches (1961). Malpezzi et al. (2003) and Zabel (2015) also note several applications of hedonic models of house value, including improving house price indices, estimating the value of local amenities such as public goods, (e.g., quality and crime), and job accessibility, measuring environmental quality, and appraising homes for sale.

The hedonic equation is a reduced-form equation that is determined by the interaction of supply and demand (Hill 2013). The most popular functional form is semi-log, that is, the dependent variable is in logarithmic form. Given that the hedonic model is in the form of a reduced-form equation, it is not possible to determine its appropriate functional form by theoretical reasoning alone (Rosen 1974).

1. Arlington County, Virginia

A hedonic regression was first estimated for housing units in Arlington County using the 2013 1% ACS PUMS adjusted sample of owner-occupied housing units, as described in the previous section. For this application, 20 additional units (from the remaining 453) were eliminated for which the value of an independent variable (house characteristic) was imputed by the Census Bureau. As a result, a sample of 433 units was used for the analyses.

Table 6.1 presents the summary statistics for the 2013 ACS PUMS sample included in the model. The median home value in Arlington County is \$600,000 in 2013 and the values range from \$50,000 to over \$2.5 million.

The *modal* owner-occupied unit was in North Arlington on a less than one acre lot, had 3 bedrooms, was heated by utility gas, and was a single-family detached unit built before 1960. Of the other units, 30 percent had 4 or more bedrooms, 36 percent were heated by electricity, 18 percent were in buildings of 50 or more units (condominiums), and one-sixth (17 percent) were built between 1980 and 1999, with another 13 percent built in 2000 or later.

Table 6.2 shows the semi-log regression of house prices on housing characteristics, estimated using weighted least squares. The adjusted R^2 is 0.619. The adjusted R^2 for the same specification and sample, but using an untransformed dependent variable, is 0.486. The residuals depicted in Figure 6.7 (A) do not have a pattern. Many of the coefficients were significantly different from zero and all were in the expected direction. A location in North Arlington increased the housing price, as did having a large lot, more rooms and bedrooms, having the unit be a single-family detached unit, and being built after 1979. The more recent the year built, the higher the price premium.

A hedonic regressions based on the local 2013 Arlington County data sources were estimated using as many of the housing characteristics variables as were profiled and prepared for

Table 6.1: Summary Statistics for Owner-Occupied Housing Units, ACS PUMS, 2013

	Variable	Mean	Minimum	Maximum
Arlington County	Property value	$682,691	$50,000	$2,552,000
	No. of Rooms	6.952	1	19
James City County	Property value	$359,100	$5,000	$2,552,000
	No. of Rooms	7.60	1	19

	Arlington County	James City County
Variable	Frequency	Frequency
Unit has 0 bedrooms	0.007	0.00
Unit has 1 bedroom	0.113	0.00
Unit has 2 bedrooms	0.247	0.10
Unit has 3 bedrooms	0.335	0.42
Unit has 4+ bedrooms	0.298	0.48
Single-family detached	0.550	0.86
Single-family attached	0.166	0.11
Building has 2-19 units	0.079	NA
Building has 20-49 units	0.025	NA
Building has 50+ units	0.180	NA
Building has 2+ units		0.02
Heating fuel: utility gas	0.584	0.55
Heating fuel: electricity	0.363	0.35
Heating fuel: other	0.053	0.09
Built before 1960	0.552	0.09
Built 1960-1979	0.146	0.21
Built 1980-1999	0.171	0.43
Built 2000-2004	0.049	0.13
Built 2005-2009	0.062	0.10
Built 2010-2013	0.021	0.03

Note: Summary statistics of owner-occupied Housing Units in Arlington County PUMAs and the James City County-York County-Williamsburg City-Poquoson City PUMA. NA means the data were not available. Arlington County: Sample size: 433 owner-occupied housing units with no imputed values. Median property value = $600,000, with 16 units topcoded at $2,552,000; property values are rounded. Median number of rooms = 7. James City County: Sample size: 359 owner-occupied housing units with no imputed values. Median property value = $300,000, with 3 units top-coded at $2,552,000; property values are rounded. Median number of rooms = 7. Statistics are unweighted.
Source: ACS Public Use Microdata Sample (PUMS) 2013.

this analysis. The summary statistics for the four data sources for Arlington County examined in this section are listed in Table 6.3. For each data source, condominiums and single family properties (attached and detached) with assessed values and geocoded addresses placing them within the Arlington County census tracts were included. These number differ sightly from the raw property counts on the collaborative wiki (Keller et al. 2016).

Each of the data sources had some unusual data. These data were not eliminated. They are highlighted here for completeness. In the Arlington County real estate assessment data, the home with the highest assessed value was coded by the county as having zero bedrooms, though examination of its detailed property record online indicates it had three bedrooms, but zero on the main level.

Table 6.2: Weighted Hedonic Semi-Log Regression for House Value, Arlington County, American Community Survey Data, 2013

Dependent Variable: Natural Log of (Value / $10,000)	Coefficient	Std. Error	Significance
Intercept	3.7179	0.0869	***
North Arlington	0.2755	0.0412	***
Lot is >1 acre	0.3463	0.1394	**
Unit has 0 bedrooms	-0.3248	0.2339	
Unit has 1 bedrooms	-0.4640	0.0879	***
Unit has 2 bedrooms	-0.1030	0.0676	
Unit has 4+ bedrooms	0.0813	0.0577	
Single-family attached	-0.0752	0.0735	
Building has 2-19 units	-0.4238	0.1034	***
Building has 20-49 units	-0.8750	0.1320	***
Building has 50+ units	-0.2880	0.0944	***
Heating fuel: electricity	-0.0913	0.0489	*
Heating fuel: other	-0.0302	0.0888	
Rooms	0.0439	0.0098	***
Built 1960-1979	0.0056	0.0625	
Built 1980-1999	0.1454	0.0584	**
Built 2000-2004	0.2666	0.0901	***
Built 2005-2009	0.4067	0.0850	***
Built 2010-2013	0.4170	0.1393	***

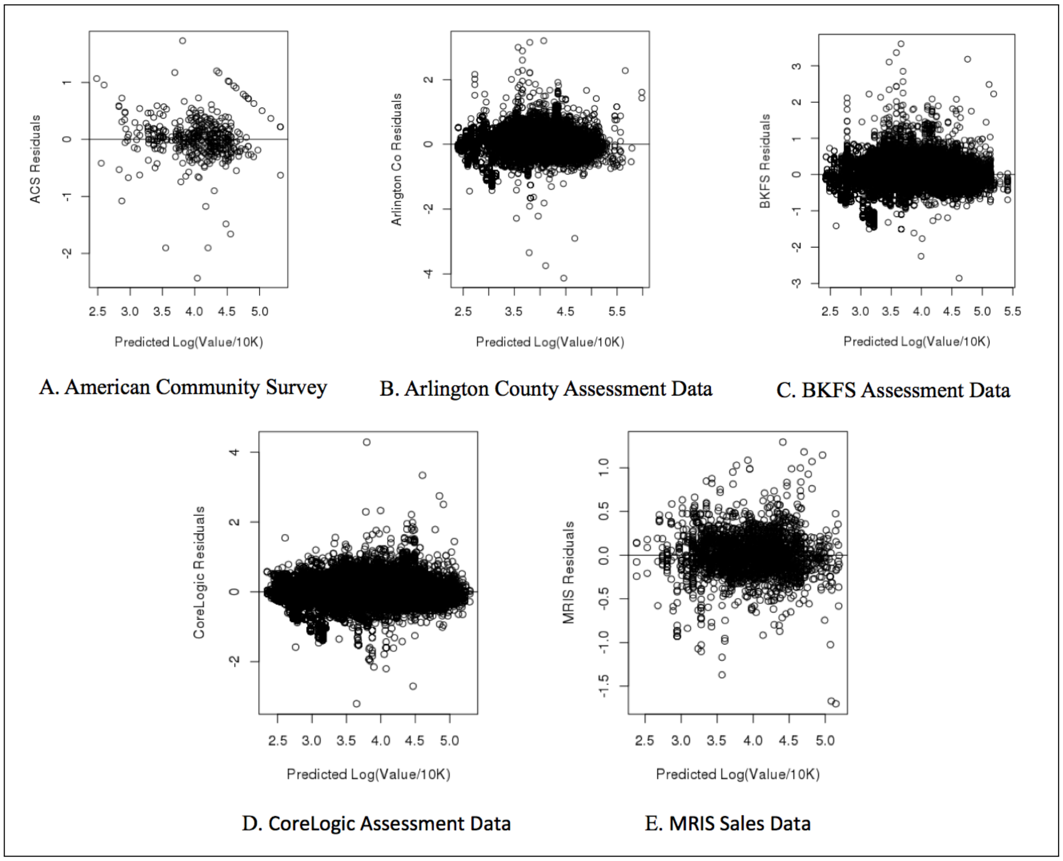
Note: Sample size: 433 owner-occupied housing units with no imputed values. Adjusted R^2 is 0.619. The intercept for the reference unit (single-family detached unit with 3 bedrooms heated with utility gas, and built before 1960 on a lot of less than one acre in South Arlington) is $412,000. */**/*** = Significant at the 0.10/0.05/0.01 level.
Source: American Community Survey 2013.

Table 6.4 presents the results of the ordinary least squares (OLS) hedonic regression using the Arlington County data sources. Even without all the variables measured by the ACS, the hedonic regression using only Arlington County assessment data for single-family housing units or condominiums fit better, with an adjusted R^2 of 0.654, versus the ACS R^2 of 0.619. For comparison, the adjusted R^2 for the ACS without North Arlington variable is 0.599.

Adding dummy variables identifying the 59 census tracts increased the fit to 0.800, and adding ten selected block group characteristics measured for the 181 block groups with housing units over the 2009-2013 period by the ACS increased the fit to 0.821. A description of the characteristics added is in Table 6.5.

Finally data from Location, Inc. census tract indices of school quality, walkability, crime, and quiet score, were added to the model. Since the combinations of the four scores uniquely identified all census tracts, the tract identifiers were removed in the regression. The adjusted R^2 for the hedonic regression including the base variables, the block group characteristics, and the Location, Inc. indices was 0.754, below the fit for the regression without those indices but

Figure 6.7: Residuals from Hedonic Regressions on House Value, Arlington County, 2013.



A. American Community Survey

B. Arlington County Assessment Data

C. BKFS Assessment Data

D. CoreLogic Assessment Data

E. MRIS Sales Data

Residuals are from models that include census tract ID and block group characteristics, except for the ACS model. No particular patterns are apparent.

including the tract identifiers. As was true for the ACS regression, the corresponding residuals in Figure 6.7(B) show no particular pattern.

The OLS hedonic regressions for 2013 BKFS data are described next. There were some differences in variables from the Arlington County assessment records. First, BKFS does not allow properties to have zero bedrooms. Arlington County data had 7,955 units with zero bedrooms. While the BKFS data had 0 parcels with zero bedrooms, they had 7,693 with missing bedroom data. Consequently, a dummy variable was included for missing bedroom information instead of a variable for zero bedroom units. Second, recall that the number of units in the building for the Arlington County assessment data for condominium units was inferred from their GPS locations; this was not available for BKFS, so the number of units variables were replaced by a condominium dummy variable. Third, an additional variable was available for BKFS nd CoreLogic: whether the unit is owner-occupied. BKFS inferred this variable from whether the mailing address was the same as the unit's address, and CoreLogic used a proprietary method.

Table 6.3: Summary Statistics for the *External* data for Single-Family Housing Units, Arlington County, 2013

Variable	Arlington County	BKFS	CoreLogic	MRIS
Mean Property Value	\$529,570	\$529,620	\$528,190	\$598,740
Median Property Value	\$497,500	\$496,650	\$496,950	\$525,000
Number of Units	59,289	59,484	59,742	2,773
Unit is in N. Arlington	0.58	0.58	0.58	0.55
Lot is >1 acre	0.00	0.02	0.13	NA
Unit has 0 bedrooms	0.12	NA	NA	0.01
Unit has 1 bedroom	0.18	0.17	0.18	0.21
Unit has 2 bedrooms	0.28	0.28	0.29	0.32
Unit has 3 bedrooms	0.26	0.26	0.26	0.24
Unit has 4+ bedrooms	0.15	0.15	0.15	0.21
Number of bedrooms unknown	NA	0.13	0.12	NA
Single-family detached	0.46	0.46	0.46	0.37
Single-family attached	0.10	0.10	0.20	0.19
Unit is a duplex	NA	NA	NA	0.02
Building has 2-19 units	0.01	NA	0.11	NA
Building has 20-49 units	0.04	NA	0.02	NA
Building has 50+ units	0.39	NA	0.17	NA
Unit is a condominium	NA	0.44	NA	0.44
Building type unknown	NA	NA	.03	NA
Built before 1960	0.55	0.40	0.55	0.46
Built 1960-1979	0.13	0.06	.013	0.11
Built 1980-1999	0.17	0.07	0.16	0.18
Built 2000-2004	0.04	0.02	0.04	0.05
Built 2005-2009	0.10	0.02	0.10	0.16
Built 2010-2013	0.01	0.01	0.01	0.05
Year built unknown	NA	0.44	0.00	NA
Unit is owner-occupied	NA	0.74	0.79	NA

Notes:This data includes property units identified as condominiums, single family-attached, and single family-detached housing units. Only housing units with assessed values $> \$15,000$ and addresses falling withing the Arlington County census tracts were included. NA means data were not available.
Source: Arlington County Real Estate Assessment Data 2013, Black Knight Financial Services (BKFS) 2013, CoreLogic 2013, Metropolitan Regional Information System (MRIS) Real Sales Data 2013.

Table 6.6 presents the hedonic regression results for Arlington County using the BKFS data. In this case, the adjusted R^2 for the BKFS variables alone (0.605) was slightly less than the fit for ACS (0.619), despite having 59,484 housing units in the BKFS regression versus 433 in the ACS regression. Yet, the addition of the additional sets of variables had the same effect for BKFS as it did for the Arlington County assessment records, adding significantly to the adjusted R^2. Adding the census tract identifiers increased the fit to 0.753, then adding in the block group characteristics increased the fit to 0.794. Replacing the census tract identifiers with the Location Inc. indices did not do as well as including the census tract identifiers (0.732). Figure 6.7 (C) shows that as was true for the ACS and Arlington County real estate assessment regressions there is no pattern for the residuals.

The assessment records provided by CoreLogic for 2013 were analyzed next. As with BKFS, CoreLogic did not allow properties to have zero bedrooms. Arlington County data had

Table 6.4: Hedonic Semi-Log Regressions for House Value, Arlington County Real Estate Assessment Data, 2013

Dependent Variable: Natural Log of (Value / $10,000)	ACS-available variables only	With Census Tract identifiers	With 2009-2013 Block Group characteristics from the ACS	With Block Group characteristics and CT indices from Location Inc.
Intercept	4.2592	3.7376	3.1234	0.7709
North Arlington	0.1644			
Lot is >1 acre	1.0928	0.9261	0.9245	0.9706
Unit has 0 bedrooms	-0.2894	-0.2550	-0.2216	-0.2306
Unit has 1 bedroom	-0.3133	-0.2958	-0.2968	-0.2810
Unit has 2 bedrooms	-0.0370	-0.0336	-0.0438	-0.0403
Unit has 4+ bedrooms	0.1239	0.0753	0.0819	0.1003
Single-family attached	-0.4102	-0.3567	-0.3234	-0.2861
Building has 2-19 units	-0.5123	-0.5893	-0.5467	-0.4057
Building has 20-49 units	-0.6479	-0.8079	-0.7736	-0.6000
Building has 50+ units	-0.7834	-0.8917	-0.8377	-0.7079
Built before 1960	-0.2129	-0.1440	-0.1267	-0.1283
Built 1960-1979	-0.1942	NS -0.0150	NS -0.0183	NS-0.0507
Built 1980-1999	0.1417	0.2169	0.1559	0.1668
Built 2000-2004	0.2566	0.3859	0.3669	0.3172
Built 2005-2009	0.2764	0.4486	0.3973	0.3171
Built 2010-2013	0.3456	0.4898	0.4697	0.4934
Block Group Characteristics				
% married couple households			0.0017	0.0049
% those 25 or older without a high school diploma			0.0031	0.0071
% those 25 or older with bachelor's degree			0.0020	0.0068
% those 3 and older with enrolled in school			-0.0027	-0.0016
% families in poverty			0.0195	0.0168
Median household income / $10,000			-0.0059	-0.0030
% households receiving SNAP benefits			0.0078	0.0062
% 16 and older unemployed			0.0069	0.0043
% housing units vacant			0.0041	0.0032
Median year built (minus 1939)			0.0030	0.0006
Census Tract Characteristics				
LI Quiet score				-0.0011
LI Walk score				-0.0010
LI Crime Index				-0.0004
LI School Quality score				0.0280
Dummy variables for tracts	excluded	included	included	excluded
Adjusted R^2	0.654	0.800	0.821	0.754

Notes: Sample size is 59,289 single-family housing units or condominiums. The reference unit is a single-family detached unit with 3 bedrooms with unknown year built on a lot of less than one acre in South Arlington (census tract 1028.01 when dummy variables for tracts are included). The following ACS variables were not available: heating fuel, number of rooms. SNAP = Supplemental Nutritional Assistance Program. LI = Location, Inc. All coefficients significant at the 0.01 level except where noted (NS = not significant at the 0.10 level; S5 = significant at the 0.05 level).
Source: Arlington County Real Estate Assessment Data, 2013.

7,955 units with zero bedrooms. Perhaps not coincidently, CoreLogic data had 0 parcels with zero bedrooms, and 6,938 with missing bedroom data. Consequently we included a dummy variable for missing bedroom information instead of a variable for zero bedroom units. Second, again as with BKFS, an additional variable was available from CoreLogic namely, whether the unit was owner-occupied.

Table 6.5: Characteristics of Supplemental Data for Arlington County, Census Tracts and Block Groups, 2013

	Median	Mean	Minimum	Maximum
2009-2013 Block Group Characteristics from ACS (n=181)				
% married couple households	84.38	80.82	0.00	100.00
% those 25 or older without a high school diploma	35.03	36.18	6.68	89.85
% those 25 or older with bachelor's degree	59.50	56.74	5.52	92.47
% those 3 and older enrolled in school	21.37	21.56	0.00	58.76
% families in poverty	0.00	5.31	0.00	54.84
Median household income/$10,000	10.94	11.80	0.99	25.00
% households receiving SNAP benefits	0.00	3.42	0.00	64.03
% 16 and older unemployed	2.88	4.19	0.00	28.30
% housing units vacant	7.74	12.25	0.00	100.00
Median year built (minus 1939)	19.00	23.97	0.00	66.00
"2013" Census Tract Characteristics from Location, Inc. (n=59)				
Quiet score	33.00	35.54	3.00	100.00
Walk score	83.00	78.93	0.00	92.00
Crime Index	55.00	59.14	15.00	160.00
School Quality score	83.28	81.54	68.81	83.86

Notes: Substitutes year built as measured for census tract 1802.02 for missing data for block group 2 in that census tract. Medians and means are not weighted for the number of housing units or parcels.
Source: American Community Survey (ACS) and Location, Inc.

The descriptive statistics for CoreLogic showed an anomaly for the lot size variable. Arlington County data show that 0.2% of units (94 units) are on at least one acre, the CoreLogic data showed that 13% were on large lots (BKFS data showed 2% were on large lots; see Table 6.3). This appeared to be an error; for example, on the CoreLogic data file there were 303 units with exactly 16.1141 acres, 333 units with exactly 28.023 acres, and 755 units with exactly 43.0738 acres. This variable was used in the model but warrants more exploration.

Table 6.7 presents the hedonic regression results for Arlington County using the CoreLogic data. In this case, the adjusted R^2 for the CoreLogic variables alone for 59,752 housing units (0.644) was higher than the fit for ACS (0.619). The addition of the additional sets of variables had the same effect for CoreLogic as it did for the Arlington County assessment records and for BKFS, adding significantly to the adjusted R^2. Adding the census tract identifiers increased the fit to 0.776, then adding in the block group characteristics increased the fit to 0.797, and replacing the census tract identifiers with the Location Inc. indices did not do as well as including the census tract identifiers (0.732). Figure 6.7 (D) shows that as was true for ACS and Arlington County real estate assessment regressions there is no pattern for the residuals.

The next set of analysis used a data source generated by home sales through the Multiple Listing Service for Arlington County in 2013, as assembled by the Metropolitan Regional Information System (MRIS) sales data. The MRIS closing sales price is the actual value of the housing unit on the date the sale closed. This may differ from the county's assessed value and the respondent's estimate of home value on the ACS. Even though the MLS data collected

Table 6.6: Hedonic Semi-Log Regressions for House Value for Arlington County, Black Knight Financial Services Assessment Records 2013

Dependent Variable: Natural Log of (Value / $10,000)	ACS-available variables only	With Census Tract identifiers	With 2009-2013 Block Group characteristics from the ACS	With Block Group characteristics and CT indices from Location Inc.
Intercept	3.8299	3.4491	2.7007	0.1353
North Arlington	0.2432			
Lot is >1 acre	0.1163	0.0587	0.0578	0.0269
Unit has 1 bedroom	-0.2423	-0.2787	-0.3117	-0.2728
Unit has 2 bedrooms	S5 0.0099	-0.0156	-0.0416	-0.0290
Unit has 4+ bedrooms	0.0850	0.0663	0.0687	0.0708
Unknown number of bedrooms	-0.3195	-0.2600	-0.1962	-0.2263
Single-family attached	-0.4137	-0.3320	-0.3313	-0.3437
Unit is condominium	-0.4937	-0.5810	-0.5044	-0.4218
Built before 1960	NS 0.0628	NS 0.0345	S5 0.1062	S5 0.1133
Built 1960-1979	0.1773	0.1240	0.1713	0.1944
Built 1980-1999	0.4425	0.3392	0.3674	0.4424
Built 2000-2004	0.5909	0.4825	0.5281	0.5834
Built 2005-2009	0.6236	0.5951	0.6510	0.6863
Built 2010-2013	0.5846	0.5780	0.6499	0.6915
Unit is owner-occupied	0.1104	0.1019	0.0755	0.0645
Block Group Characteristics				
% married couple households			0.0014	0.0047
% those 25 or older without a high school diploma			0.0035	0.0069
% those 25 or older with bachelor's degree			0.0016	0.0060
% those 3 and older enrolled in school			-0.0040	-0.0022
% families in poverty			-0.0064	-0.0031
Median household income / $10,000			0.0262	0.0233
% households receiving SNAP benefits			0.0091	0.0073
% 16 and older unemployed			0.0088	0.0050
% housing units vacant			0.0043	0.0040
Median year built (minus 1939)			0.0069	0.0051
Census Tract Characteristics				
LI Quiet score				-0.0012
LI Walk score				0.0015
LI Crime Index				-0.0008
LI School Quality score				0.0288
Dummy variables for tracts	excluded	included	included	excluded
Adjusted R^2	0.605	0.753	0.794	0.732

Notes: Sample size is 59,484 single-family housing units or condominiums (first column); other columns exclude 64 cases with no tract identifier. The reference unit is a single-family detached unit with 3 bedrooms with unknown year built on a lot of less than one acre in South Arlington (tract 1028.01 when dummy variables for tracts are included). The following ACS variables are not available: heating fuel, number of rooms. SNAP = Supplemental Nutritional Assistance Program. LI = Location, Inc. All coefficients significant at the 0.01 level except where noted (NS = not significant at the 0.10 level; S5 = significant at the 0.05 level). Black Knight Financial Services does not code zero bedroom units.
Source: Black Knight Financial Services Assessment Record 2013.

for the James City County-Williamsburg area included lot size, the MRIS data did not, and therefore that variable was excluded from the regressions. It was not possible to determine the number of units in a building, except if it was a duplex, since the MRIS data only indicate the number of floors in a building and not the number of units. We designated any units not identified as single-family detached or attached as duplexes or multifamily units.

Table 6.7: Hedonic Semi-Log Regressions for House Value for Arlington County, CoreLogic Assessment Records, 2013

Dependent Variable: Natural Log of (Value / $10,000)	ACS-available variables only	With Census Tract identifiers	With 2009-2013 Block Group characteristics from the ACS	With Block Group characteristics and CT indices from Location Inc.
Intercept	4.0718	3.4906	2.9373	0.5814
North Arlington	0.1812			
Lot is >1 acre	-0.2315	-0.1853	-0.1172	-0.1058
Unit has 1 bedroom	-0.4197	-0.4071	-0.3959	-0.3893
Unit has 2 bedrooms	-0.1058	-0.1098	-0.1119	-0.1107
Unit has 4+ bedrooms	0.1305	0.0765	0.0821	0.1017
Unknown number of bedrooms	-0.2853	-0.2656	-0.2340	-0.2340
Single-family attached	-0.5356	-0.5202	-0.4624	-0.4357
Building has 2-19 units	-0.5707	-0.6287	-0.6165	-0.5271
Building has 20-49 units	-0.4890	-0.5756	-0.5366	-0.3904
Building has 50+ units	-0.5716	-0.5833	-0.5535	-0.4921
Number of units unknown	-0.5619	-0.4034	-0.3811	-0.3785
Built before 1960	-0.0983	NS -0.0397	NS-0.0208	S10 -0.0459
Built 1960-1979	-0.1137	NS 0.0284	NS 0.0347	NS -0.0017
Built 1980-1999	0.2341	0.3123	0.2616	0.2659
Built 2000-2004	0.2829	0.4360	0.4414	0.3595
Built 2005-2009	0.2931	0.4561	0.4261	0.3378
Built 2010-2013	0.4458	0.5845	0.5778	0.5910
Unit is owner-occupied	0.0916	0.0720	0.0586	0.0532
Block Group Characteristics				
% married couple households			0.0029	0.0051
% those 25 or older without a high school diploma			0.0020	0.0050
% those 25 or older with bachelor's degree			NS 0.0002	0.0032
% those 3 and older enrolled in school			-0.0026	-0.0018
% families in poverty			-0.0052	-0.0023
Median household income / $10,000			0.0217	0.0231
% households receiving SNAP benefits			0.0092	0.0078
% 16 and older unemployed			0.0056	0.0035
% housing units vacant			0.0040	0.0028
Median year built (minus 1939)			0.0024	0.0004
Census Tract Characteristics				
LI Quiet score				-0.0007
LI Walk score				0.0012
LI Crime Index				-0.0006
LI School Quality score				0.0289
Dummy variables for tracts	excluded	included	included	excluded
Adjusted R^2	0.644	0.776	0.797	0.732

Notes: Sample size is 59,752 single-family housing units or condominiums. The reference unit is a single-family detached unit with 3 bedrooms with unknown year built on a lot of less than one acre in South Arlington (tract 1028.01 when dummy variables for tracts are included). SNAP = Supplemental Nutritional Assistance Program. LI = Location, Inc. All coefficients significant at the 0.01 level except where noted (NS = not significant at the 0.10 level; S10 = significant at the 0.10 level).
Source: CoreLogic Assessment Records, 2013.

Table 6.8 presents the hedonic regressions for the for the 2,773 single-family housing units in the MRIS data for Arlington County. These data provided a better fit for the first specification, i.e., adjusted R^2 of 0.684, as compared to 0.619 for ACS and 0.654 for Arlington County real estate assessment data, presented in earlier tables. This fit increased to 0.793 with census tract identifiers and 0.802 with census tract identifiers and block group characteristics (both slightly lower than for Arlington County data). The familiar pattern emerged similar to the other models when including Location Inc. indices. The residuals also performed similar to the other models (Figure 6.7 (E)).

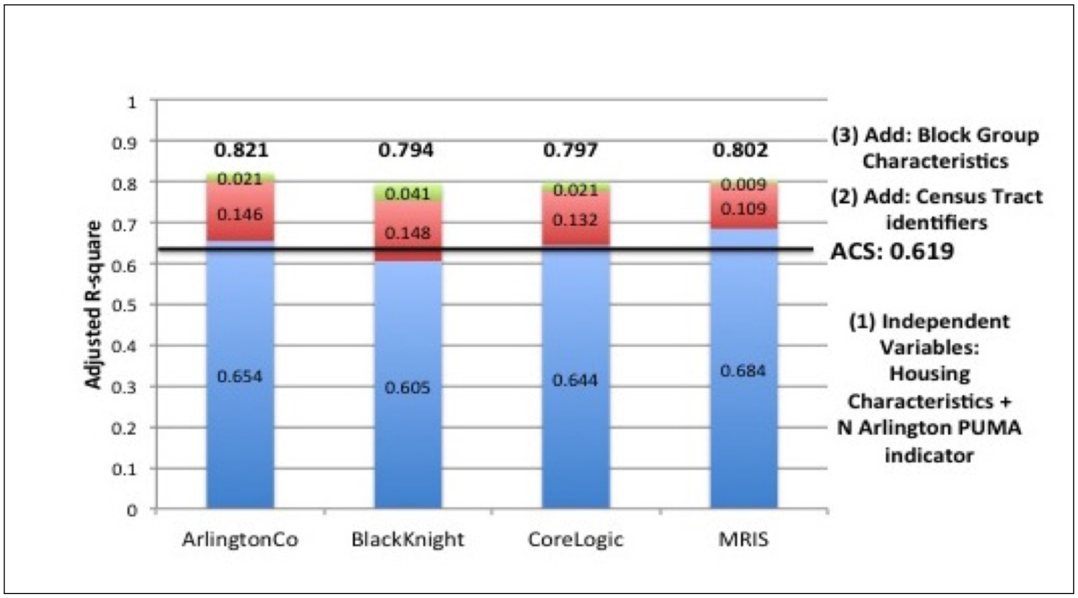
Table 6.8: Hedonic Semi-Log Regressions for House Value for Arlington County, Metropolitan Regional Information System Real Estate Sales Data, 2013

Dependent Variable:	ACS-available	With	With 2009-2013	With Block Group
Natural Log of (Value / $10,000)	variables only	Census	Block Group	characteristics
		Tract	characteristics	and CT indices
		identifiers	from the ACS	from Location Inc.
Intercept	4.0803	3.7255	3.4525	0.6363
North Arlington	0.1927			
Unit has 0 bedrooms	-0.9523	-0.9996	-0.9260	-0.8872
Unit has 1 bedroom	-0.5662	-0.5669	-0.5653	-0.5642
Unit has 2 bedrooms	-0.2337	-0.2309	-0.2321	-0.2397
Unit has 4+ bedrooms	0.1829	0.1191	0.1184	0.1500
Single-family attached	-0.1687	-0.2538	-0.2310	-0.1648
Unit is duplex	-0.2203	NS -0.0381	NS -0.0506	-0.1160
Unit is multifamily	-0.4113	-0.4918	-0.4609	-0.3896
Built 1960-1979	NS -0.0206	0.0663	0.0582	NS 0.0295
Built 1980-1999	0.1739	0.2161	0.1920	0.1844
Built 2000-2004	0.3926	0.4607	0.4356	0.3780
Built 2005-2009	0.4495	0.5564	0.5306	0.4185
Built 2010-2013	0.5122	0.5965	0.5801	0.5938
Block Group Characteristics				
% married couple households			S10 0.0012	0.0033
% those 25 or older without a high school diploma			NS 0.0018	0.0086
% those 25 or older with bachelor's degree			NS -0.0005	0.0060
% those 3 and older enrolled in school			S10 -0.0015	S10 -0.0014
% families in poverty			-0.0045	NS 0.0000
Median household income / $10,000			0.0161	0.0183
% households receiving SNAP benefits			NS 0.0028	NS 0.0041
% 16 and older unemployed			0.0046	0.0036
% housing units vacant			S10 0.0013	0.0050
Median year built (minus 1939)			NS 0.0004	NS 0.0005
Census Tract Characteristics				
LI Quiet score				NS 0.0003
LI Walk score				NS 0.0002
LI Crime Index				0.0011
LI School Quality score				0.0274
Dummy variables for tracts	excluded	included	included	excluded
Adjusted R^2	0.684	0.793	0.802	0.734

Notes: Sample size is 2,773 single-family housing units or condominiums. The reference unit is a single-family detached unit with 3 bedrooms with year built before 1960 in South Arlington (tract 1028.01 when dummy variables for tracts are included). The following ACS variables are not available: heating fuel, number of rooms, large lot. SNAP = Supplemental Nutritional Assistance Program. LI = Location, Inc. All coefficients significant at the 0.01 level except where noted (NS = not significant at the 0.10 level; S5 = significant at the 0.05 level; S10 = significant at the 0.10 level).
Source: Metropolitan Regional Information System Sales Data 2013.

The overall pattern of results is presented by Figure 6.8, showing the incremental value of adding additional variables to the hedonic regressions beyond those available from the ACS. Clearly, ***external*** data can usefully augment the ACS data for characterizing house value. In

Figure 6.8: Comparison of Goodness-of-Fit Statistics for Hedonic Regressions, Arlington County, 2013



The colors on the bars show the incremental improvements in the adusted R^2 for the addition of finer levels of geography. The North Arlington PUMA indicator is removed when census tract identifiers are added.

particular, census tract identifiers and block group characteristics were useful, even if the latter were averages over the 5 years culminating in the year to which the administrative data pertain. Because it was necessary to know the unit's census tract to take advantage of the Location Inc. data, substituting the four indices for the 59 dummy tract indicator variables did not add significantly to the explanatory power.

2. James City County, Virginia

James City County is a mostly rural county near Williamsburg, the colonial capital of Virginia. The hedonic analysis performed for Arlington County is repeated in this new locale. Table 6.1 presents the summary statistics for the single-family units in the county.

The first step was to estimate a hedonic model for James City County using 2013 ACS microdata. However, James City County has too small a population to be separately identified on the ACS PUMS. Therefore the analysis was done using Public Use Microdata Area (PUMA) that also includes York County, and the independent cities of Williamsburg and Poquoson. Table 6.9 presents the hedonic regression for the James City County PUMA. Its adjusted R^2 was lower than that for Arlington County (an urban location), 0.422 versus 0.619.

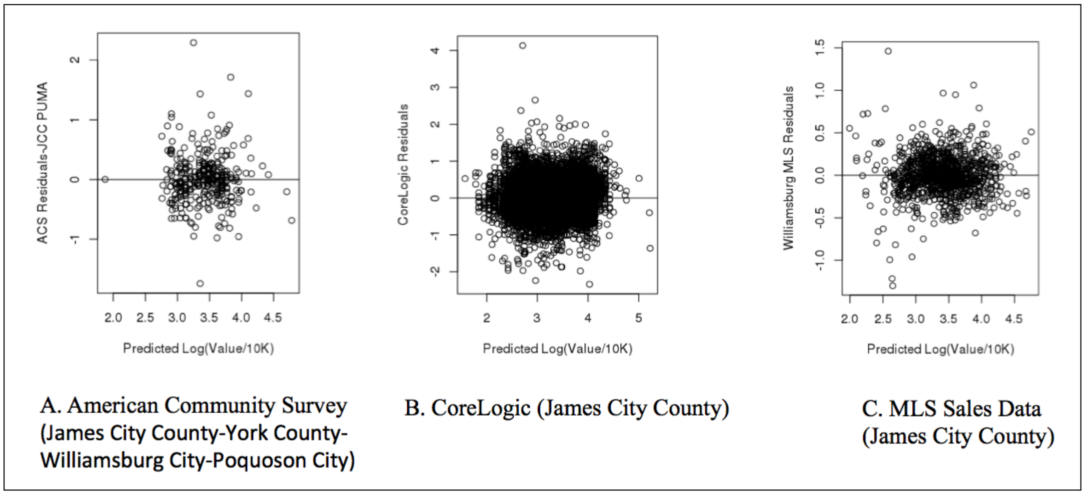
Also, there were only 8 single-family housing units in buildings of two or more units in James City County, preventing that independent variable from explaining much of the variation. The residuals in Figure 6.9 show no patterns.

Table 6.9: Hedonic Semi-Log Regression for House Value, James City County-York County-Williamsburg City-Poquoson City, American Community Survey data, 2013

Dependent Variable: Natural Log of (Value / $10,000)	Coefficient	Std. Error	Significance
Intercept	2.5946	0.1085	***
Lot is >1 acre	0.2605	0.0694	***
Unit has 1 bedrooms	-0.8302	0.4164	**
Unit has 2 bedrooms	0.0583	0.0921	
Unit has 4+ bedrooms	0.1225	0.0513	**
Rooms	0.0753	0.0115	***
Single-family attached	-0.2195	0.0753	***
Building has 2+ units	-0.2754	0.1699	
Heating fuel: electricity	-0.1407	0.0465	***
Heating fuel: other	-0.0933	0.0917	
Built 1960-1979	0.0874	0.0882	
Built 1980-1999	0.2975	0.0843	***
Built 2000-2004	0.3704	0.0974	***
Built 2005-2009	0.3732	0.1006	***
Built 2010-2013	0.4123	0.1389	***

Note: Sample size: 359 owner-occupied housing units with no imputed values (of 713 observations). Adjusted $R^2 = 0.422$. The intercept for the reference unit (single-family detached unit with 3 bedrooms heated with utility gas and built before 1960 on a lot of less than one acre) is $134,000. */**/*** = Significant at the 0.10/0.05/0.01 level.
Source: American Community Survey (ACS) Public Use Microdata Sample 2013.

Figure 6.9: Residuals from Hedonic Regression with Census Tract Identifiers and Block Group Characteristics, James City County, 2013



A. American Community Survey (James City County-York County-Williamsburg City-Poquoson City)

B. CoreLogic (James City County)

C. MLS Sales Data (James City County)

Residuals are from models that include census tract ID and block group characteristics, except for the ACS model. Residuals do not show any apparent patterns.

James City County does not archive its historical data so only 2015 data could be accessed. Consequently, a hedonic regression was not estimated for James City County's real estate assessment data.

BKFS does not distinguish among single-family detached, single-family attached, and condominium units for James City County. Therefore, a hedonic analysis using BKFS data was not estimated.

We were able to estimate the hedonic regression for the county using CoreLogic data. The results are given in Table 6.10. There were 11 census tracts in James City County, and 27 block groups. The goodness-of-fit for James City County alone using CoreLogic data and ACS variables was 0.349, lower than the adjusted R^2 for the ACS PUMA that includes James City County (that is, for a non-comparable geographic area.) But, the pattern that census tract identifiers and block group characteristics can add significantly to the explanatory power of independent variables holds for James City County, as was found for Arlington County. The adjusted R^2 rose to 0.537 when adding only census tract identifiers, and to 0.574 when block group characteristics were added. Figure 6.9 shows the residuals from the regression that includes these other variables. Given the results for the Location Inc. supplemental data for Arlington County, the regression approach that replaced the tract identifiers with the four Location Inc. indices was not repeated.

Finally, we used the data for 2013 from the Williamsburg Multiple Listing Service real estate sales, which covers Williamsburg and James City County. After eliminating the two home sales in Williamsburg, Table 6.11 presents the results of the hedonic regression for James City County. The fit of the regression without any geographical information was better than that for the ACS microdata covering a wider geographic area with an adjusted R^2 of 0.640 versus 0.422. Showing the same pattern as in Arlington County, the census tract identifiers alone added a substantial amount to the explanatory power increasing the adjusted R^2 to 0.733, and the addition of block group characteristics increased the adjusted R^2 yet again, to 0.763. Again, given the results for the Location, Inc. supplemental data for Arlington County, we did not repeat the regression approach that replaced the tract identifiers with the four Location Inc. indices.

Though the goodness-of-fit is initially lower for James City County data than for Arlington County data, the pattern of results is quite similar, and the fit including the location-based variables is comparable across the two locales.

Table 6.10: Hedonic Semi-Log Regressions for House Value, James City County, CoreLogic Assessment Data, 2013

Dependent Variable: Natural Log of (Value / $10,000)	ACS-available variables only	With Census Tract identifiers	With 2009-2013 Block Group characteristics from the ACS
Intercept	2.7039	3.1536	2.0715
Lot is >1 acre	0.2282	0.3143	0.2931
Unit has 1 bedroom	-0.2211	-0.2636	-0.2803
Unit has 2 bedrooms	-0.1814	-0.2037	-0.2207
Unit has 4+ bedrooms	0.3623	0.2448	0.2415
Unknown number of bedrooms	-0.0288	-0.0498	-0.0273
Single-family attached	-0.2664	-0.2706	-0.2841
Building has 2-19 units	-0.4352	-0.3483	-0.3784
Number of units unknown	1.9482	1.9432	1.8567
Built before 1960	-0.1801	-0.2020	-0.0704
Built 1960-1979	0.2059	0.1830	S5 0.2047
Built 1980-1999	0.4737	0.4069	0.4165
Built 2000-2004	0.5941	0.5427	0.5406
Built 2005-2009	0.5986	0.6351	0.6188
Built 2010-2013	0.4973	0.5740	0.5752
Unit is owner-occupied	0.1093	0.1034	0.1065
Block Group Characteristics			
% married couple households			0.0053
% those 25 or older without a high school diploma			-0.0059
% those 25 or older with bachelor's degree			-0.0015
% those 3 and older enrolled in school			-0.0065
% families in poverty			0.0060
Median household income / $10,000			0.0331
% households receiving SNAP benefits			0.0062
% 16 and older unemployed			S10 0.0019
% housing units vacant			NS 0.0003
Median year built (minus 1939)% housing units vacant			0.0107
Dummy variables for tracts	excluded	included	included
Adjusted R^2	0.349	0.537	0.574

Notes: Sample size is 25,977 single-family housing units or condominiums. The reference unit is a single-family detached unit with 3 bedrooms with unknown year built on a lot of less than one acre (tract 801.01 when dummy variables for tracts are included). SNAP = Supplemental Nutritional Assistance Program. All coefficients significant at the 0.01 level except where noted (NS = not significant at the 0.10 level; S10 = significant at the 0.10 level).
Source: CoreLogic Assessment Records 2013.

Table 6.11: Hedonic Semi-Log Regressions for House Value, James City County, Multiple Listing Service Real Estate Sales Data, 2013

Dependent Variable: Natural Log of (Value / $10,000)	ACS-available variables only	With Census Tract identifiers	With 2009-2013 Block Group characteristics from the ACS
Intercept	2.2431	2.6026	1.9324
Lot is >1 acre	S5 0.0936	0.2483	0.2218
Unit has 1 bedroom	-0.6734	-0.6152	-0.6110
Unit has 2 bedrooms	-0.1209	-0.1152	-0.1334
Unit has 4+ bedrooms	0.5048	0.4053	0.3702
Unit has 4+ bathrooms	-0.1728	-0.1584	-0.1429
Rooms	0.0663	0.0526	0.0493
Unknown number of rooms	0.5152	0.4046	0.3945
Single-family attached	-0.2831	-0.3192	-0.3766
Condominium	0.3483	0.3798	0.4027
Built 1960-1979	NS 0.1009	S5 0.1408	S5 0.1200
Built 1980-1999	0.2987	0.3118	0.2825
Built 2000-2004	0.3358	0.4328	0.3840
Built 2005-2009	0.3742	0.5017	0.4368
Built 2010-2013	0.3516	0.5001	0.4719
Block Group Characteristics			
% married couple households			S10 0.0037
% those 25 or older without a high school diploma			NS -0.0059
% those 25 or older with bachelor's degree			NS-0.0028
% those 3 and older enrolled in school			-0.0052
% families in poverty			NS -0.0016
Median household income / $10,000			NS 0.0158
% households receiving SNAP benefits			NS -0.0036
% 16 and older unemployed			NS 0.0042
% housing units vacant			S5 0.0019
Median year built (minus 1939)			0.0099
Dummy variables for tracts	excluded	included	included
Adjusted R^2	0.640	0.733	0.763

Note: Sample size is 1,146 single-family housing units or condominiums. The reference unit is a single-family detached unit with 3 bedrooms with year built before 1960 (tract 801.01 when dummy variables for tracts are included). SNAP = Supplemental Nutritional Assistance Program. All coefficients significant at the 0.01 level except where noted (NS = not significant at the 0.10 level; S5 = significant at the 0.05 level; S10 = significant at the 0.10 level).

Source: Tabulation of 2013 Multiple Listing Service Sales Data.

C. Summary

The housing use cases (representative research studies) focused on measuring the diversity of housing value as a surrogate for wealth and the effect of housing characteristics on housing value, measured with hedonic regressions. The use cases illustrate the potential benefits of the local external data sources to characterize diversity at a lower level of granularity, to identify determinants of housing values, and for characterizing the spatial distribution of other variables, such as year built, number of bedrooms, and heating type.

7. Education Data Framework

The education case study focused on external data obtained from state administrative longitudinal educational records that provide information about student enrollment by grade, school, demographic characteristics, and other attributes. The study sought to use administrative educational data that, when combined with a subset of the American Community Survey (ACS) education variables, and analyzed through valid statistical procedures, provide useful insight into the underlying factors and processes affecting education outcomes.

Most state longitudinal administrative records programs grew out of the Statewide Longitudinal Data Systems Grant Program. This program is administered by the Institute of Education Sciences (IES) of the U.S. Department of Education and was first authorized in 2002. The program provides funding for state education agencies in the 50 states, District of Columbia, and U.S. territories to develop and maintain longitudinal data systems from administrative data collected during each school year.

The first round of awards was made in November of 2005. Since then four more rounds of grants were awarded, one in 2007, two in 2009, and one in 2012. A total of $613 million have been awarded to 47 states, the District of Columbia, Puerto Rico, and the Virgin Islands. The three states that have not received any SLDS funding are Wyoming, Alabama, and New Mexico, although these states have developed SLDS systems.

Early rounds of funding focused on developing K-12 systems to track students through a unique student identifier from kindergarten to 12th grade. Later rounds focused on developing P-20W SLDS systems. P-20W refers to data from pre-kindergarten (early childhood), K-12, and postsecondary through post-graduate education, along with workforce and other outcomes data (e.g., public assistance and corrections data). The specific agencies and other organizations that participate in the P-20W initiative vary from state to state.

Most states track students as long as they remain in the state. Washington is an exception in that they track higher education students who attend school out of state.

Each state is in a different stage of development in terms of what data they provide for research and the processes researchers must go through to obtain this data. This chapter discusses the insights and understanding of these differences gained through the data discovery, inventory, acquisition, profiling, preparation, linking and exploration of the the SLDS systems.

A. Data Discovery, Inventory, and Selection

1. Data Discovery

The data discovery in this study began by applying the screening process, as described in Chapter 3, to the Statewide Longitudinal Data Systems (SLDS) data. An inventory was completed for SLDS data sources for the 50 states and the District of Columbia. Many states had separate data sources for K-12, higher education, and workforce data.

For completeness other commercial, state, U.S. federal, and international education sources of data were also identified and screened. These additional sources are listed in Table 7.1. The collaborative wiki (Keller et al. 2016) documents the data discovery and screening for all of the education data sources.

2. Data Inventory

Table 7.1: Federal, International, and Commercial Data Inventory for Education Case Study

Federal	**International**
Bureau of Labor Statistics	Organization for Economic Co-operation and Development
Business Employment Dynamics	Program for International Student Assessment (PISA)
Current Population Survey	Teaching and Learning International Survey (TALIS)
Employment Projections	
Job Openings and Labor Turnover Survey (JOLTS)	
Occupational Employment Statistics	
Quarterly Census of Employment and Wages (QCEW)	
State and Metro Area Employment	
State and Local Area Unemployment Statistics	
Department of Education	**Commercial**
Academic Library Survey	
Adult Literacy and Life Skills Survey	College Board
Common Core of Data	Donors Choose
Consolidated State Performance Report	eSparks
EDFacts	Glassdoor
Fast Response Survey System	Great School Ratings
Federal Student Loan Program	LinkedIn
Free Application for Federal Student Aid (FAFSA)	Location Inc (Neighborhoodscout)
High School and Beyond Survey	Maponics - School Boundaries
National Assessment of Educational Progress	Monster Resume Database
National Household Education Surveys Program	National Student Clearinghouse
National Postsecondary Student Aid Study	School Attendance Boundary Information System (SABINS)
National Study of Postsecondary Faculty	School Digger
Private School Universe Survey	US News and World Report Rankings

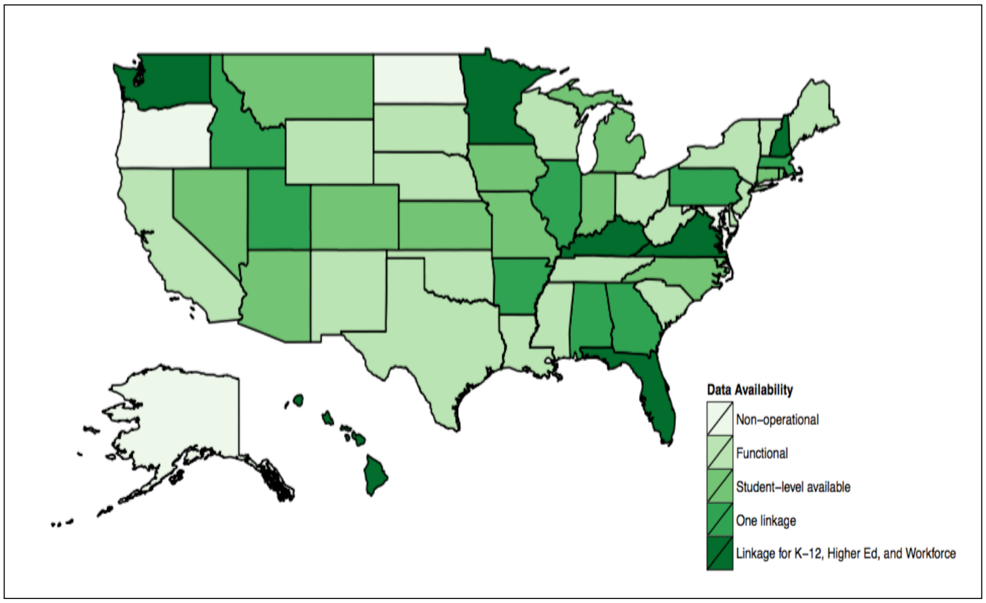
3. Inventory of SLDS Data Sources

The screening process highlighted the various stages of maturity of the SLDS systems across the 50 states and the District of Columbia. The states were categorized into the following five categories:

- States with linkages across K-12, higher education, and workforce (7 states)
- States with linkage across K-12 to higher education (8 states)
- States with only K-12 data (13 states)
- States with functional SLDS that have challenges, e.g., very restrictive interpretation of FERPA or required onsite presence to use some of the data, especially workforce data (18 states)
- States that do not have SLDS or are in the process of creating an SLDS (4 states and the District of Columbia)

The map in Figure7.1 summarizes this categorization and helps to indicate which states warranted a full inventory at this stage in the project. SDLS data sources from thirty-three states passed the initial screening and went through a full inventory process. Throughout the course of the project, a full inventory for all 50 states and the District of Columbia was conducted. The results are documented on the collaborative wiki (Keller et al. 2016).

Figure 7.1: Maturity of State Longitudinal Data Systems.



The status of each SLDS system is displayed across the 50 States.

4. Selection of SLDS Data Sources

Based on the full data inventories, 10 states were selected as potential candidates for this case study. After further review, five states were included in the study for the following reasons:

- Kentucky (linkage for K-12, higher education, and workforce; a straightforward review process although it took longer than expected)
- North Carolina (K-12; we already had the data)
- Texas (K-12; fast process to obtain the data)
- Virginia (linkage for K-12, higher education, and workforce; used publically available data about students in each K-12 public school)
- Washington (linkage for K-12, higher education, and workforce; links in-state students and out-of-state students; a straightforward review process, although it took longer than expected to get the data)

The other five states considered were Arkansas, Massachusetts, New Hampshire, New York, and Ohio. These five states, with the exception of Arkansas, link K-12 to higher education data, but not to workforce data. These states were not chosen for this case study because of their lengthy acquisition processes and the expectation that the data would not be received in enough time to meet project deadlines.

B. Acquiring Education Data Sources

This section provides a brief description and summary of the acquisition process necessary to obtain the data from the five selected states for the study period of 2009-2013. The process involved discussions with the state liaison for each of the selected states and iterations in completing the forms.

1. Kentucky Data

Acquiring Kentucky Longitudinal Data System (KLDS) data required a Memorandum of Understanding (MOU) with the Kentucky Center for Education and Workforce Statistics to describe the purpose of the research and data needed. Institutional Review Board (IRB) approval was not required. The cost for 5 years of data was $8500.

KLDS data are available for 2009-2014 and each subsequent year's data are released in the following year, e.g., 2015 data will be released in 2016. KLDS data were gathered from various agencies including the Kentucky Department of Education, the Council on Postsecondary Education, Educational Professional Standards Board, the Kentucky Higher Education Assistance Authority, and the Kentucky Education and Workforce Development Cabinet.

These data cover populations of students who attend:

- Preschool and Early Childhood
- K-12 students in public schools in Kentucky
- Higher education, public and independent institutions
- Technical education

The KLDS data can be linked from pre-school through workforce.

Initially officials estimated the time to receive the data following submission of the MOU to be about one month. We were discouraged from asking for children's free and reduced price eligibility information as that would delay delivery of the data by several months. We were also told that we would receive K-12 data but would have to request the KLDS programmers to run programs for Higher Education and Workforce analysis. In the end, the receipt of data took three months, in part due to many agency heads leaving their job since this was an election year and signatures were required from each agency providing the data; we received all the data from K-12 through higher education, with the understanding that workforce data outputs would be obtained at a future date; and we did not ask for the children's free and reduced price eligibility information at this time.

2. North Carolina Data

North Carolina data are maintained and distributed through the North Carolina Education Research Data Center at Duke University's Center for Child and Family Policy. Detailed longitudinal data were available from 1995 to 2014 but the time frame varies by the type of data. Data are available for students, teachers, classrooms, schools, and school districts. Our case study focuses on students.

Acquiring North Carolina Longitudinal Data Systems (NC SLDS) data required completion of several forms:

- Data Use Application and Agreement
- Proposal for Using the Data
- IRB approval, required to be updated annually
- Confidentiality Agreement for Investigators, which must be signed by everyone working on the project
- Data request with rationale required for each variable requested
- Disclosure and Data Destruction Agreement, data must be destroyed after three years from initial receipt of the data

The process was straightforward and data were received within a few days of submitting the required paperwork. The cost was $2500.

3. Texas Data

The Texas Education Agency requires minimal paper work for obtaining data from the Texas Longitudinal Data System (TLDS). The cost was $486. The process took about 6 weeks.

Unlike other states, Texas interprets the Family Educational Rights and Privacy Act (FERPA) very strictly for research purposes. The agency will not allow release of data with fewer than 5 cases in a cell. We requested enrollment counts by grade level, ethnicity, gender, reason left school, and other variables. The result was that we received approximately 1 million out of 5 million records due to the exclusion of records that had cells containing fewer than 5 cases. The 1 million records were not a representative sample so even weighting the data did not summarize to the total enrolled student population because certain cells were empty, and therefore, could not be weighted.

4. Virginia Data

The Virginia Longitudinal Data System (VLDS) has longitudinal data that are probabilistically linked from pre-school to K-12 to higher education to workforce. Virginia requires that requests for data be made through the Virginia Department of Education (VDOE) or the State Council of Higher Education for Virginia (SCHEV). The work must be of direct interest to the VDOE and SCHEV contacts and they may require specific research as a condition for receiving the data. After a quick agreement on the research conditions, it took months to receive the data. What we received was a subset of the data requested, for reasons unknown. We received selected higher education data from SCHEV.

Fortunately, Virginia releases aggregate data for K-12 schools that includes enrollment counts and demographic information by grade (Virginia Department of Education). These data were used for the K-12 enrollment comparisons to ACS estimates in Chapter 8.

5. Washington Data

Washington requires a data sharing agreement that lasts for 2 years. The agreement has multiple parts: purpose, timeframe, requirements for data transmission and storage, Statement of Confidentiality and Non-Disclosure, and Certification of Data Disposition of Data. The request was made in early June and the data were received by mid-November. Washington Longitudinal Data Sets (WLDS) data are very popular for use by researchers and thus our request was put into the queue and took 5 months to implement. Washington does not charge for the files and they do not require an IRB approval.

Washington administrative records include data on the population from pre-school through workforce. A dataset of high school graduates that links to higher education was provided, although we had requested K-12 data as well. We expect to receive linkable workforce dataset once an initial analysis of the received data is completed.

C. Education Data Profiling, Cleaning, and Transformation

The data profiling, cleaning, and transformation activities for the state-level education data sources provided input to and helped to formalize the data framework described in Chapter 3. As discussed in that chapter the steps are designed to assess whether the data correctly represent the real-world construct to which it refers.

Data profiling was undertaken to assess the quality, consistency, uniqueness, and duplicates in the data sources. Through this process, we identified questions that resulted in emails and discussions with the state data providers. The resolution of each question is documented on the collaborative wiki (Keller et al. 2016). This step also involved reviewing the documentation, primarily codebooks, and profiling the data. The results of this step are summarized here for each of the states used in this case study.

Data cleaning primarily involved removing duplicate records and ensuring consistency of demographic characteristics. In general, the first duplicate record was kept and the others removed. For demographics, generally, the most recent demographic characteristic (by year) for that variable was used. Records showing students in the same grade for 3 or more years were flagged but not changed.

Data transformation varied by state but involved adding school district or county codes to the records, computing age based on birth dates, and for Texas, computing count adjustment weights since only 1/5th of the records were obtained.

1. Kentucky

Kentucky provides three reports: a data dictionary, a set of postsecondary comprehensive database reporting guidelines, and additional codebook questions. Officials provided maps of the county and school districts and a special map of the Appalachian region. FERPA guidelines are also provided to indicate conditions that may require suppression of data. The data received for this project however were not suppressed or redacted since we have an MOU in place. The data are composed of Pre-school/early education, K-12, higher education, technical education, and demographics datasets.

Data profiling. The following variables were profiled for K-12: race/ethnicity, gender, birth year, grade, student identification number, district code, year, high school dropout code, high school dropout reason, high school graduation indicator, and limited English proficiency indicator. The K-12 data appeared to be of high quality with fewer than 0.2% duplicates and less than 1% missing values.

For higher education, the following variables were profiled: race/ethnicity, gender, birth year, student identification number, institution, year, level of school (e.g., undergraduate, graduate student), graduation degree (e.g., BA, Masters), and degree area (e.g., business, education). These data had few duplicates and missing values (< .1%).

Data cleaning. Out of the 2,147 duplicates (in terms of student ID by year only), 830 (39%) were identical duplicate cases across variables of interest including district, grade, reported graduated, dropout reason, and limited English proficiency. Among the duplicates that were left, most had duplicate student identification numbers and inconsistent districts or grade. There was not a clear pattern to these duplicates. Given the small number of duplicates and the lack of a consistent pattern, the first entry for a student in the dataset was retained and the second duplicate record deleted. For instance, among the duplicates with inconsistent dropout reasons, 97% (94 out of 96) had "NULL" for one entry, which signified that they did not drop out, and a dropout reason for the other entry, which signified that they did drop out. In these cases, the first entry for a student was retained. Small numbers of demographics were inconsistent: birth date (0.31%), gender (0.35%), race/ethnicity (2.54%), and grade (0.9%). In general, the most recent characteristic (by year) for that variable was used.

Data transformation. Age was computed based on birth date. Counties were also added by matching school district with county five-digit Federal Information Processing Standard (FIPS) code since student information was compared to the ACS estimates at the county-level.

2. North Carolina

North Carolina provides several codebooks: student demographics and attendance; graduates, dropouts, growth, end of course, course membership, and Masterbuild which is the student level academic summary documented at the end of the school year. The files are provided separately by student, teacher, classroom, school, and district levels. This study profiled data from the Masterbuild, Growth, Course Membership, and Demographics files. The most useful file for this project was the Demographic file.

Data profiling. The data files were well-constructed with most variables of interest being complete and valid (i.e., sex, race/ethnicity, limited English proficiency, economically disadvantaged, reporting year, and Local Education Agency). The student identification number and grade variables were 99% complete.

Data cleaning. There were a very small number of duplicate records (<.0003%). The duplicate records were removed. There was also a small number of inconsistencies in demographic variables across 2009-2013 (<2.5%). In order to create consistent demographic variables, the most frequent value was used for birth date and race/ethnicity. If counts were equal, then the most recent value was used. For gender, the most recent value was used. There were <1% missing values with the exception of grade, which was missing for 18% of the students in 2009.

The North Carolina Education Research Data Center reported that including demographic information for students in grades pre-kindergarten through 2nd grade in the datasets was not required and therefore the information for students in these grades may not be complete. In 2010-2013, the amount of missing grade information was <1%. As shown in Table 7.2, the student enrollment was lower as a result of cleaning the grade variable. Ultimately, the student count based on those who had a valid grade was used to get counts for all other demographics, because students without a valid grade value could not be used for ACS direct estimates.

Table 7.2: Student Enrollment Counts for North Carolina

Year	Overall Student Count	Student Count for those with Valid Grade
2009	1,457,411	1,229,276
2010	1,456,969	1,456,952
2011	1,471,276	1,470,599
2012	1,478,000	1,478,084
2013	1,492,353	1,492,353

Source: Unadjusted counts and adjusted counts based on Valid Grade for School Year, North Carolina SLDS, 2008-2013.

Data transformation. Age was computed based on birth dates. Counties were also added by matching school district with county FIPS code since student information was compared to the ACS estimates at the county-level.

3. Texas

Texas provides three types of codebooks. The first is the main codebook that describes the 326 variables; the second provides geographic information for each school district; and the third provides the codes delineating the reasons why students leave a school. There are separate data sources for:

- Student demographics (student-level)
- Class (class-level)
- Course (course-level, which includes multiple classes for one type of course)
- Discipline (student-level)
- Employment (staff-level)
- Non-Class Employment (staff-level)
- Assessement (student-level)

Data profiling. Fewer than 2% of the rows were duplicates (1.64% to 1.73% by year). There were zero percent missing for the Scrambled ID (linkable student identifier), year, school district, grade, race/ethnicity, gender, and exit reason. The data for 2009-2011 contained less than one percent missing for economic status. Some additional reports were helpful in understanding the Texas data, including the Texas Comptroller tax information about county codes, the Texas Education Agency's description of county and school districts, and the National Center for Education Statistics digest for state-level enrollment counts.

Data cleaning. For the duplicate records (<2%), the first entry for a student in a data set was retained. For the small number of inconsistent demographic variables, i.e., gender code (0.07%) or race/ethnicity (0.33%), the most recent code was used. Students in grades for more than three years were flagged.

Data transformation. Two data transformations were necessary: (1) adding county information and (2) weighting. Counties were not included in the dataset. Since student information was compared to the ACS estimates at the county-level, this information was matched with school districts and added to the dataset. The Texas Education Agency provided codebooks matching counties and school districts for 2009-2013. These were profiled and cleaned before matching to the datasets.

Texas suppressed any cells with fewer than 5 students and thus provided 1 million of the 5 million student records. We created simple weights to compare the restricted Texas SLDS data to ACS estimates. The weights were created by taking the state-level aggregate student enrollment counts by grade, gender, and race/ethnicity from published reports by the Texas Education Agency (TEA 2015) and dividing this number by the equivalent grade, gender, and race/ethnicity original enrollment counts that were tabulated from the restricted Texas SLDS files.

The ACS tables report student enrollment estimates in grade by gender groups and in grade by race/ethnicity groups. Therefore, we calculated overall weights to compare SLDS data to the ACS data for these demographic characteristics using

$$w_{gr} = \frac{w_g n_g + w_r n_r}{n_g + n_r}, g = 1, \ldots, 4, r = 1, \ldots, 4,$$

where

- g is grade (PK-KG, 1-4, 5-8, 9-12)
- r is race/ethnicity group (White, Black, Hispanic, Other)
- n_g is the enrollment count tabulated from the restricted Texas SLDS data for the grade group g

- n_r is the enrollment count tabulated from the restricted Texas SLDS data for the race/ethnic group r
- w_g is the weight for the grade group g calculated by dividing the TEA official (total) enrollment counts by n_g
- w_r is the weight for the race/ethnic group r calculated by dividing the TEA official enrollment counts by n_r

The weights, w_{gr}, are multiplied by the total enrollments tabulated from the restricted Texas SLDS data to create estimates to compare to the ACS tabulations. Table 7.3 presents an example of the weight calculations for 2013 for grade and race/ethnicity groups.

Table 7.3: Averaged Weights Used to Match Texas State-Level Counts by Race/Ethnicity and Grade

Race/Ethnicity	Grade	Grade Weight	Texas SLDS Grade Counts	Race/Ethnicity Weight	Texas SLDS Race Counts	Overall Weights
Black	PK-KG	5.91	108,472	6.67	96,939	6.27
Hispanic	PK-KG	5.91	108,472	3.02	862,735	3.34
Other race	PK-KG	5.91	108,472	14.01	21,553	7.25
White	PK-KG	5.91	108,472	4.63	328,515	4.95
Black	1-4	5.40	286,806	6.67	96,939	5.72
Hispanic	1-4	5.40	286,806	3.02	862,735	3.61
Other race	1-4	5.40	286,806	14.01	21,553	6.00
White	1-4	5.40	286,806	4.63	328,515	4.99
Black	5-8	2.95	508,133	6.67	96,939	3.55
Hispanic	5-8	2.95	508,133	3.02	862,735	2.99
Other race	5-8	2.95	508,133	14.01	21,553	3.40
White	5-8	2.95	508,133	4.63	328,515	3.61
Black	9-12	3.41	406,331	6.67	96,939	4.04
Hispanic	9-12	3.41	406,331	3.02	862,735	3.14
Other race	9-12	3.41	406,331	14.01	21,553	3.94
White	9-12	3.41	406,331	4.63	328,515	3.96

Note: The Grade and Race Weights were created by taking the TEA published student enrollment counts for grade and race/ethnicity and dividing this number by the enrollments tabulated from the restricted Texas SLDS data. (**Sources:** Texas Education Agency published total enrollment count data 2013; SLDS Texas data received 2013).

4. Virginia

Data profiling. Publically available data at the school level were used for PK-12 data. The data appeared to be of high quality. The number of unique school numbers and names varied slightly by year. School numbers are not unique across the state; i.e., the same school number can be used in different districts. Cells are suppressed with fewer than 10 cases for economically disadvantaged, limited English proficiency status, and/or disability status.

For the PK-12 Fall membership counts and the high school graduation counts, there were no duplicates and the variables were consistent - these variables are school year, district number, district name, fall membership count, grade code, gender, federal race code, limited English proficiency flag, and disadvantaged flag. The postsecondary education file was similar and has two additional variables, institution type (2-year, 4-year, public, private) and enrollment counts.

Data transformation. The Fall membership dataset was used to create three main tables: (1) a race/ethnicity table that provides student count's by race/ethnicity and grade; (2) a gender table that provides student counts by gender and grade; and (3) a disadvantaged table that provides student counts by disadvantaged status and grade. Because of how the Fall membership dataset is structured, creating these tables involved filtering the Fall membership dataset based on the variables of interest. For example, for the race/ethnicity table, the variables of interest were school year, district number, school, grade, and race/ethnicity. The dataset was filtered for observations where the values for these variables were not blank or "NULL". The collaborative wiki (Keller et al. 2016) provides the rules.

D. Summary

The Statewide Longitudinal Data Systems (SLDS) systems were the primary source of data examined for this education case study. After completing an inventory of the SLDS systems for each of the 50 states, 5 states were selected for this study based on the described quality of the data and ability to obtain the data within the timeframe for this study: Kentucky, North Carolina, Texas, Virginia, and Washington. Implementing the data framework involved profiling and cleaning student data; e.g., student identification number, district code, year, gender, race/ethnicity, grade, age, and other characteristics such as indicators for limited English proficiency and high school dropouts. Most variables were valid and consistent and there were very few missing values or duplicates.

8. Education Comparisons

A. Data Comparisons

A central focus of this study was to determine if ***external*** data, that is, data external to the federal statistical system, can provide estimates comparable to American Community Survey (ACS) statistics. This chapter presents some of the benchmarking of the external state-level administrative educational data with the 2009-2013 ACS estimates. ACS tabulations used for comparison are at the state, county, and school district levels accessed through the US Census Bureau's *American FactFinder* (Census Bureau 2013).

It is useful to recognize some fundamental differences between ACS data and the State Longitudinal Data System (SLDS) data sources to be used in the education benchmarking. First, ACS data come from a carefully designed national study. The data provide estimates and margins of error for several variables of interest to the government and the populace. Methods for weighting and imputation of non-response have been developed and applied to the ACS data to provide official statistical estimates at the national, state, multi-county, county, and sub-county levels.

Educational data from SLDS provide a rich source of data for complementing or supplementing ACS data about student enrollment by state, school district, grade, and demographic characteristics. In contrast to the ACS data, the SLDS data sources acquired for this study are not samples. Rather, these data represent administrative data on ***all*** students enrolled in public schools in the state. These data could be thought of as the census of the complete population of public school students from which the ACS samples. For some states, some of the SLDS data that are made available to researchers and those outside of the state education departments may be suppressed due to the Family Educational Rights and Privacy Acts (FERPA). Weights may need to be developed if data are suppressed.

ACS data on school enrollment counts and characteristics of students are available for the U.S., states, counties, selected metropolitan and micropolitan statistical areas, cities, school districts, census tracts, and selected zip codes from 2000 to the present. There are more than 20 tables that provide school enrollment counts and characteristics of students. Many of the ACS enrollment counts include those attending public and private schools, although some tabulations, such as the school enrollment differentiate between public and private schools. The SLDS enrollment counts are for public schools. The primary SDLS and ACS data comparisons presented in this chapter are the enrollment counts for public schools for the ACS tables listed in Table 8.1.

Table 8.1: ACS Tables used for Benchmarking by SLDS Data Sources

ACS Table	Kentucky	North Carolina	Texas	Virginia
B14002 Sex by Enrollment by Level of School by Type of School	X	X	X	X
B14003 Sex by Enrollment by Type of School by Age	X	X		
B14007 School Enrollment by Detailed Level of School				
B14007B Black Alone	X	X	X	X
B14007C American Indian and Alaska Native Alone	X	X	X	X
B14007D Asian Alone	X	X	X	X
B14007E Native Hawaiian and Pacific Islander Alone	X	X	X	X
B14007G Two or More Races	X	X	X	X
B14007H White Alone, Not Hispanic or Latino	X	X	X	X
B14007I Hispanic or Latino	X	X	X	X

Source: US Census Bureau's *American FactFinder*

B. Reasons for Discrepancies in Enrollment Counts

When comparing ACS data with SLDS data, differences in enrollment counts are expected. There are four sources of differences. The first is that the geographic location of the ACS data is based on where an individual resides. The geographical location of the SLDS education data is based on school location. Second, the geographical boundaries of school districts change which can affect enrollment counts. These boundaries can potentially change every year, however, ACS only updates the boundary information every other year.

Third, some ACS tables group students by public and private schools, while other ACS tables do not differentiate between the school types. Public schools include charter schools and private schools include students who are home-educated. The SLDS data are reported for public schools only. While the ACS Public Use Microdata Sample (PUMS) education data differentiates between public and private school, this data is geographically aligned with Public Use Microdata Areas (PUMAs) and state data are aligned with school district and county areas. PUMA areas do not always align with the geographic boundaries of the school districts or counties, therefore, new geographic areas would need to be created to compare PUMS data with SLDS data. Fourth, charter schools are often public schools, however, they are not geographically bound to a school district. Charter school enrollment counts from the SLDS data would need to be geographically assigned to a school district or county in order to include them in the benchmark comparisons to ACS.

A potential fifth difference is due to timing. ACS data are collected year-round, while SLDS enrollment counts are collected at different times during the year depending on the state (e.g., the Fall, Spring, and Summer terms). For this study, the Fall counts are used for all states with the exception of North Carolina where Spring counts were available. The ACS questionnaire asks if the person has been enrolled in the past three months. ACS data are based on an average of two school years which could result in differences with SLDS counts. A study conducted by Census Bureau researchers that compared ACS estimates with the Common Core of Data collected by the Department of Education concludes that timing differences in data collection do not completely explain the differences between data sources (Davis and Bauman 2011). They found that the ACS underestimates school enrollments compared to public school enrollment counts from the Common Core of Data. This is relevant because the Common Core of Data comes from the SLDS data.

Others have documented deficiencies in ACS data. For instance, the data do not provide information on the location of school, attendance, or migration during schooling (Plunk et al. 2014, Goodman 2012). There are large errors in areas with small populations and for race/ethnic groups that comprise a small share of those sampled (e.g., Native Hawaiian and Pacific Islanders). In addition, tracking individuals longitudinally as they progress through school is not possible, although the ACS was not designed to do this. The SLDS data could supplement ACS data products by providing geographic and longitudinal data.

C. Benchmark Comparisons

Examples of Benchmark comparisons are provided for four states selected through the data discovery and inventory process described in Chapter 7. A more comprehensive examination is provided on the collaborative wiki (Keller et al. 2016).

The following ratio is used as a figure of merit or fitness throughout the comparisons:

$$Fitness\ Ratio = FR = \frac{ACS\ estimate - SLDS\ estimate}{90\%\ ACS\ margin\ of\ error}$$

When *FR* falls outside the ±1 range, the benchmark estimates are not within the 90% ACS margin of error (MOE). It is important to note that falling outside the MOE does not mean the estimates are bad. The ground truth or the gold standard is unknown for the quantities being estimated. Therefore, estimates that do not match could be equally valid or invalid. The question that needs to be answered is whether they are useful for the intended purposes or not.

To compare ACS and SLDS data, data are presented as (1) enrollment counts, (2) differences in counts using the *Fitness Ratios*, (3) boxplots of the *Fitness Ratios* by grade or grade clusters and by demographics (male/female or race/ethnicity), and (4) geographic distributions of *Fitness Ratios*.

1. Summary of Enrollment Counts for SLDS States: Kentucky, North Carolina, Texas, and Virginia

Table 8.2 presents state-level enrollment counts for ACS and SLDS for four states, Kentucky, North Carolina, Texas, and Virginia for individual years 2009-2013 and the 5-year estimate for 2009-2013. The comparisons presented in this chapter compare ACS and SLDS enrollment counts for public schools. Exceptions are noted in the text if the comparison is ACS total (public and private) enrollment counts to SLDS public schools enrollment counts.

Very few of the state estimates are "within the 90% ACS MOE," which means the estimates are different. At the state level, SLDS enrollment counts are generally lower across the four states than ACS estimates. This relationship generally holds when examining the demographic groups within each of the four states examined in this study. As will be shown, there are many counties (and school districts) within each state for which the estimates are "within the 90% ACS MOE" and thus essentially comparable.

Table 8.2: Comparison of ACS and SLDS State Enrollment Counts for North Carolina, Texas, Virginia, and Kentucky, 2009-2013

2009-2013	ACS and SLDS Comparison for North Carolina, Texas, Virginia, and Kentucky					
Description of Variable	ACS Estimate	90% MOE (+/-)	SLDS Estimate	FR	Difference	Within 90% MOE?
North Carolina 2009	1,555,812	18,828	1,229,276	17.3	326,536	NO
North Carolina 2010	1,586,213	17,692	1,456,725	7.3	129,488	NO
North Carolina 2011	1,604,308	20,903	1,469,945	6.4	134,363	NO
North Carolina 2012	1,604,960	17,793	1,477,229	7.2	127,731	NO
North Carolina 2013	1,615,186	17,955	1,491,204	6.9	123,982	NO
North Carolina 2009-2013	1,593,638	7,164	1,424,882	23.6	168,756	NO
Texas 2009	4,819,279	31,819	4,666,700	4.8	152,579	NO
Texas 2010	4,957,932	32,287	4,752,354	6.4	205,578	NO
Texas 2011	5,008,904	35,823	4,841,500	4.7	167,404	NO
Texas 2012	5,065,706	31,803	4,908,614	4.9	157,092	NO
Texas 2013	5,111,506	33,903	4,961,187	4.4	150,319	NO
Texas 2009-2013	4,987,979	14,048	4,826,076	11.5	161,903	NO
Virginia 2009	1,244,864	14,951	1,246,088	-0.1	-1224	YES
Virginia 2010	1,275,668	15,376	1,254,112	1.4	21,556	NO
Virginia 2011	1,281,586	16,729	1,260,816	1.2	20,770	NO
Virginia 2012	1,279,721	14,975	1,267,073	0.8	12,648	YES
Virginia 2013	1,297,388	14,734	1,275,401	1.5	21,987	NO
Virginia 2009-2013	1,277,023	6,570	1262149	2.3	14,874	NO
Kentucky 2009	690,239	10,307	700,161	-1.0	-9,922	YES
Kentucky 2010	702,630	11,682	707,972	-0.5	-5,342	YES
Kentucky 2011	713,287	11,200	713,847	0.0	-560	YES
Kentucky 2012	701,847	10,579	716,120	-1.3	-14,273	NO
Kentucky 2013	699,568	10,875	717,644	-1.7	-18,076	NO
Kentucky 2009-2013	700,031	4,674	711,152	-2.4	-11,121	NO

Note: "FR" stands for *Fitness Ratio*; MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error. Enrollment counts are for public schools. (**Sources:** SLDS data for North Carolina, Texas, Virginia, and Kentucky data 2009-2013; ACS 2009-2013 1-year and 5-year estimates).

2. Benchmark Comparisons for North Carolina

Table 8.3 and Figure 8.1 present ACS and North Carolina SLDS data comparisons by gender and grade group, employing the same categories the ACS uses to tabulte the data (i.e., pre-kindergarten, kindergarten, grades 1-4, 5-8, and 9-12). ACS estimates are generally higher for all groups. The *Fitness Ratio* ranges from 0.7 to 19.8 for the enrollment counts by gender and grade group.

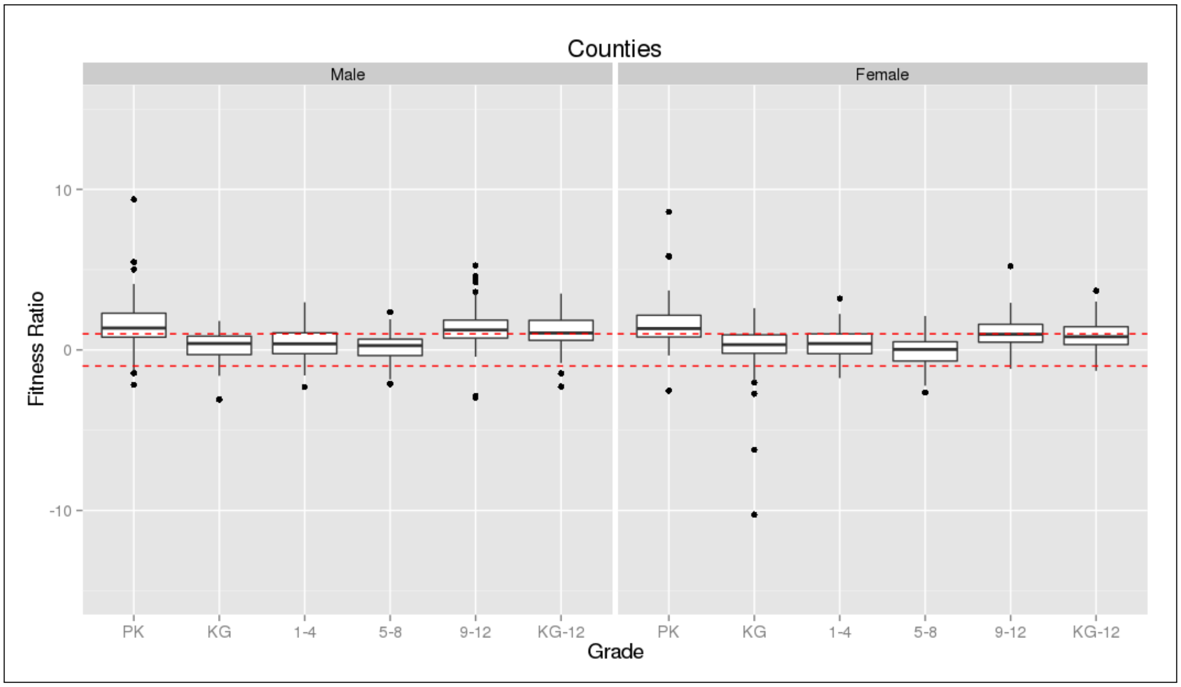
As shown in Figure 8.2, there are many counties in which the ACS and SLDS enrollment counts for females and males are comparable, (i.e., *Fitness Ratio's* between ±1), but many counties where they are not. In particular, the ACS estimates tend to be higher in the more urban areas, as indicated by the blue areas in Figure 8.2.

Table 8.3: ACS and North Carolina SLDS State Enrollment Counts Comparison, by Gender and Grade Group, 2013 5-year estimate, 2009-2013

	ACS			North Carolina		Comparison		
Variable	Estimate	90% MOE (+/-)	Dist.	Estimate	Dist.	FR	Difference	Within 90% MOE?
Total	1,593,638	7,164	100%	1,424,882	100%	24	168,756	NO
Male Pre-Kindergarten	40,848	1,539	3%	10,312	1%	19.8	30,536	NO
Male Kindergarten	60,086	1,759	4%	53,157	4%	3.9	6,929	NO
Male grades 1-4	239,215	3,009	15%	224,248	16%	5.0	14,967	NO
Male grades 5-8	238,311	2,707	15%	232,303	16%	2.2	6,008	NO
Male grades 9-12	246,159	2,628	15%	208,431	15%	14.4	37,728	NO
Female Pre-Kindergarten	35,336	1,453	2%	8,101	1%	18.7	27,235	NO
Female Kindergarten	58,117	1,513	4%	49,898	4%	5.4	8,219	NO
Female grades 1-4	223,903	2,262	14%	213,581	15%	4.6	10,322	NO
Female grades 5-8	224,231	2,654	14%	222,258	16%	0.7	1,973	YES
Female grades 9-12	227,432	2,443	14%	202,593	14%	10.2	24,839	NO

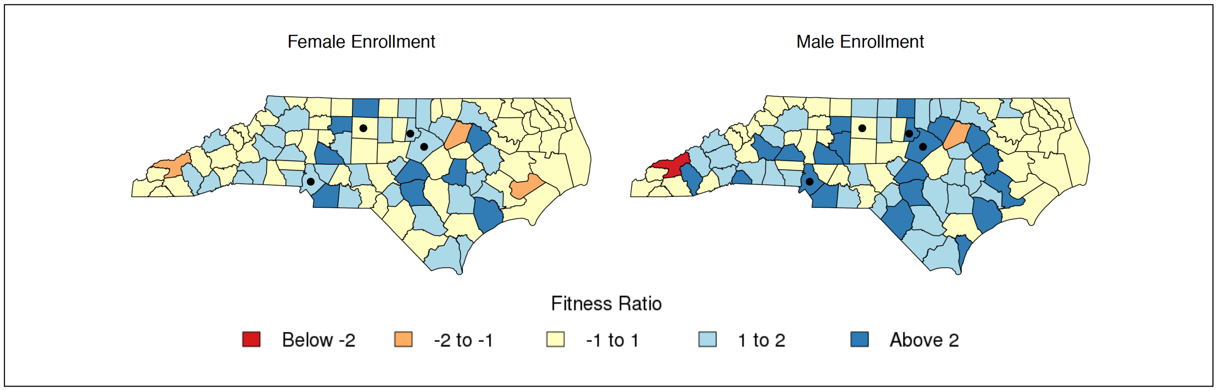
Note: "Dist." stands for Distribution; "FR" stands for *Fitness Ratio*; "Difference" refers to the difference between the ACS and SLDS estimates; MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error. Enrollment counts are for public schools. (**Sources:** North Carolina SLDS data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.1: *Fitness Ratio* Scores for North Carolina Counties by Gender and Grade



Boxplots compare the distribution of the *Fitness Ratios* at the county level and their relation to the ACS MOE. Estimates falling outside the red reference lines of ±1 are not within the 90% ACS margins of error. PK refers to pre-kindergarten and KG refers to kindergarten. Enrollment counts are for public schools. The KG-12 category represents the total enrollment counts across all grades. Total number of counties=100. There are outliers for the male and female pre-kindergarten and female kindergarten groups, but the number of outliers is low. (**Sources:** SLDS North Carolina data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.2: *Fitness Ratios* across North Carolina Counties for Females and Males



Enrollment counts are for public schools and do not include pre-kindergarten. Black points represent the four largest cities in North Carolina (from west to east): Charlotte, Greensboro, Durham, and Raleigh. Total number of counties=100. Many of the enrollment counts are similar between ACS and SLDS data (yellow counties) and the counties in which the counts tend to be different (dark blue counties) are around urban areas. (**Sources:** SLDS North Carolina data 2009-2013; ACS 2009-2013 5-year estimates).

3. Benchmark Comparisons for Texas

Texas provided 1 million of the 5 million student records due to their interpretation of FERPA. The Texas Education Agency suppressed all cells with fewer than 5 students. Section 7.C.3 described the weights we created in an attempt to adjust the Texas SLDS data.

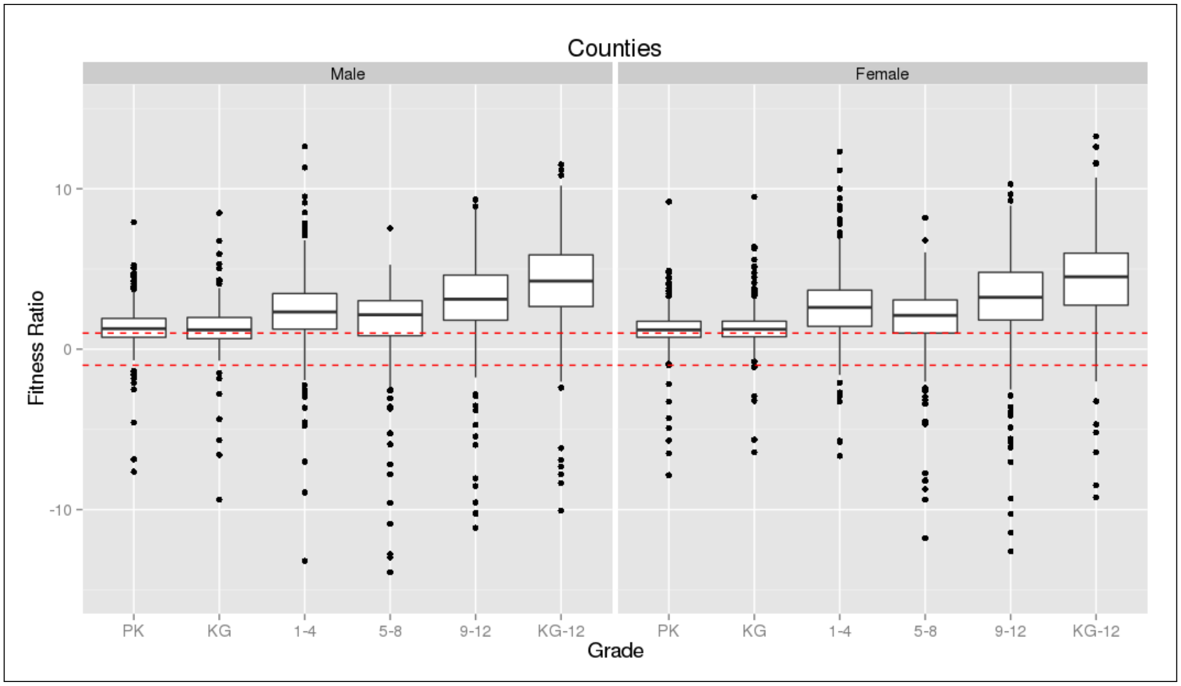
Table 8.4 and Figure 8.3 present enrollment counts for ACS and the adjusted Texas SLDS data by gender and grade group for the 5-year estimates (2009-2013). The *Fitness Ratio* ranges from -23.9 to + 22.0 and none of the groups fall within the ACS 90% MOE. As shown in Figure 8.4, the ACS estimates tend to be higher than the adjusted SLDS estimates for the majority of counties in Texas, as indicated by a *Fitness Ratio* above 2. The opposite is true, however, for the areas around cities where the adjusted SLDS estimates tend to be higher than the ACS estimates. This discrepancy is likely due to the amount of Texas data that was suppressed. The Texas SLDS data received was a non-representative sample of Texas student records. We weighted the SLDS data in an effort to create enrollment count estimates that could be compared to ACS tabulations. It appears that the weighting scheme cannot overcome the non-representative nature of the Texas SLDS data extract.

Table 8.4: ACS and Texas SLDS State Enrollment Counts Comparison, by Grade and Grade Group, 2013 5-year estimate, 2009-2013

	ACS			Texas		Comparison		
Variable	Estimate	90% MOE (+/-)	Dist.	Estimate	Dist.	FR	Difference	Within 90% MOE?
Total	4,987,979	14,048	100%	4,826,076	100%	11.5	161,903	NO
Male Pre-Kindergarten	148,463	3,075	3%	97,050	2%	16.7	51,413	NO
Male Kindergarten	196,545	3,176	4%	128,474	3%	21.4	68,071	NO
Male grades 1-4	758,378	5,668	15%	665,237	14%	16.4	93,141	NO
Male grades 5-8	737,317	5,304	15%	864,183	18%	-23.9	- 126,866	NO
Male grades 9-12	728,966	4,042	15%	748,991	16%	-5.0	- 20,025	NO
Female Pre-Kindergarten	135,806	3,029	3%	97,089	2%	12.8	38,717	NO
Female Kindergarten	183,549	3,014	4%	117,182	2%	22.0	66,367	NO
Female grades 1-4	717,560	5,717	14%	606,147	13%	19.5	111,413	NO
Female grades 5-8	699,068	5,595	14%	823,862	17%	-22.3	- 124,794	NO
Female grades 9-12	682,327	4,344	14%	677,861	14%	1.0	4,466	NO

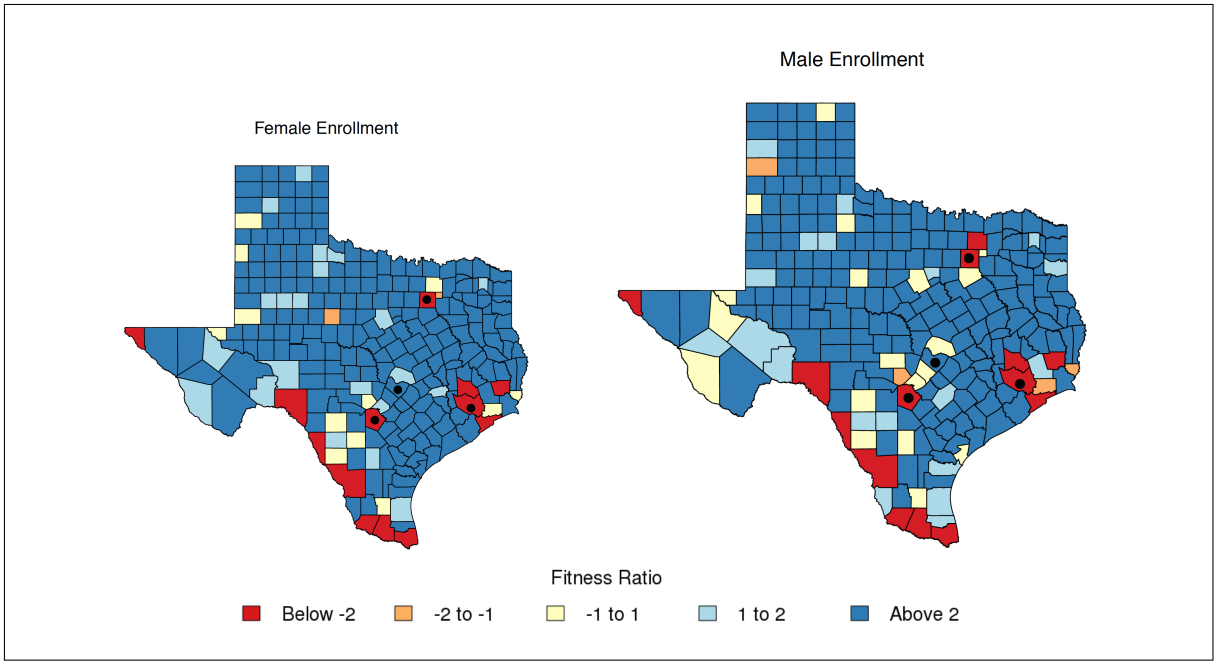
Note: “Dist.” stands for Distribution; “FR” stands for *Fitness Ratio*; “Difference” refers to the difference between the ACS and SLDS estimates; MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error. Enrollment counts are for public schools. (**Sources:** SLDS Texas data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.3: *Fitness Ratio* Scores for Texas Counties by Gender and Grade



Boxplots compare the distribution of the *Fitness Ratios* at the county level and their relation to the ACS MOE. Estimates falling outside the red reference lines of ±1 are not within the 90% ACS margins of error. PK refers to pre-kindergarten and KG refers to kindergarten. Enrollment counts are for public schools. The KG-12 category represents the total enrollment counts across all grades. Total number of counties=254. Extreme outliers were removed from the following grade groups: male 1-4 (2 outliers, ratio range= -32 to 15); female 1-4 (3 outliers, ratio range= -27 to 18); male 5-8 (4 outliers, ratio range= -32 to 8); female 5-8 (5 outliers, ratio range= -31 to 8); male 9-12 (5 outliers, ratio range= -34 to 9); and female 9-12 (4 outliers, ratio range= -29 to 10). Most of the *Fitness Ratios* fall above the upper 90% ACS MOE. (**Sources:** SLDS Texas data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.4: ***Fitness Ratios*** **across Texas Counties for Females and Males**



Enrollment counts are for public schools and do not include pre-kindergarten. Black points represent the four largest cities in Texas (from west to east): San Antonio, Austin, Dallas, and Houston. Total number of counties=254. The majority of counties in the state have *Fitness Ratios* above 2, except for in uban areas and southern parts of the state, which have *Fitness Ratios* below -2. (**Sources:** SLDS Texas data 2009-2013; ACS 2009-2013 5-year estimates).

4. Benchmark Comparisons for Virginia

School districts align with counties or cities in the Commonwealth of Virginia. The *Fitness Ratios* of enrollment counts between ACS and the Virginia Longitudinal Data System (VLDS) data given in Table 8.5 and Figure 8.5 indicate the estimates are comparable, with some exceptions. The extremes are the pre-kindergarten *Fitness Ratios* of 7.6 and 9.9 for male and female, respectively.

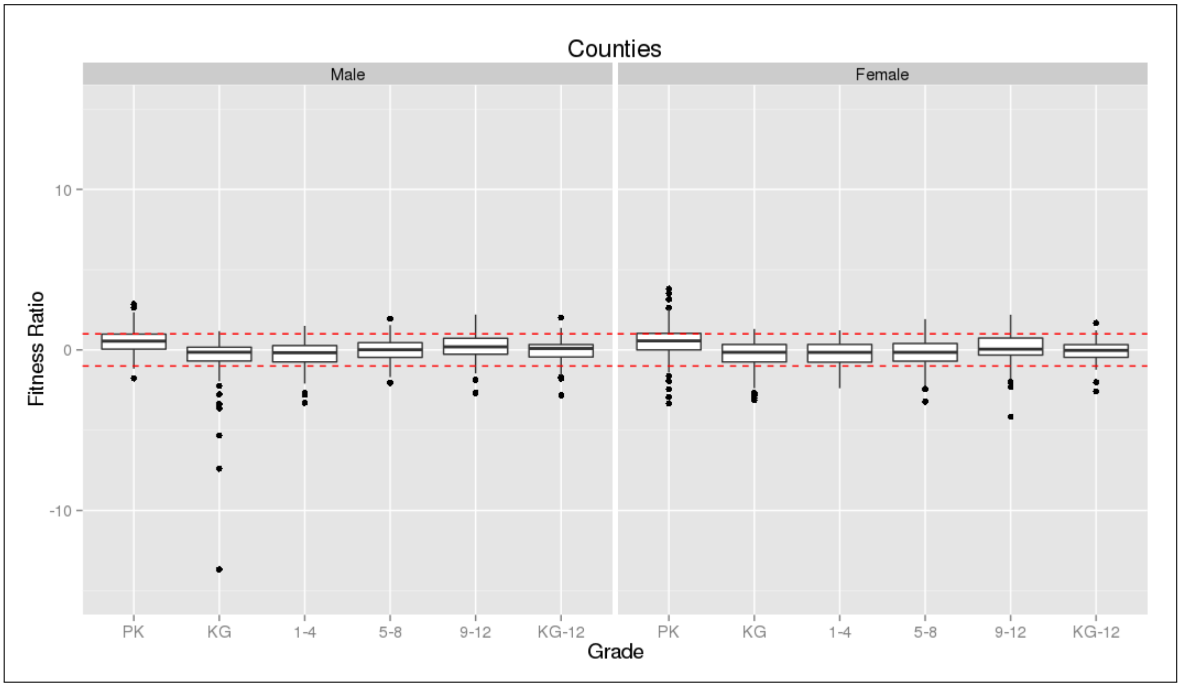
Table 8.5: ACS and Virginia SLDS State Enrollment Counts Comparison, by Gender and Grade Group, 2013 5-year estimate, 2009-2013

	ACS			Virginia		Comparison		
Variable	Estimate	90% MOE (+/-)	Dist.	Estimate	Dist.	FR	Difference	Within 90% MOE?
Total	1,277,023	6,570	100%	1,262,149	100%	2.3	14,874	NO
Male Pre-Kindergarten	28,590	1,296	2%	18,777	1%	7.6	9,813	NO
Male Kindergarten	47,489	1,717	4%	49,295	4%	-1.1	-1,806	NO
Male grades 1-4	187,534	2,403	15%	195,271	15%	-3.2	-7,737	NO
Male grades 5-8	193,213	2,819	15%	192,608	15%	0.2	605	YES
Male grades 9-12	200,040	1,983	16%	193,961	15%	3.1	6,079	NO
Female Pre-Kindergarten	25,744	1,124	2%	14,589	1%	9.9	11,155	NO
Female Kindergarten	45,251	1,407	4%	45,818	4%	-0.4	-567	YES
Female grades 1-4	180,805	2,318	14%	185,193	15%	-1.9	-4,388	NO
Female grades 5-8	177,270	2,726	14%	182,261	14%	-1.8	-4,991	NO
Female grades 9-12	191,087	2,191	15%	184,376	15%	3.1	6,711	NO

Note: "Dist." stands for Distribution; "FR" stands for *Fitness Ratio*; "Difference" refers to the difference between the ACS and SLDS estimates; MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error. Enrollment counts are for public schools. (**Sources:** SLDS Virginia data 2009-2013; ACS 2009-2013 5-year estimates).

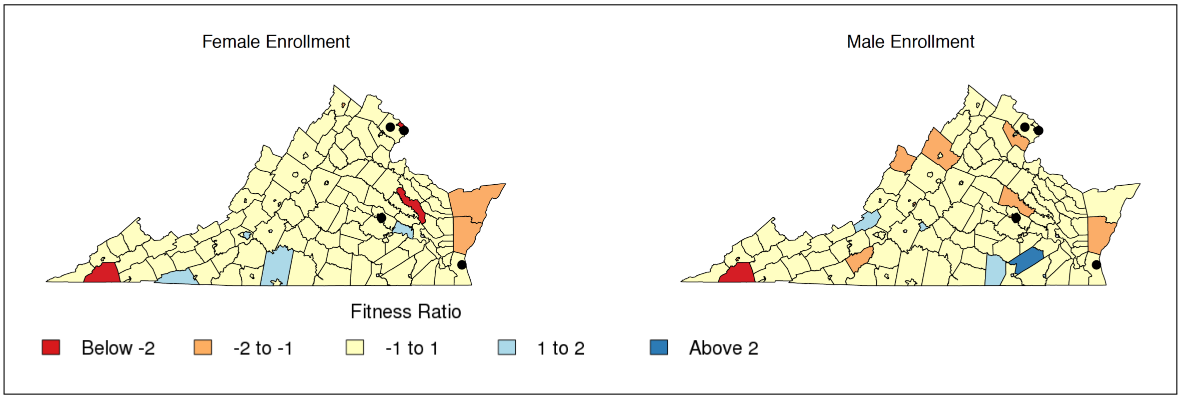
Figure 8.6 geographically displays the *Fitness Ratios*, showing there are many counties in which the ACS and VLDS enrollment counts for females and males are comparable and only a few counties that are not scattered throughout Virginia. Figures 8.7 and 8.8 present the data by four race/ethnicity groups. The ACS and VLDS enrollment counts for Black, Hispanic, and Asian students are similar. There are more differences in the enrollment counts for White students. This could be due to higher private school enrollment for White students. The ACS includes both private and public school enrollment estimates for race/ethnic groups, whereas the SLDS data only includes public school enrollment. These differences are particularly large in the Northern Virginia, Richmond, and Norfolk areas.

Figure 8.5: *Fitness Ratio* Scores for Virginia Counties by Gender and Grade



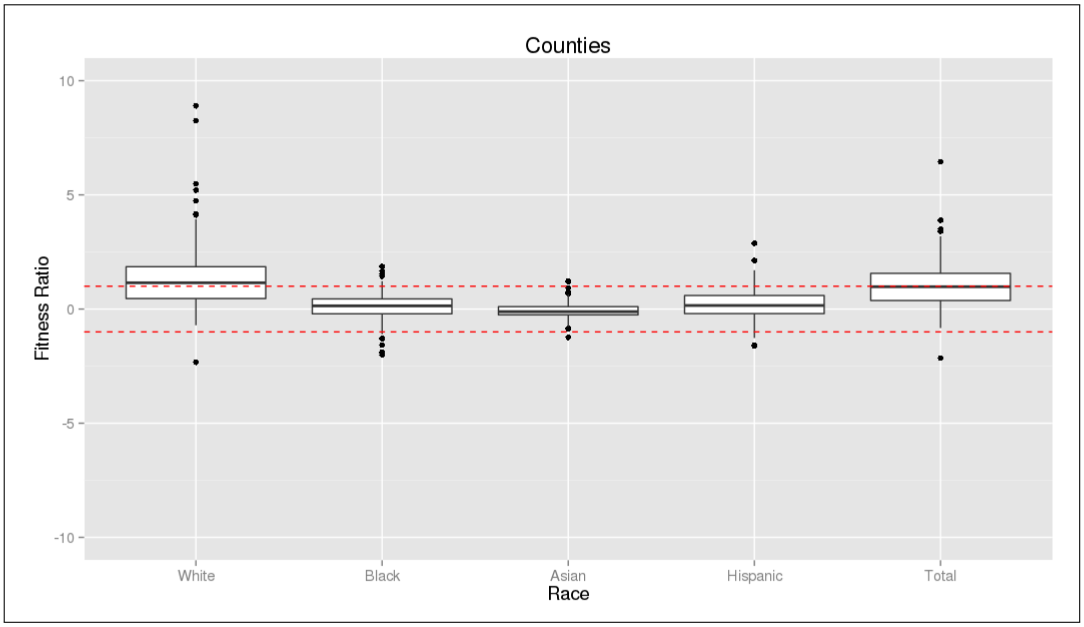
Boxplots compare the distribution of the *Fitness Ratios* at the county level and their relation to the ACS MOE. Estimates falling outside the red reference lines of ±1 are not within the 90% ACS margins of error. PK refers to pre-kindergarten and KG refers to kindergarten. Enrollment counts are for public schools. The KG-12 category represents the total enrollment counts across all grades. Total number of counties=129. Most of the counties are within the 90% ACS MOE, although there are several outliers for the Male Kindergarten group. (**Sources:** SLDS Virginia data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.6: *Fitness Ratios* across Virginia Counties for Females and Males



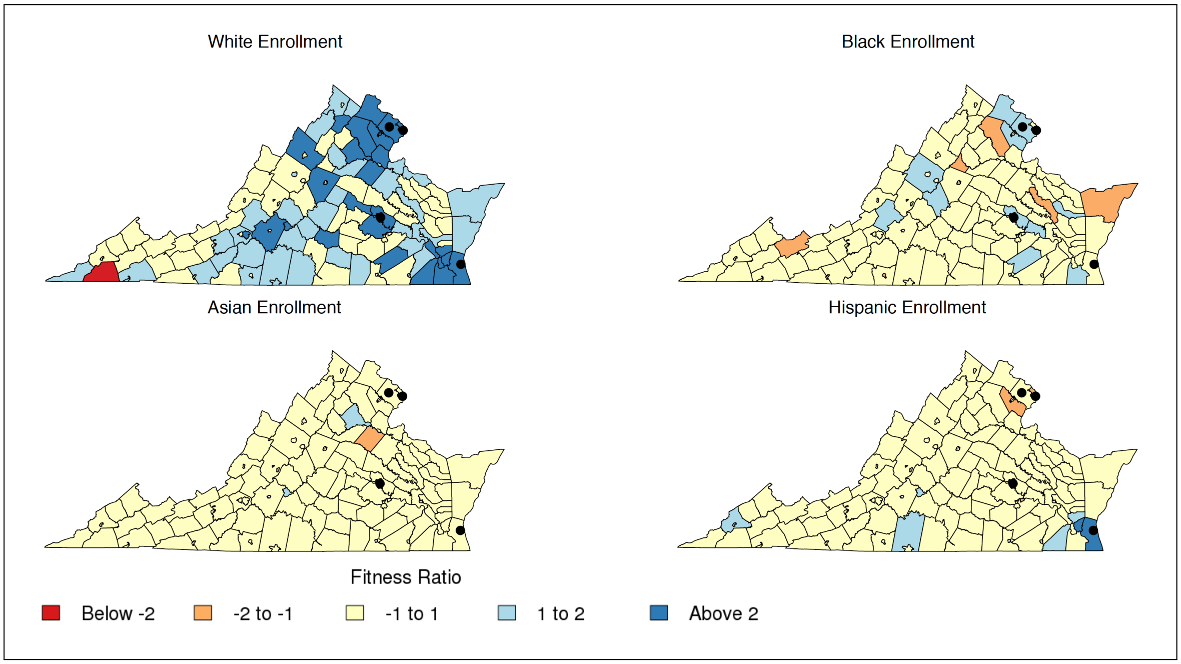
Enrollment only includes public school and does not include pre-kindergarten. Black points represent four cities in the urban area of Virginia (from north to south): Fairfax, Alexandria, Richmond, and Virginia Beach. Total number of counties=129. (**Sources:** SLDS Virginia data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.7: *Fitness Ratios* for Virginia Counties for Race/Ethnicity



Enrollment includes both private and public school for the ACS data and only includes public schools for the SLDS data. Enrollment does not include pre-kindergarten. Total number of counties=129. (**Sources:** SLDS Virginia data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.8: *Fitness Ratios* across Virginia Counties for Race/Ethnicity



Enrollment includes both private and public school for the ACS data and only includes public schools for the SLDS data. Enrollment does not include pre-kindergarten. Black points represent four cities in the urban area of Virginia (from north to south): Fairfax, Alexandria, Richmond, and Virginia Beach. Total number of counties=129. The majority of counties have *Fitness Ratios* within the 90% ACS MOE for Asian, Black, and Hispanic students (yellow shading); however ACS estimates tend to be higher than SLDS counts for white student enrollment (dark blue shading). (**Sources:** SLDS Virginia data 2009-2013; ACS 2009-2013 5-year estimates).

5. Benchmark Comparisons for Kentucky

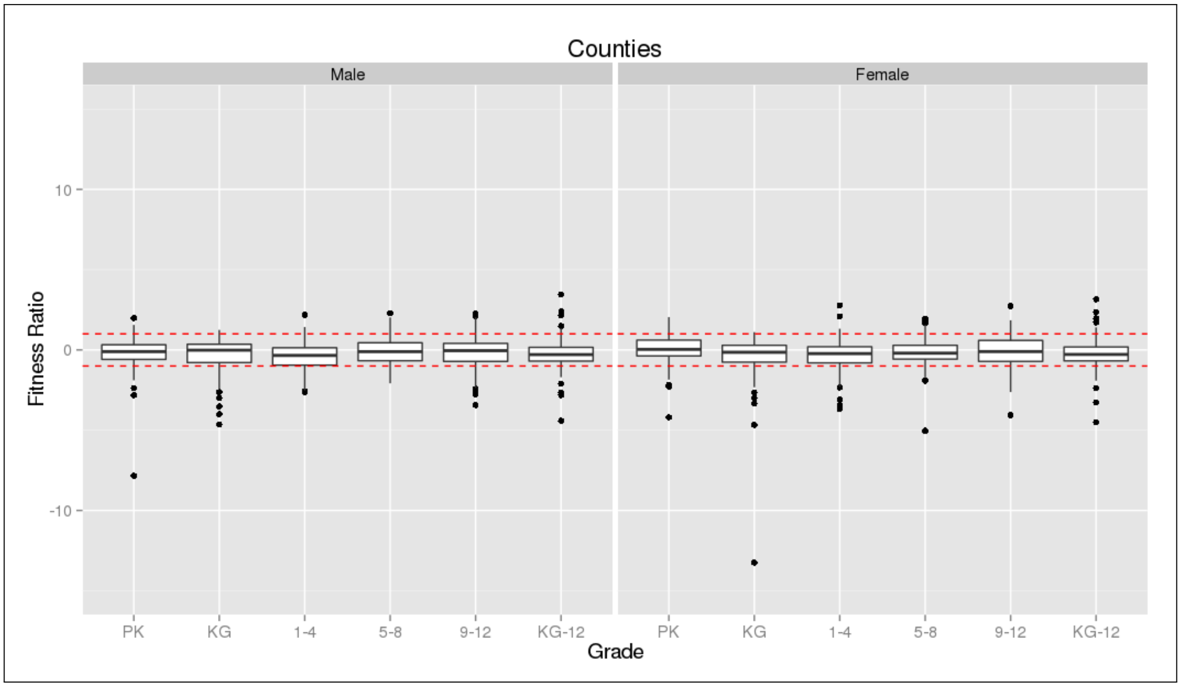
School districts align with counties or cities in the Kentucky SLDS (KLDS). The *Fitness Ratios* of enrollment counts between ACS and KLDS data given in Table 8.6 indicate the estimates are comparable, with some exceptions. The groups that do not fall within the 90% ACS MOE are female pre-kindergarten, males and females grades 1-4, and females grades 5-8. The maps in Figure 8.10 indicate most counties have similar enrollment count estimates as the ACS data and the Kentucky SLDS data.

Table 8.6: ACS and Kentucky SLDS State Enrollment Counts Comparison, by Gender and Grade Group, 2013 5-year estimate, 2009-2013

	ACS			North Carolina		Comparison		
Variable	Estimate	90% MOE (+/-)	Dist.	Estimate	Dist.	FR	Difference	Within 90% MOE?
Total	700,031	4,674	100%	711,152	100%	-2.4	-11,121	NO
Male Pre-Kindergarten	20,888	1,026	3%	20,413	3%	0.5	475	YES
Male Kindergarten	27,748	1,133	4%	27,998	4%	-0.2	- 250	YES
Male grades 1-4	101,270	1,727	14%	109,169	15%	-4.6	- 7,899	NO
Male grades 5-8	106,575	2,057	15%	105,570	15%	0.5	1,005	YES
Male grades 9-12	103,945	1,344	15%	104,400	15%	-0.3	- 455	YES
Female Pre-Kindergarten	19,567	1,050	3%	16,593	2%	2.8	2,974	NO
Female Kindergarten	25,199	1,072	4%	26,102	4%	-0.8	- 903	YES
Female grades 1-4	98,668	1,849	14%	102,909	14%	-2.3	- 4,241	NO
Female grades 5-8	97,665	1,660	14%	99,632	14%	-1.2	- 1,967	NO
Female grades 9-12	98,506	1,425	14%	98,366	14%	0.1	140	YES

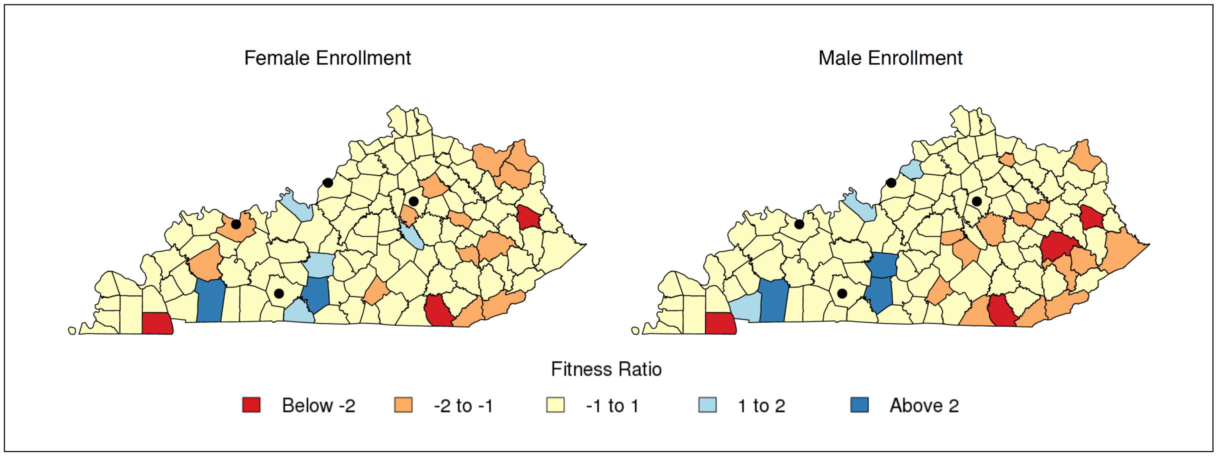
Note:“Dist.” stands for Distribution; “FR” stands for *Fitness Ratio*; “Difference” refers to the difference between the ACS and SLDS estimates; MOE is margin of error. Green highlights estimates that fall within 90% ACS margins of error. Enrollment counts are for public schools. (**Sources:** SLDS Kentucky data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.9: *Fitness Ratio* Scores for Kentucky Counties by Gender and Grade



Boxplots compare the distribution of the *Fitness Ratios* at the county level and their relation to the ACS MOE. Estimates falling outside the red reference lines of ±1 are not within the 90% ACS margins of error. PK refers to pre-kindergarten and KG refers to kindergarten. Enrollment counts are for public schools. The KG-12 category represents the total enrollment counts across all grades. Total number of counties=120. One extreme lower outlier (-20.67) is removed from the male kindergarten grade group in the boxplot. The majority of counties fall within the 90% ACS MOE, with the exception of several outliers which are for the pre-kindergarten and kindergarten groups. (**Sources:** SLDS Kentucky data 2009-2013; ACS 2009-2013 5-year estimates).

Figure 8.10: *Fitness Ratios* across Kentucky Counties for Females and Males



Enrollment counts are for public schools and do not include pre-kindergarten. Black points represent the four largest cities in Kentucky (from west to east): Owensboro, Bowling Green, Louisville, and Lexington. Total number of counties=120. The majority of counties are within the 90% ACS MOE, and those counties with higher *Fitness Ratios* tend to be in the southern and eastern parts of the state. (**Sources:** SLDS Kentucky data 2009-2013; ACS 2009-2013 5-year estimates).

D. Summary

The fitness ratios comparing the SLDS student enrollment counts to ACS estimates were generally within or close to the 90% ACS margins of error for Kentucky and Virginia. The enrollment estimates were more likely to be outside of the ACS 90% margins of error for North Carolina and Texas. When comparing the estimates by gender and grade clusters, the differences are larger for the earlier grades (PK, K, and 1-4) and smaller for the older grades (5-8 and 9-12). Mapping the fitness ratios by county shows that the SLDS enrollment estimates were similar to ACS estimates for many counties. There are reasons for differences between ACS and SLDS estimates as well as deficiencies in ACS data. The SLDS data could enhance ACS data products in providing geographic and longitudinal estimates.

9. Representative Education Use Cases

This chapter explores the use of statewide longitudinal data sources to enhance or possibly replace the American Community Survey (ACS) data for some education research applications. Two specific representative use cases were developed. The first described and modeled the number and characteristics of students between grades 3 and 12 in North Carolina identified as Limited English Proficient (LEP). The second described and modeled the number and characteristics of students that drop out of high school in Kentucky. The development of both representative use cases utilized three logistic regression models:

- Model 1: School District-Level Model Using American Community Survey (ACS) data
- Model 2: Student-Level Model Using Statewide Longitudinal Data System (SLDS) data
- Model 3: Combined Hierarchical Model Using ACS and SLDS data

Three motivations drove the application of logistic regression in these applications. First, the logistic regression is well-suited to handle the binary outcome variable (e.g., dropped out=1, did not dropout=0) found in both use cases. Second, the models include both categorical and numeric predictor variables, and logistic regression is the most common modeling approach for binomial data with a mix of categorical and numeric predictor variables. Third, the expected monotonic relationship between the outcome and the predictor variables fits the assumptions of logistic regression.

The models include a random effect variable for school district since we expect the modeled probabilities to vary by district. The motivation for inclusion of random effects is further explored for each of the representative use cases. Box 9.1 provides definitions of some of the common variables. Models 2 and 3 include both student-level and district-level predictor variables.

Box 9.1: Definitions of Variables Used in LEP and Dropout Models

Some of the variables in the models described in this chapter are defined in the same way. To reduce repetition within each model description, the ones that are the same are defined below.

Limited English Proficiency (LEP) Rate. The LEP rate is the proportion of students in grades 3-12 within a school district that were identified as limited English proficient out of total number of students enrolled in grades 3-12.

Dropout Rate. The dropout rate is the proportion of students within a school district that drop out of high school out of total number of students enrolled in grades 9-12. The percentage share of dropouts is the percentage of dropouts within each school district out of the total number of dropouts in the state of Kentucky.

Race. Race was defined as White, Black, Asian, and Other Race and did not include individuals who identified their ethnicity as Hispanic. In Models 2 and 3, White serves as the reference level for student race groups.

Ethnicity. The Hispanic category included all individuals who identified their ethnicity as Hispanic, regardless of their race. The Hispanic and Race categories are mutually exclusive, so those in the Hispanic category were not double counted in the race categories.

Poverty Percentages. The percentage of individuals under age 18 in poverty was calculated using the total number of individuals under the age of 18. This variable was calculated using ACS data.

Economically Disadvantaged. In North Carolina, students are economically disadvantaged if they receive free or reduced school lunch.

Grade. For the LEP models, 10th grade served as the reference level. For the high school dropout models, 9th grade served as the reference level.

School Districts. The school districts in the models are those assigned a Local Education Agency number.

A. Limited English Proficiency (LEP) Application in North Carolina

A major challenge facing the U.S. public school system is the increasingly large number of students identified as Limited English Proficient (LEP) (Chin et al. 2013). In the 2008-09 school year, approximately 1 in 9 students enrolled in pre-kindergarten to grade 12 across the U.S. were identified as LEP. In contrast, 1 in 13 students were classified as LEP a decade earlier (National Clearinghouse for English Language Acquisition (NCELA) 2011). Additionally, with over 400 languages represented (Kindler 2002), the diversity of languages spoken has also increased (Migration Policy Institute 2011). Studies have shown that LEP is a significant barrier to learning in U.S. schools (Chin et al. 2013) and can result in lower grades, lower standardized test scores, and higher dropout rates (Moss and Puma 1995, Bennici and Strang 2015, Ruiz-de Velasco and Fix 2000).

In addition to the lagging academic achievement of students identified as LEP (Christian 2006), studies have found that LEP is associated with potential issues in socioemotional development, including effects on learning, interpersonal skills, and the internalization or externalization of behaviors (Kang et al. 2014), as well as low health literacy, and subsequently, a higher frequency of poor health status (Sentell and Braun 2012). The challenges associated with LEP are of particular significance to North Carolina because the state experienced the second highest growth in the U.S. in the LEP population between 1990 and 2010, a 395% percent increase (Migration Policy Institute 2011). This provided motivation for focusing these LEP analyses on North Carolina.

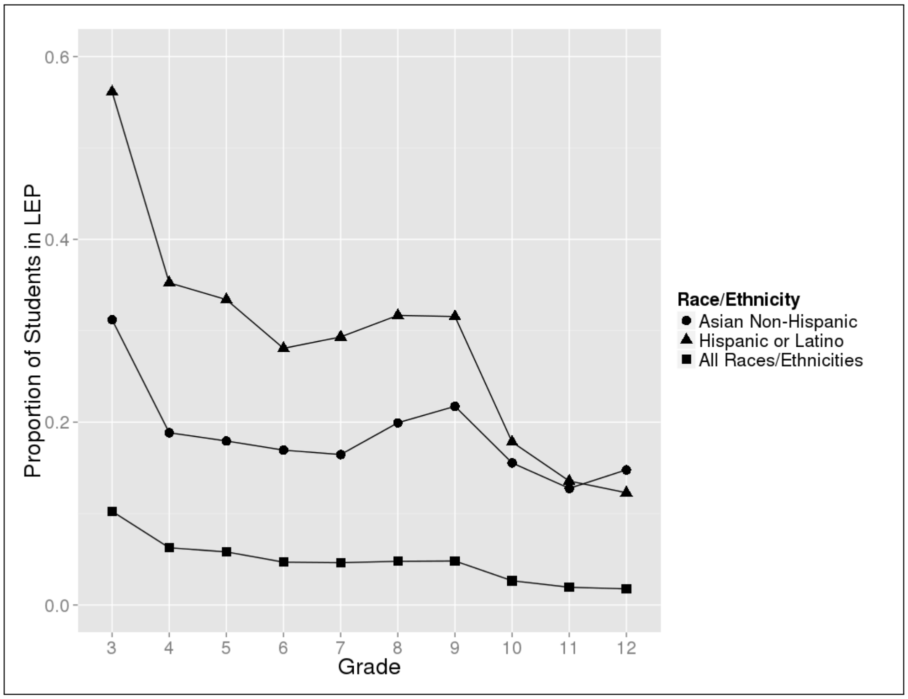
1. Selected descriptive results.

Descriptive statistics are presented using North Carolina SLDS data. Figure 9.1 shows the proportion of students in LEP in North Carolina in 2013 by race/ethnicity and grade. While Hispanic or Latino students generally had the highest proportion of students identified as LEP, Hispanic and Asian student proportions showed similar trends across grades, with the exception of 12th grade where the proportion of Asian students in LEP increased slightly from the previous grade.

In North Carolina, LEP students in grades 3 to 12 accounted for 6% of students in 2009-11 and 5% of students in 2012-13. The proportion of Asian students and Hispanic students who were identified as LEP declined between 2009 and 2013 (Asian students from 27% to 19% and Hispanic students from 48% to 31%). A slightly higher proportion of males than females were in LEP classes (5% versus 4% in 2013). The proportion of students in LEP classes declined with each grade (10% of 3rd graders and 2% of 12th graders were in LEP classes in 2013). This trend holds across demographic groups although there were still a large proportion of Asians and Hispanics in LEP classes in 12th grade (15% and 12% in 2013). Figure 9.2 and Figure 9.3 show the proportion of students who were identified as limited English proficient in grades 3 to 12 in 2013 by school district. Areas with 5% or more LEP students were more likely to be in school districts closer to one of the four major cities in North Carolina.

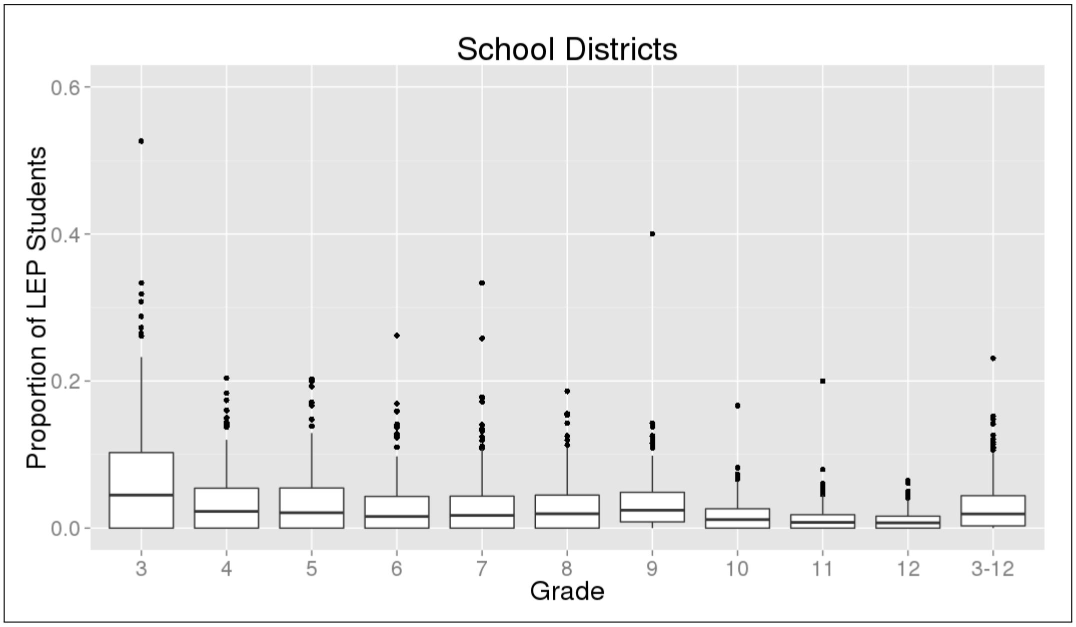
The North Carolina SLDS data in 2013 were compared to the ACS 2013 5-year estimates from the "Language Spoken at Home" tabulation. This ACS tabulation provides the count of individuals by age group and English proficiency when a language other than English is spoken at home. The data for individuals aged 5 to 17 who "speak English less than very well" were compared to the LEP enrollment counts in the North Carolina SLDS data.

Figure 9.1: Proportion of Students Identified as Limited English Proficient, North Carolina, by Race/Ethnicity and Grade, 2013



Source: North Carolina Statewide Longitudinal Data System (SLDS) administrative education data.

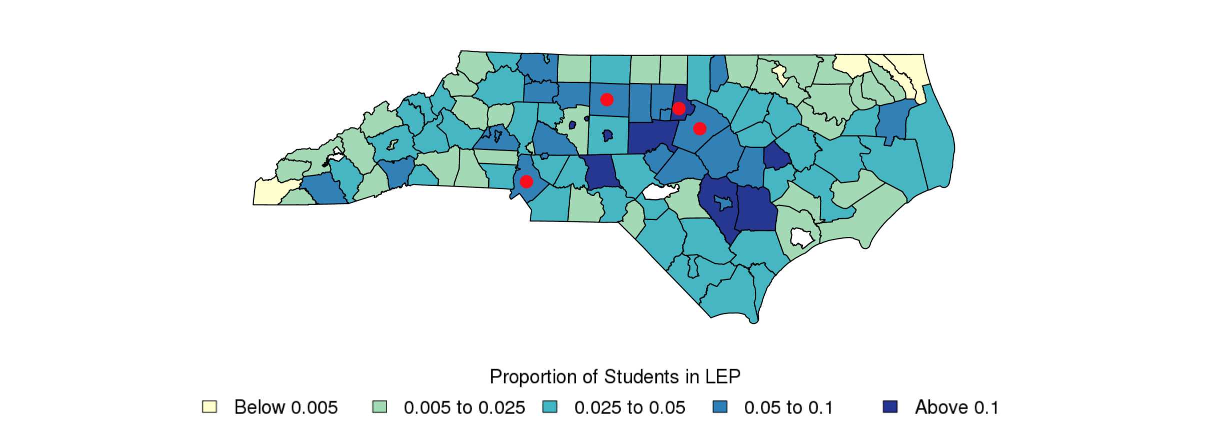
Figure 9.2: Proportion of Students Identified as Limited English Proficient in Grades 3 to 12, North Carolina School Districts, 2013



Boxplot compares the the distribution of the proportion of students in LEP by Grade at the district level. (**Source:** North Carolina Statewide Longitudinal Data System (SLDS) administrative education data.)

At the state-level, in 2013 there were 53,487 students identified as LEP as compared to 56,401 (±529) individuals age 5 to 17 that “speak English less than very well” according to 2013 5-year ACS estimates. Applying the *Fitness Ratios* described in Chapter 8, the corresponding *Fitness Ratio* is 5.5. At the school district-level, the *Fitness Ratios* for each school district are shown in Figure 9.4. Three of the four major cities are in school districts that have *Fitness Ratios* $> |2|$. These differences may be due to the way the data are collected. Judging how well a family member speaks English (based on the ACS question) is different from how the school assesses this (based on the SLDS data). The ACS data are based on where one lives which may or may not align with school districts.

Figure 9.3: Proportion of Students Identified as Limited English Proficient in Grades 3 to 12 across North Carolina School Districts, 2013



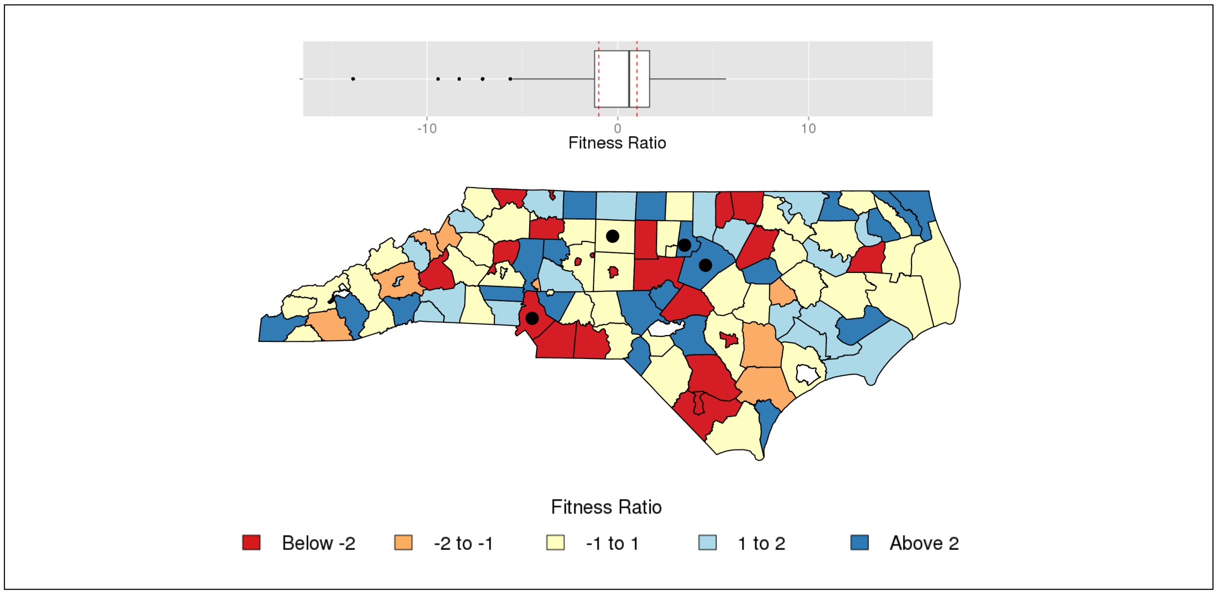
The black dots represent the following major cities: Charlotte, Greensboro, Durham, and Raleigh (moving west to east). (**Source:** North Carolina Statewide Longitudinal Data System (SLDS) administrative education data.)

2. Model Description

This application builds on a study conducted by Ramani and Noel (2014) that compared the Department of Education Common Core of Data (CCD) estimates of English Language Learner (ELL) students and ACS estimates of individuals who “speak English less than very well” at the state-level. While CCD data had limited capability to link to ACS estimates at the school district level, the state education administrative records for North Carolina provide LEP enrollment data at the student level that can be aggregated up to the school district, county, and/or state, thus providing a greater degree of flexibility.

The analyses combined state education administrative records for North Carolina and ACS data to better understand the socioeconomic and demographic factors at the student and school district levels that are associated with LEP. The ACS provides data on the number of individuals

Figure 9.4: *Fitness Ratios* for Students Identified as Limited English Proficient, North Carolina School Districts, 2013



Boxplot compares the distribution of the *Fitness Ratios* at the district level. Estimates falling outside the red reference lines of ±1 are not within the 90% ACS margins of error. Three extreme lower outliers (-68.5, -69, and -80) are removed in the boxplot. The black points on the map are Charlotte, Greensboro, Durham, and Raleigh. (**Sources:** American Community Survey (ACS) 2013 5-year estimates, North Carolina Statewide Longitudinal Data System (SLDS) administrative education data.)

that "speak English very well" and "less than very well" while the state education data provides an LEP indicator for each student.

Three models were developed to explore predictors of LEP using logistic regressions. The models used data from the ACS 2013 5-year estimates and the 2013 North Carolina administrative records. The nature of the data from these sources placed certain constraints on the data that could be used for each of the analyses. Five-year estimates were used in order to have more coverage at the school district level for the ACS data. Data for the population in ages 5 and over were used because the ACS "Language Spoken at Home" tabulation does not provide the count of individuals who "speak English very well" by more granular age ranges. Data for grades 3 through 12 were used because the student enrollment data in North Carolina for pre-kindergarten through grade 2 are not complete.

A random effects factor for the school district was included in each model. To explore the need for the random effect, a simple logistic regression was fit to several of the district-level predictor variables (e.g., percent Hispanic, percent Asian). As an example, Figure 9.5 plots the logit proportion of LEP, $\log(\frac{\hat{p}_k}{1-\hat{p}_k})$ where $\hat{p}_k$ is the recorded proportion of students enrolled in LEP within school district k, by the percent Hispanic within the school district. A logistic regression was fit (given the the line).

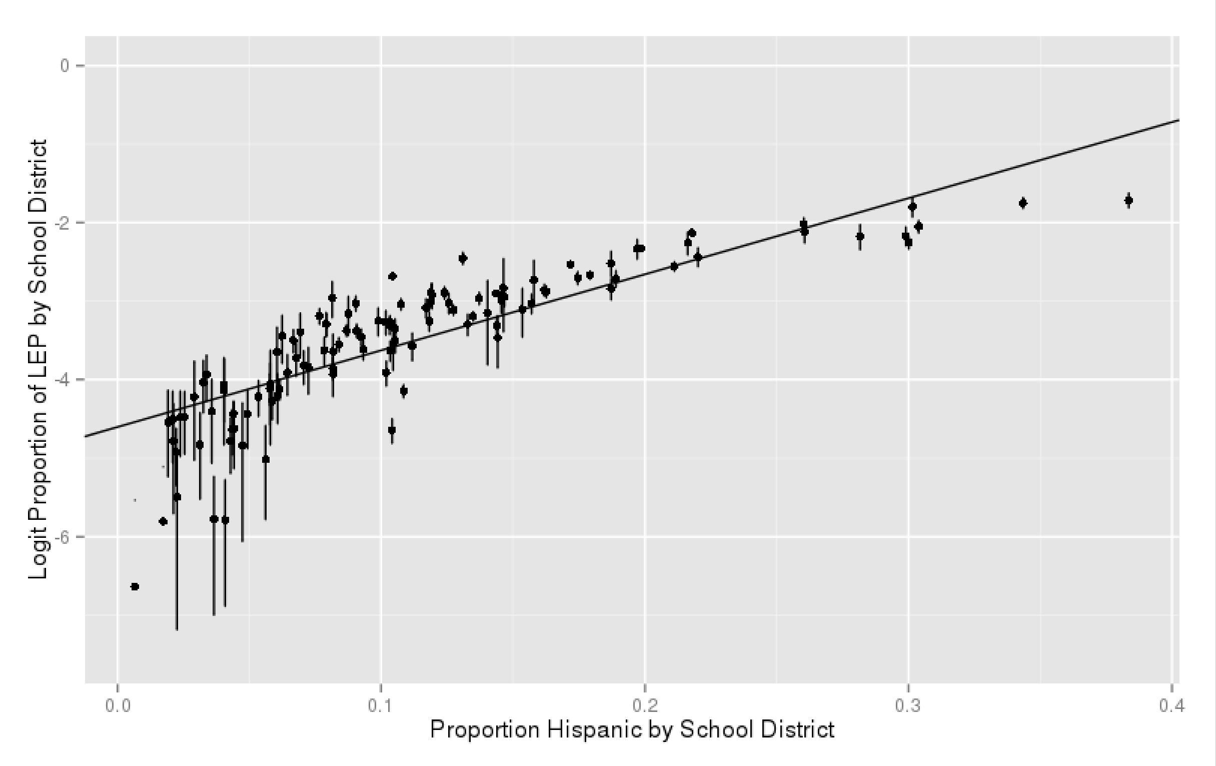
Simple univariate logistic regression models the recorded number of LEP cases y_k from a sample of size n_k within a school district k as

$$\begin{aligned} y_k &\sim \text{Binomial}(n_k, p_k), \text{ with} \\ logit(p_k) &= \alpha + x_k\beta \end{aligned}$$

where x_k is the percentage of Hispanic students within school district k. Estimates for α and β are obtained via maximum likelihood, producing the line in Figure 9.5. The binomial model dictates the uncertainty in the fitted model since $\hat{p}_k = y_k/n_k$ has a standard error of $\sqrt{\hat{p}_k(1-\hat{p}_k)/n_k}$.

For each district, the estimate and uncertainty ($\hat{p}_k \pm 2 \cdot \text{se}$) is given as well. If the basic simple logistic regression model were adequate here, the logistic regression line should be within a couple of se's of each $\hat{p}_k$. This is clearly not this case since a substantial number of $\hat{p}_k$'s are many more se's away from the fitted line. Thus this model requires the inclusion of a school district random effect in the model to account for this extra binomial variation. Although the estimated logistic regression line does not pass neatly through all the data points, the plot does show that a linear model on the logit scale is reasonable, supporting the use of a logistic regression in this case. A similar result was true for the percent Asian data.

Figure 9.5: Logit Proportion of Students Identified as Limited English Proficient by Percentage of Hispanic Students in North Carolina School Districts in 2013



Each point in the plot represents a school district. The estimated logistic regression line shown in the figure does not pass through all the data points, which means an extra error term is required. This is given by the error bars around each point. (**Source:** North Carolina Statewide Longitudinal Data System (SLDS) administrative education data.)

3. Models

Model 1: Predicting LEP at the School District Level Using ACS Data

Model 1 uses ACS data and is a mixed effects logistic regression, with data collected for each of the $N = 115$ school disticts: $y = (y_1, ..., y_N)$ holds the number of LEP cases for each school district; $n = (n_1, \ldots, n_N)$ holds the total number in the sample for each school district. Using vector-wise notation we have

$$\begin{aligned} y &\sim \text{Binomial}(n, p), \text{ with} \\ \text{logit}(p) &= \alpha + X_{D1}\beta_{D1} + X_{D2}\beta_{D2} + X_{D3}\beta_{D3} + X_{D4}\beta_{D4} + X_{D5}\beta_{D5} + z, \end{aligned}$$

where:

- $p = (p_1, \ldots, p_N)$ is the modeled probability that an individual “speaks English less than very well” for each school district.

- X_{D1} - X_{D5} are design matrices for the district-level fixed effects predictor variables: percent Black, percent Hispanic, percent Asian, percent Other Race, and percent under 18 in poverty
- $z = (z_1, \ldots, z_N)$ holds the random effect for each school district, modeled as mean zero normal effects $z \sim N(0, \sigma_z^2 I_N)$.

The number of cases y and the enrollment counts n, as well as all of the predictor variables, were from the ACS data at the school district level. Both private and public school were used in this case because the ACS tables did not differentiate enrollment counts for race/ethnicity by type of school. This model was fit to 115 observations, one for each school district, and 6 variables.

Model results

School districts were identified by their Local Education Agency (LEA) numbers. This was used to link school districts across the ASC and NC SLDS and to compute the proportions and percentages at the school district level. School district was included in the model as a random effects variable to account for extra binomial variation and to account for the variation in LEP enrollment between school districts. When models were fit with and without this variable, the addition of school districts as a random effects variable significantly improved the fit of the model. As shown in Table 1.1, the deviance for the model with the random effects variable is significantly lower (chi-square=18,805; p<0.001). These results provide further evidence for the need to include random effects in the model.

Table 9.1: Model 1, Logistic Regression ANOVA Results - With and Without Random Effects, American Community Survey (2013 5-year estimate)

Model	DF	Deviance	Chisq	Chi DF	Significance
Model 1 without Random Effect	7	20545.1			
Model 1 with Random Effect	8	1738.7	18806	1	***

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Source:** American Community Survey 2013 5-year estimate.)

Table 9.2 presents the full model results. The percentage of Hispanic and Asian students in a school district were significant predictors of students that “speak English less than very well”. School districts were more likely to have greater number of LEP students if they had a higher percentage of Hispanic students and Asian students and a lower percentage of individuals under 18 in poverty.

Table 9.2: Model 1. Logistic Regression Results for Predicting Limited English Proficiency, North Carolina Students, American Community Survey, 2009-2013

Predictor	Coefficient	Std. Error	Significance
Intercept	-4.1943	0.1455	***
Percent Black Non-Hispanic	0.4201	0.2329	
Percent Hispanic	8.4988	0.5319	***
Percent Asian Non-Hispanic	7.1347	1.9665	***
Percent Other Race Non-Hispanic	1.2785	0.6890	
Percent Under 18 in Poverty	-1.4764	0.5117	**

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Source:** American Community Survey 2009-2013 5-year estimates).

Model 2. Predicting Limited English Proficiency at the Student Level Using North Carolina SLDS Data

Model 2 uses North Carolina SLDS data and is a mixed effects logistic regression, with data collected for $N = 1,081,659$ students: $y = (y_1, ..., y_N)$ holds a 0 or 1, depending on whether or not the student is LEP. Using vector-wise notation we have

$$\begin{aligned} y &\sim \text{Bernoulli}(p), \text{ with} \\ \text{logit}(p) &= \alpha + X_{S1}\beta_{S1} + X_{S2}\beta_{S2} + X_{S3}\beta_{S3} + X_{S4}\beta_{S4} + X_{D1}\beta_{D1} + X_{D2}\beta_{D2} \\ &\quad + X_{D3}\beta_{D3} + X_{D4}\beta_{D4} + X_{D5}\beta_{D5} + X_{D6}\beta_{D6} + Uz, \end{aligned}$$

where:

- X_{S1} - X_{S4} are design matrices holding student-level fixed effects predictor variables: race/ethnicity, gender, grade, and economically disadvantaged status
- X_{D1} - X_{D6} are design matrices holding district-level fixed effects predictor variables: enrollment counts, percent Black, percent Hispanic, percent Asian, percent Other Race, and percent economically disadvantaged
- U is a design matrix assigning students to school districts, and
- $z = (z_1, \dots, z_K)$ is the random effect for each of the $K = 115$ school districts, modeled as mean zero normal effects $z \sim N(0, \sigma_z^2 I_K)$

All variables in this model were fit to the North Carolina SLDS data. The outcome variable in this model is a binary indicator for LEP status at the student level (1=current LEP student, 0=not in LEP). As in the previous model, school districts were identified by their LEA numbers and were included in the model as random effects. This model was fit to 1,081,659 observations, one for each student, and 11 variables.

Model results

Table 9.3 presents the model results. Predictors that were significant at the student level include gender, race, grade, and economically disadvantaged, while at the district-level they include percent Asian, percent Hispanic, percent Other Race, and percent economically disadvantaged. Students were more likely to be identified as LEP if they were male and Hispanic, Asian, or some Other Race (as compared to White), with Hispanic and Asian having the largest effects.

Students were more likely to be identified as LEP if they were in grades 3 to 9, as compared to grade 10. School districts having higher percentages of Hispanics, Asians, and economically disadvantaged were predictors of a larger proportion of LEP students. Students were less likely to be identified as LEP if they were in higher grades, e.g., grades 11 to 12, and were in school districts with a higher percentage of Other Race.

The coefficients for economically disadvantaged were positive in this model, an opposite effect from the percentage under 18 in poverty in the first model. This finding was not surprising since this model adjusted for covariates at the individual student-level, one of which is a measure of poverty. In addition, definitions of the two poverty variables differ between ACS and the state data. In ACS, individuals are in poverty if they are in families who fall under the federal poverty line. In the North Carolina data, students are economically disadvantaged if they receive free or reduced school lunch. (The criteria for receiving free or reduced school lunch are defined by the U.S. Department of Agriculture). The mean percentage of individuals under 18 in poverty is 29.5%, while the mean percentage of students who are economically disadvantaged is substantially higher, at 52.4%.

Table 9.3: Model 2. Logistic Regression Results for Predicting Students Identified as Limited English Proficient in North Carolina, Grades 3-12, NC SLDS data, 2013

Predictor	Coefficient	Std. Error	Significance
Intercept	-8.128	0.08528	***
Student level:			
Asian Non-Hispanic	4.237	0.02874	***
Hispanic	4.376	0.02423	***
Black Non-Hispanic	-0.0641	0.03454	
Other Race Non-Hispanic	0.9043	0.04518	***
Male	0.3097	0.01074	***
Economically disadvantaged	1.39	0.01701	***
Grade 3	1.595	0.02405	***
Grade 4	0.7449	0.02452	***
Grade 5	0.6661	0.02473	***
Grade 6	0.4301	0.02535	***
Grade 7	0.4923	0.02549	***
Grade 8	0.6269	0.02556	***
Grade 9	0.6902	0.02521	***
Grade 11	-0.2633	0.03155	***
Grade 12	-0.2811	0.03344	***
District level:			
Enrollment Count (in 10,000s)	0.01394	0.01976	
Percent Asian Non-Hispanic	4.845	0.168	***
Percent Hispanic	1.392	0.1727	***
Percent Black Non-Hispanic	-0.1585	0.1328	
Percent Other Race Non-Hispanic	-0.8668	0.2196	***
Percent Economically Disadvantaged	0.6886	0.1585	***

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Sources:** American Community Survey (ACS) 2013 5-year estimates, North Carolina Statewide Longitudinal Data System (SLDS) administrative education data.)

Model 3. Predicting Limited English Proficiency at the Student Level Using ACS and North Carolina SLDS Data

Model 3 outcome and predictor variables are the same as for Model 2, with the addition of the ACS predictor variable "speaks English less than very well", measured at the school district level. The percent of foreign-born from the ACS tabulation "Selected Characteristics of the Native and Foreign-born Populations" was also considered, but in the 2013 5-year estimate, the data were only available for 20 school districts (out of 115 North Carolina school districts).

Model 3 was fit to 1,081,659 observations, one for each student. Given the additional ACS predictor variable, this model has 12 variables. Data were collected for $N = 1,081,659$ students: $y = (y_1, ..., y_N)$ holds a 0 or 1, depending on whether or not the student is LEP. Using vector-wise notation we have

$$\begin{aligned} y &\sim \text{Bernoulli}(p), \text{ with} \\ \text{logit}(p) &= \alpha + X_{S1}\beta_{S1} + X_{S2}\beta_{S2} + X_{S3}\beta_{S3} + X_{S4}\beta_{S4} + X_{D1}\beta_{D1} + X_{D2}\beta_{D2} \\ &\quad + X_{D3}\beta_{D3} + X_{D4}\beta_{D4} + X_{D5}\beta_{D5} + X_{D6}\beta_{D6} + X_{D7}\beta_{D7} + Uz, \end{aligned}$$

where:

- X_{S1} - X_{S4} are design matrices holding student-level fixed effects predictor variables: race/ethnicity, gender, grade, and economically disadvantaged status
- X_{D1} - X_{D7} are design matrices holding district-level fixed effects predictor variables: enrollment counts, percent Black, percent Hispanic, percent Asian, percent Other Race, and percent economically disadvantaged
- U is a design matrix assigning students to school districts, and
- $z = (z_1, \ldots, z_K)$ is the random effect for each of the $K = 115$ school districts, modeled as mean zero normal effects $z \sim N(0, \sigma_z^2 I_K)$

Model Results

Since the ACS variable was at the district level, it had little effect on the student level variables. All of the significant predictor variables in Model 2 are also significant in Model 3 and their interpretation remains the same. In addition, the ACS variable "speaks English less than very well" was a significant predictor. Even though the approximate logistic regression significance test in Model 3 shows significance for the ACS variable "speaks English less than very well", a more reliable logistic regression deviance test showed no significant differences between the two models (Chi-square=1.4, p=0.24).

Table 9.4: Model 2 and Model 3-Logistic Regression ANOVA Results, Grades 3-12, American Community Survey (2013 5-year estimate) and North Carolina SLDS Data, 2013

Model	DF	Deviance	Chisq	Chi DF	Significance
Model 2	24	230812			
Model 3	25	230811	1.4026	1	

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Sources:** American Community Survey (ACS) 2013 5-year estimates, North Carolina Statewide Longitudinal Data System (SLDS) administrative education data.)

B. High School Dropout Rates in Kentucky

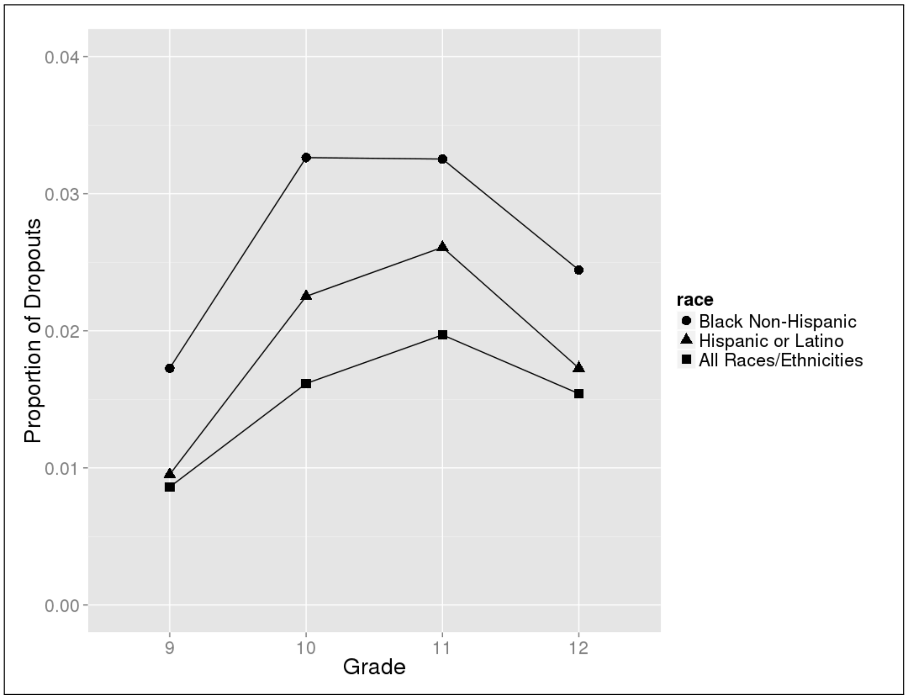
This application focused on students who drop out of high school. While the ACS collects data on student enrollment and education attainment, the ACS does not separately identify students who drop out of high school. This information is included in state education administrative records. Identifying students who drop out is important as students who fail to graduate have significantly worse life outcomes compared to those who graduate, including higher unemployment, lower earnings, greater risk of incarceration, and lower life expectancy (Jemal et al. 2008, Moretti 2007, Muenning 2007, Rouse 2007, Swanson 2009, Waldfogel et al. 2007). Understanding the factors that predict the characteristics of students that drop out can help inform policy and prevention efforts in schools. Kentucky, for instance, recently passed a new compulsory attendance policy that raised the compulsory school attendance age from 16 to 18 in order to address the number of students dropping out of high school. This policy will take effect for most districts in Kentucky beginning with the 2015-16 school year. The attention Kentucky is paying to dropout as evidenced by the new policy provides motivation to focus this analysis on the socioeconomic and demographic factors associated with dropout rates in Kentucky. The analyses provided here for 2009-2013 data provides a baseline to measure the change in policy that is occurring in 2015-2016.

1. Selected Descriptive Results

Descriptive statistics are presented using Kentucky SLDS data. Figure 9.6 shows the proportion of students who dropped out in 2013 by race/ethnicity and grade. While Black students generally had the highest proportion of dropouts, Hispanic or Latino students showed a similar trend.

In Kentucky, 1.5% of students in grades 9 to 12 dropped out of school in 2013. Slightly more males than females drop out of school (1.7% versus 1.3%) and the dropout rate was higher for Black (non-Hispanic) (2.6%), Hispanic (1.8%), and students of other minority races (1.9%), such as Hawaiian and American Indian. The dropout rate increased with each grade (0.9% of 9th graders, 1.6% of 10th graders, and 2.0% of 11th graders), but then declined for 12th graders (1.5%). This holds across demographic groups except for Asian students who had the highest dropout rate in grades 9 (1.3%) and 11 (1.1%). Figure 9.7 and Figure 9.8 show the proportion of students in grades 9 to 12 who dropped out of high school.

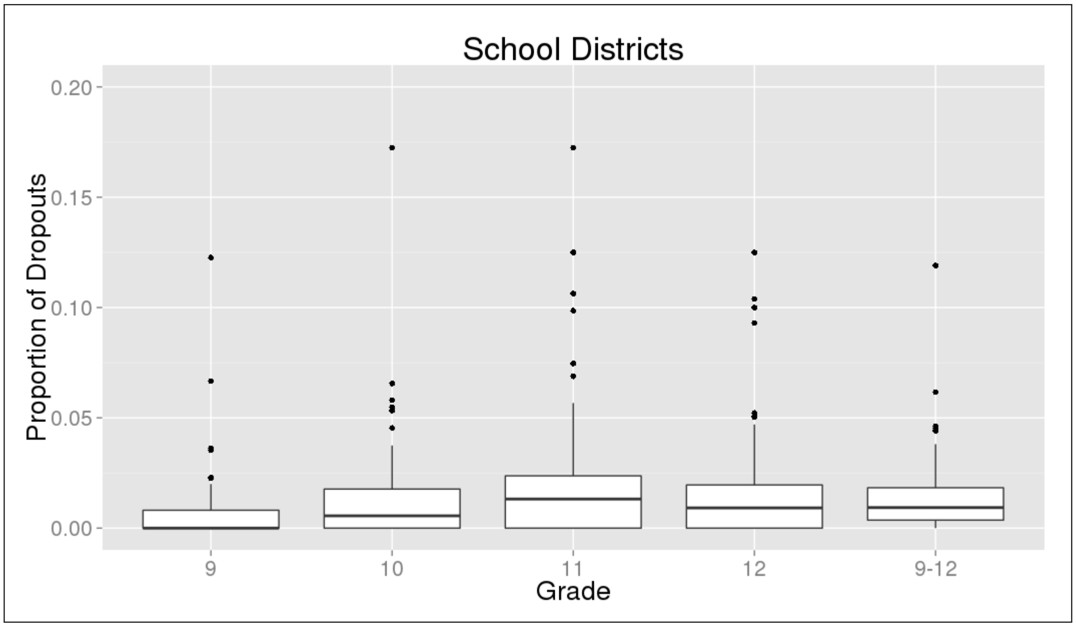
Figure 9.6: Proportion of Students who Dropped Out of High School by Race/Ethnicity and in Grades 9-12, Kentucky, 2013



Source: Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.

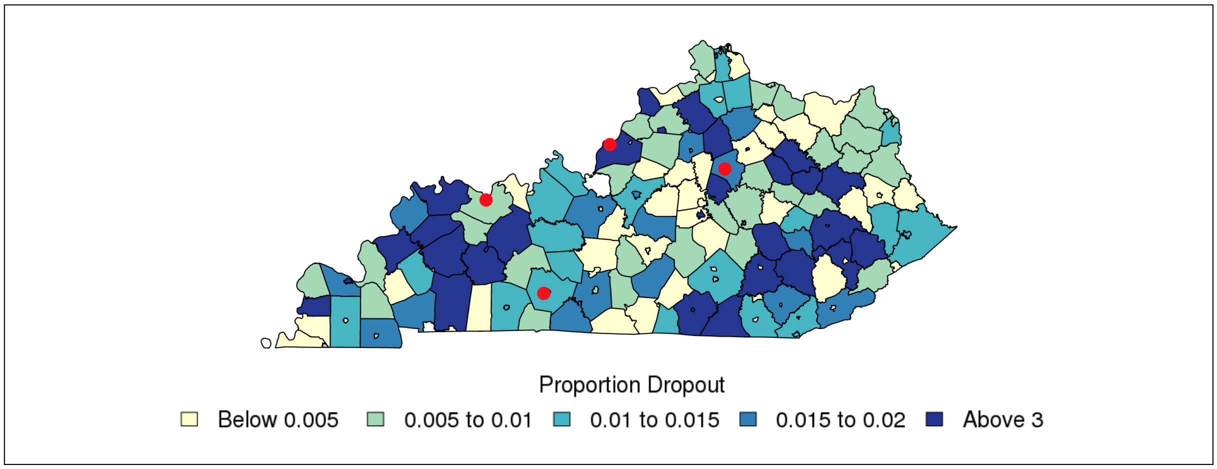
Figure 9.7 shows the proportion of the students within a school district who dropped out of high school out of the total number of students who dropped out in the state of Kentucky in grades 9-12. While there does not appear to be a pattern to the dropout rates by school district, the map in Figure 9.9 indicates that among the students who drop out, the largest percentage (above 10%) are located in Jefferson County School District where Louisville is located and which has the highest student enrollment.

Figure 9.7: Proportion of Students who Dropped Out of High School, Grades 9-12, Kentucky School Districts, 2013



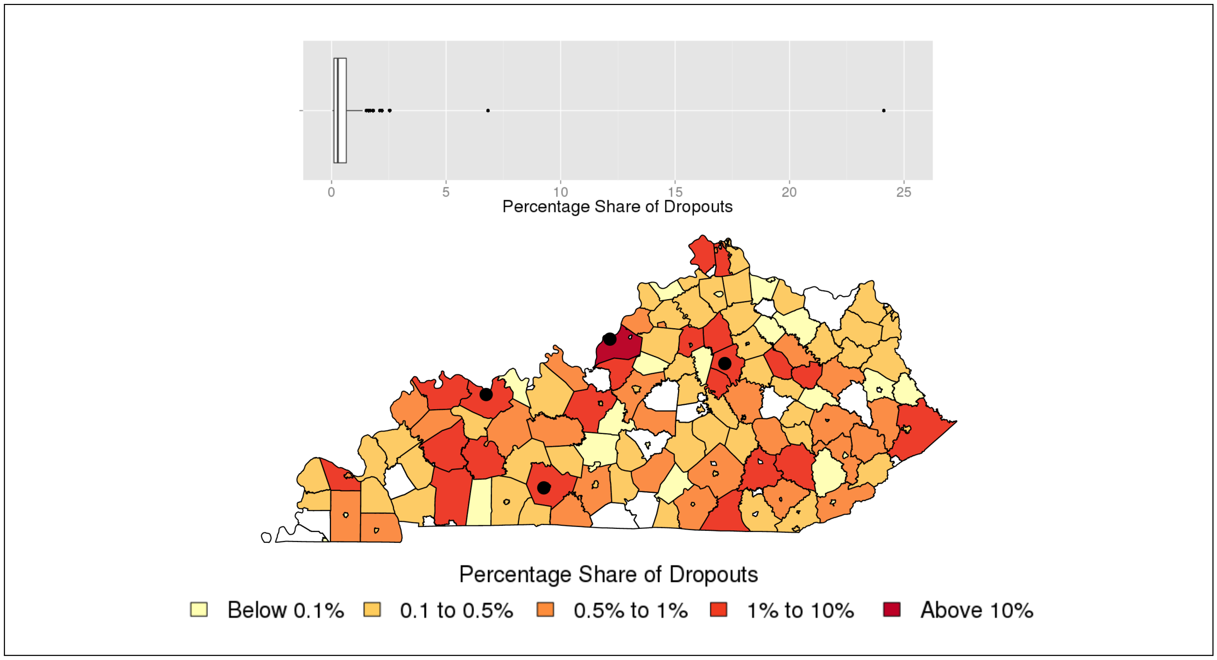
Boxplot compares the the distribution of the proportion of students who dropped out by grade at the district level. (**Source:** Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

Figure 9.8: Proportion of Students who Dropped Out of High School, Kentucky, 2013



The red dots are Owensboro, Bowling Green, Louisville, and Lexington (moving west to east). (**Source:** Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

Figure 9.9: Percentage of the Total Number of Students who Dropped Out of High School across Kentucky School Districts, Grades 9-12, Kentucky, 2013



Note: that the red dots are Owensboro, Bowling Green, Louisville, and Lexington (moving west to east). The school districts and counties that are white did not have any students drop out in 2013. (**Source:** Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

2. Model description

Previous studies have investigated the demographic characteristics of dropouts (Davis and Bauman 2011) and the predictors of dropout (Angrist and Krueger 1991, **?**). This application goes beyond the current literature by exploring predictors of dropouts using different data sources (ACS and SLDS) and accounting for the variation in dropouts within school districts. For this analysis, state education administrative records for Kentucky and ACS data were used to better understand what socioeconomic and demographic factors are associated with students who drop out. In this representative use case, "dropout" is defined as a student who has an indicator in the Kentucky SLDS data that he or she dropped out of public high school. This information was then used to calculate the dropout rate, which is defined as the number of students who dropped out of public high schools divided by the total number of students enrolled in public schools for each school district. Note that the dropout rate would exclude student-age individuals who moved to Kentucky after dropping out of their previous high school.

Three models were developed to explore predictors of dropout using logistic regressions. The models were fit to data from the ACS 2013 5-year estimates and the 2013 Kentucky SLDS data. Five-year estimates were used in order to have more coverage at the school district level for the ACS data. Only students in grades 9 through 12 were included in the analysis and all state data were from the Kentucky SLDS.

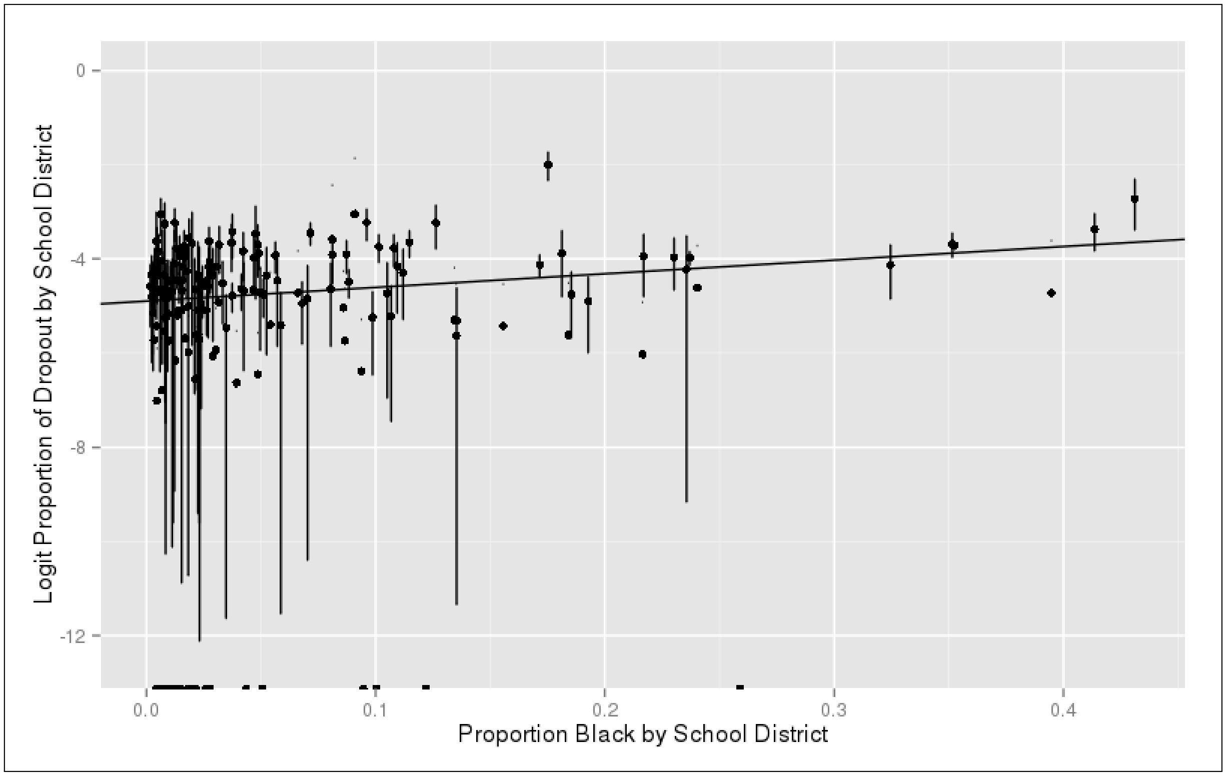
Similar to the LEP application, a random effects variables for the school district were included in each model. To explore the need for this, a simple logistic regression was fit to several of the district-level predictor variables (e.g., percent Black). The simple univariate logistic regression models the recorded number of dropouts y_k from a sample of size n_k within a school district k as

$$\begin{aligned} y_k &\sim \text{Binomial}(n_k, p_k), \text{ with} \\ logit(p_k) &= \alpha + x_k\beta \end{aligned}$$

where x_k is the percentage of Black students within school district k. Estimates for α and β are obtained via maximum likelihood, producing the line in Figure 9.10. The binomial model dictates the uncertainty in the fitted model since $\hat{p}_k = y_k/n_k$ has a standard error of $\sqrt{\hat{p}_k(1-\hat{p}_k)/n_k}$.

For each school district, the estimate and uncertainty ($\hat{p}_k \pm 2 \cdot \text{se}$) is given as well. If the basic simple logistic regression model were adequate here, the logistic regression line should be within a couple of standard error's of each $\hat{p}_k$. This is clearly not this case since a substantial number of $\hat{p}_k$'s are many more standard error's away from the fitted line. Thus this model requires the inclusion of a school district random effect in the model to account for this *extra* *binomial variation*. Although the estimated logistic regression line does not pass neatly through all the data points, the plot does show that a linear model on the logit scale is reasonable, supporting the use of a logistic regression in this case.

Figure 9.10: Logit Proportion of Students who Dropped out of High School by Percentage of Black Non-Hispanic Students, Kentucky School Districts, 2013



Each point in the plot represents a school district. The estimated logistic regression line shown in the figure does not pass through all the data points means an extra error term is required which is given by the error bars around each point. The points below -12.5 represent school districts that had no students dropout in 2013. (**Source:** Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

3. Models

Model 1. Predicting High School Dropout Rates at the School District Level using ACS Data and SLDS for the Dropout Indicator for Kentucky

Model 1 is a mixed effects logistic regression fit at the school district level. Data were collected for each of the $N = 168$ school districts: $y = (y_1, ..., y_N)$ holds the number of dropout cases for each school district; $n = (n_1, \ldots, n_N)$ holds the total number in the sample for each school district. We take $p = (p_1, \ldots, p_N)$ to denote the probability of dropout within each school district. Using vector-wise notation we have

$$\begin{aligned} y &\sim \text{Binomial}(n,p), \text{ with} \\ \text{logit}(p) &= \alpha + X_{D1}\beta_{D1} + X_{D2}\beta_{D2} + X_{D3}\beta_{D3} + X_{D4}\beta_{D4} + X_{D5}\beta_{D5} + z, \end{aligned}$$

where:

- p is the modeled probability an individual drops out,
- X_{D1} - X_{D5} are design matrices for the district-level fixed effects predictor variables: percent Black, percent Hispanic, percent Asian, percent Other Race, and percent under 18 in poverty
- $z = (z_1, \ldots, z_N)$ holds the random effect for each school district, modeled as mean zero normal effects $z \sim N(0, \sigma_z^2 I_N)$.

This logistic regression model has all of the covariates at the school district level. Hence the modeled probability of drop out is the same for all students within a district. Because data about dropouts are not available in the ACS, y and n were calculated from the Kentucky SLDS data. All of the predictor variables were from the ACS data at the school district level and all were included in the model as fixed effects. The predictor variables include percent Black, percent Hispanic, percent Asian, percent Other Race, and percent of individuals under age 18 in poverty.

The race/ethnicity percentages were calculated from the total number of students enrolled in grades 9 to 12 in both private and public schools. Both private and public school were used in this case because the ACS tables did not differentiate enrollment counts for race/ethnicity by type of school. All percentages were calculated at the school district level. This model was fit to 168 observations, one for each school district, and 6 variables.

Model Results

School districts were identified by their LEA numbers and used in the model as random effects to account for extra binomial variation and to account for the variation in dropout rates between school districts. Models were fit to the data with and without this variable and the addition of school districts as a random effects variable significantly improved the fit of the model. As shown in Table 9.5, the deviance for the model with the random effects is significantly lower (chi-square=627; p<0.001). These results provide further support for the inclusion of random effects in the model.

Table 9.6 presents the overall model results. None of the variables were significant predictors of high school dropout. School districts were included in the model as random effects

Table 9.5: Model 1. Logistic Regression ANOVA Results With and Without Random Effects, American Community Survey 2009-2013, Kentucky SLDS 2013

Model	DF	Deviance	Chisq	Chi DF	Significance
Model 1 without Random Effect	6	1705.9			
Model 1 with Random Effect	7	1070.8	627.93	1	***

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Sources:** American Community Survey (ACS) 2013 5-year estimates, Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

and once their variation was accounted for, none of the other district-level predictors had a significant effect on dropout.

Table 9.6: Model 1. Logistic Regression Results For Predicting Student Dropouts, Kentucky, Grades 9-12, American Community Survey 2009-2013, Kentucky SLDS 2013

Predictor	Coefficient	Std. Error	Significance
Intercept	-5.0893	0.2408	***
Percent Black Non-Hispanic	1.7557	1.0242	
Percent Hispanic	0.4794	2.1631	
Percent Asian Non-Hispanic	-5.8523	3.7747	
Percent Other Race Non-Hispanic	2.2379	2.4942	
Percent Under 18 in Poverty	0.7636	0.7063	

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Sources:** American Community Survey (ACS) 2013 5-year estimates, Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

Model 2. Predicting High School Dropout Rates at the Student Level Using Kentucky SLDS Data

Model 2 was fit to Kentucky's SLDS data in a mixed effect logistic regression, with data collected for $N = 200,899$ students: $y = (y_1, ..., y_N)$ holds a 0 or 1, depending on whether or not the student dropped out. Using vector-wise notation we have

$$\begin{aligned} y &\sim \text{Bernoulli}(p), \text{ with} \\ \text{logit}(p) &= \alpha + X_{S1}\beta_{S1} + X_{S2}\beta_{S2} + X_{S3}\beta_{S3} + X_{D1}\beta_{D1} + X_{D2}\beta_{D2} \\ &\quad + X_{D3}\beta_{D3} + X_{D4}\beta_{D4} + X_{D5}\beta_{D5} + Uz, \end{aligned}$$

where:

- X_{S1} - X_{S3} are design matrices holding student-level fixed effects predictor variables: race/ethnicity, gender, and grade,

- X_{D1} - X_{D5} are design matrices holding district-level fixed effects predictor variables: enrollment counts, percent Black, percent Hispanic, percent Asian, and percent Other Race,
- U is a design matrix assigning students to school districts, and
- $z = (z_1, \ldots, z_K)$ is the random effect for each of the $K = 168$ school districts, modeled as mean zero normal effects $z \sim N(0, \sigma_z^2 I_K)$

All variables in this model were fit to the Kentucky SLDS administrative data. The outcome variable in this model is a binary indicator for dropout status at the student-level (1=dropped out, 0=did not drop out). The fixed effects predictor variables include student-level race/ethnicity, gender, grade, as well as school district-level total enrollment, percent Black, percent Hispanic, percent Asian, and percent Other Race. As in Model 1, school districts were identified by their LEA numbers and included in the model as random effects. This model was fit to 200,899 observations, one for each student, and 9 variables.

Model results. Table 9.7 presents the model results. The significant predictors at the student level include gender, race, and grade, while at the district level they included percent Black, percent Hispanic, percent Asian, and percent Other Race. Students were more likely to drop out if they were male, Black, Hispanic, or Other Race, as compared to White. Conversely, students were less likely to drop out if they were Asian, as compared to White. Students were more likely to drop out if they were in grades 10 to 12, as compared to 9th grade.

School districts with a higher percentage of Black students were a strong predictor of having a larger proportion of dropouts. Conversely, school districts with a higher percentage of Hispanic, Asian, or Other Race students were predictors of having a lower proportion of dropouts, with percent Asian having the highest effects.

Table 9.7: Model 2. Logistic Regression Results for Predicting Student Dropouts, Kentucky, Grades 9-12, Kentucky SLDS data 2013

Predictor	Coefficient	Std. Error	Significance
Intercept	-5.298	0.1135	***
Student level:			
Male	0.3189	0.0391	***
Asian Non-Hispanic	-0.6441	0.2197	**
Black Non-Hispanic	0.4143	0.0537	***
Hispanic	0.2695	0.0942	**
Other Race Non-Hispanic	0.3955	0.1092	***
Grade 10	0.6461	0.0597	***
Grade 11	0.8441	0.0587	***
Grade 12	0.5656	0.0614	***
Over Age 17 by Oct 1st	-0.1928	0.4689	
District level:			
Enrollment Count (in 10,000s)	0	0	
Percent Black Non-Hispanic	4.12	0.6596	***
Percent Hispanic	-3.877	0.8436	***
Percent Asian Non-Hispanic	-17.62	1.188	***
Percent Other Race Non-Hispanic	-6.521	1.924	***

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Source:** Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

Model 3. Predicting High School Dropout Rates at the Student-Level Model Using ACS and Kentucky SLDS Data

Model 3 outcome and predictor variables were the same as for Model 2, with the addition of the ACS predictor variable percentage of individuals under 18 in poverty. As in Model 2, White served as the reference level for student-level race/ethnicity groups and 9th grade served as the reference level for student-level grade groups. This model had 200,899 observations, one for each student, and 10 variables. Data were collected for $N = 200,899$ students: $y = (y_1, ..., y_N)$ holds a 0 or 1, depending on whether or not the student dropped out. Using vector-wise notation we have

$$\begin{aligned} y &\sim \text{Bernoulli}(p), \text{ with} \\ \text{logit}(p) &= \alpha + X_{S1}\beta_{S1} + X_{S2}\beta_{S2} + X_{S3}\beta_{S3} + X_{D1}\beta_{D1} + X_{D2}\beta_{D2} \\ &\quad + X_{D3}\beta_{D3} + X_{D4}\beta_{D4} + X_{D5}\beta_{D5} + X_{D6}\beta_{D6} + Uz, \end{aligned}$$

where:

- X_{S1} - X_{S3} are design matrices holding student-level fixed effects predictor variables: race/ethnicity, gender, and grade,

- X_{D1} - X_{D6} are design matrices holding district-level fixed effects predictor variables: enrollment counts, percent Black, percent Hispanic, percent Asian, percent Other Race, and percent under 18 in poverty, respectively,
- U is a design matrix assigning students to school districts, and
- $z = (z_1, \ldots, z_K)$ is the random effect for each of the $K = 168$ school districts, modeled as mean zero normal effects $z \sim N(0, \sigma_z^2 I_K)$

Model Results

The ACS variable was at the district level and had little effect on the student-level variables. All of the significant predictor variables in Model 2 were also significant in Model 3 and their interpretation remains the same as Model 2. In addition, a logistic regression deviance test conducted to compare Model 2 with Model 3, showed no significant differences between the two models (Chi-square=1.4, p=0.24) as shown in Table 9.8.

Table 9.8: Model 2 and Model 3-Logistic Regression ANOVA Results, Grades 9-12, American Community Survey 2009-2013, Kentucky SLDS 2013

Model	DF	Deviance	Chisq	Chi DF	Significance
Model 2	16	27054			
Model 3	17	27054	0.0384	1	

Note: */**/*** = Significant at the 0.05/0.01/0.001 level. (**Sources:** American Community Survey (ACS) 2013 5-year estimates, Kentucky Statewide Longitudinal Data System (SLDS) administrative education data.)

C. Summary of Application Results

The descriptive and model results demonstrate several ways that the SLDS data can enhance ACS data. First, SLDS data provides information that the ACS does not measure, such as data about high school dropouts. Second, the SLDS provides data at a more granular level than the ACS (i.e., student level), which may enhance its predictive power.

District-level variables from ACS alone have limited utility for inferring the relationship between demographics aggregated at the district level (e.g., percent Hispanic) and student-level outcomes, such as LEP and dropout rates. For instance, none of the ACS variables in Model 1 of the dropout representative use case were significant predictors of student dropout at the school district level. This suggests that SLDS data have more predictive utility than the ACS data for predicting student-level outcomes. This result is largely due to the fact that the SLDS data are at the student level, providing more direct information about the characteristics of students that can be used to better predict which students are identified as LEP or high school dropout.

Model 3 in both representative use cases demonstrated that non-survey data can be combined with ACS data. The ACS data provided information that was not readily available from the Kentucky SLDS (e.g., poverty) and the SLDS data provided a greater level of granularity and predictive power. Even though the ACS variable was not significant in these analyses, in other states, where such complete information may not be available, variables from ACS data could be used to augment similar analyses. Combining the two types of data allows for a more holistic picture of the populations, especially around topics and information that could be used for policy and future decision-making.

The descriptive comparisons and the model results also provided evidence that SLDS data that measure student dropouts, currently not included in the ACS, and the English proficiency of students have the potential to be used in the American Community Survey. In the case of English proficiency, this could either directly replace questions for student-aged household members or be used in imputation. The findings from the studies using North Carolina and Kentucky longitudinal administrative data can be extended to other states.

10. Conclusions

A. Overview and Findings

The U.S. Census Bureau faces increasing challenges in the development of statistical products, especially where the Bureau must elicit survey respondent participation. These products are primarily based on survey and census data. Declining participation has increased the costs and challenges of survey research. At the same time, the current data revolution has created the expectation for federal agencies to provide timelier and more granular data, especially with geographic detail. As a result, the Census Bureau has begun exploring multiple approaches to using external data to supplement and enhance their current data collection and to create new data products.

This study focused on leveraging ***external*** data sources to enhance official statistics and products. ***External*** data in the context of this study means data that are ***external to the federal statistical system***. The objective was to develop an initial data framework that encompasses the theory and methods capable of capturing, repurposing, and integrating multiple sources of data. Two specific case studies were used, housing and education, to inform the data framework development and to begin the process of characterizing the fitness-for-use of external data sources.

The study was guided by research questions developed in collaboration with the Census Bureau. These questions addressed the parallel interdependent research tracks, the data framework track and the case study track.

Data Framework

- What features are needed for a data framework that characterizes content, access, timeliness, quality, and potential uses of non-federally collected data?
- For which American Community Survey (ACS) questions and for what subpopulations can non-survey sources of direct estimates be obtained at the unit level? Can estimates be modeled at the unit level or at some aggregate geographic and/or temporal level?

Case Studies

- How can non-federally collected data sources enhance or complement a representative use of ACS data?
- What is the value of combining data sources, non-federally collected data sources and/or ACS data, to enhance or complement a representative use of ACS data?

To answer these questions, the study used external data from state and local governments and commercial vendors. The external data were benchmarked against the 2009-2013 ACS.

The housing case study focused on Arlington County and James City County, Virginia. The ***external*** data sources included two major commercial vendors of property data, CoreLogic and Black Knight Financial Services, county-level local property data, Multiple Listing Services real estate data, and commercial neighborhood livability indices from Location Inc.

The education case study first reviewed a variety of external data sources and then focused on the Statewide Longitudinal Data System (SLDS). Data were acquired from Kentucky, North Carolina, Texas, Virginia, and Washington State. These data sources provided detailed student data about demographics, curriculum, limited English proficiency, dropouts, and many other student, school, and school district characteristics.

Through the lens of the case studies, an initial data framework was developed. The findings indicate that it is possible to create a viable data framework that assesses the quality and fitness-for-use of the external sources of data examined in this study and to identify specific ACS questions that could benefit from these external sources of data. There were however challenges in using these sources of data. The housing data required iterative profiling, cleaning, transforming, and structuring to be able to create benchmarks and to use the data in statistical models. The education data provided a census of public school students in the state. The education data were high quality with a relatively few inconsistencies, so implementation of the data framework was reasonably straight forward. The longitudinal nature of the data acquired in both case studies enhances the data's usefulness in statistical analyses.

B. Leveraging Data Acquisition and Current Research

The principlal question addressed in this study was: ***how can we know if external data are useful for federal statistical needs?*** This study represents a preliminary investigation to answer that question and sets the stage for additional research to support the continual development of a data framework for the evaluation of repurposed external data. Some future research that directly leverages the knowledge, expertise, and data acquired for this project are worth noting. These research topics are organized into four areas: (1) data that have the potential to replace ACS data or could be used to impute ACS data; (2) new data that could be added to existing ACS data releases; (3) new data sources that could supply data in real time or at frequent intervals; and (4) data that could be used to create new products.

1. Potential to Replace or Impute ACS Data

a. Self-Reported Tax Assessments, Sale Prices, and Year Built

The current project provides evidence that local data, either from the local government or commercial vendors, may provide more accurate estimates of home values, tax assessments, and year built than through the ACS responses. This conjecture should to be tested in more geographic regions before the Census Bureau could consider the possibility of removing these questions from the ACS and providing comparable estimates through other sources.

b. Student Enrollment and English Language Proficiency

The state administrative education data are a census of public school students. Based on the states studied, the SLDS data could replace ACS questions about K-12 public school enrollment counts by grade, race, and sex, and interactions of these characteristics (grade and sex) with educational attainment. In addition, the ACS variable that asks about the ability to "speak English very well" and "less than very well" could be replaced by state administrative data for children in grades 3-12. This analysis should be extended to other states.

2. Potential to Add New Data Without Adding New Survey Questions

a. Housing Diversity Indices

Extensions of the analyses to measure housing diversity and value characterization at census block group levels would offer new measures of inequality at lower levels of geography. These could also be updated annually, capturing changes on finer time scales. These measures would build on research conducted in this study and applied to new localities.

b. Examination of Pre-1940 housing

Administrative data from local jurisdictions could be used to create profiles of the typical housing units by census tract by decade or groups of decades as necessary, with a focus on older, pre-1940 housing. Additional research to identify the rate of change of older housing versus newer housing is needed. The hypothesis is that the location and character of these older houses may reflect neighborhoods under change. The question becomes do these older homes influence this change or are fewer changes occurring with older houses and thus would require less work by the Census Bureau to maintain the sample and related housing information.

c. Align ACS PUMA to County/School District Areas

The SDLS data provide enormous opportunities to longitudinally track students within public schools and across school districts. These data provide a different geographic overlay to the Census Master Address File and could be used to enhance the Census Bureau's current and future student and labor force characterizations. However, the ACS are aligned with PUMA areas and the SDLS data are geographically aligned with school district and county areas. One county can equal multiple PUMAs (one-to-many relationship), multiple counties can equal one PUMA (many-to-one relationship), and multiple PUMAs can equal multiple counties (many-to-many relationship). To best leverage the SDLS data, algorithms need to be developed for mapping PUMAs to school districts and/or counties.

d. Adding New Education Variables to Published ACS Estimates

There is potential to add new variables to ACS such as the total number of dropouts by characteristics of the student. Additional variables from SLDS data would be analyzed through implementation of the data framework, benchmarking, and use cases to assess the usefulness. An inventory of potential topics of interest could be sought from local, state, and federal policy analysts and be matched with data provided through the SLDS.

3. Potential to Update ACS and Other Federal Statistics on a More Frequent Basis

a. Use of Permitting Data to Capture Rate of Change of Housing Characteristics

The rate of change of housing characteristics in an area might be measured using permitting data. Permits are required for new construction, structural, electrical, plumbing and gas, or mechanical change, yet the data are not easy to use. Additional analysis requires profiling and structuring permit data to calculate rates of changes at the county, tract, and block-group levels. Semi-yearly, quarterly, and monthly estimates could be computed and compared to evaluate the usefulness of these more granular estimates

b. Accounting for Changes in School Boundaries

The changes in geographical boundaries of school districts can affect enrollment counts. These boundaries can potentially change every year, however, the Census Bureau only updates the boundary information every other year. More up-to-date boundary information can be obtained from local data or through commercial vendors, which provides geospatial data on school attendance zones. Research using these data to assess how frequently school district boundaries change and how these changes affect school enrollment counts would be useful.

c. Changes in Student Enrollment Counts Throughout the Year

Enrollment counts and other data about students, teachers, curricmulm, schools, and school districts are updated three times per year for the Fall, Spring, and Summer semesters in most SLDS systems. These data could provide a frequent update of student enrollment counts and related information to adjust ACS estimates.

4. Potential to Create New Products

a. Housing Sales Data as a Statistical Sample

The number of housing units sold on the open market via Multiple Listing Service (MLS) listing each year is considerable, e.g., about 5% of single-family homes in Arlington County and 3% in James City County. MLS data are continuously collected administrative data. There are multiple avenues to access decades of this data. An advantage of the MLS data is that they contain the most up-to-date information of housing units for sale. The usefulness of the MLS data as a statistical sample needs to be explored. The quality of the data, statistical issues such as sample size, selection biases, weighting, and other issues are areas to examine. These data could be a new source of public use microdata.

b. Creating Profiles of Local Areas Using Local Data

Local administrative data sources are a rich source of information about local areas. The housing data provide numerous housing variables that could be used to paint a rich picture of housing in a local area. The education data provide information about student demographics, curriculum, grades, disciplinary actions, and more; as well as detailed data about teachers, schools, and school districts. Each of these sources of data allow for a more detailed description of housing and education in a local or regional or state level. Research should include engaging local government civil servants to identify useful new data products and other local data that could be leveraged such as 911 incidents (police, fire, and Emergency Medical Services), transportation statistics, and public health data.

c. Conduct Longitudinal Analysis of School to Work Transitions

Currently, many states have developed P-20 data systems that track students to the workforce. The Departments of Education and Labor are working together to build on the SLDS data sources through the development of state workforce longitudinal administrative databases. Collecting these and other data sources longitudinally will provide a comprehensive picture of students and later workers' earnings throughout their careers. These data could improve understanding of the relationship between education and training programs, as well as the contributions of other employment services to labor force earnings and career trajectories. These state longitudinal data sources have the potential to extend Longitudinal Employer-Household Dynamics (LEHD) and the Quarterly Census of Employment and Wages (QCEW) data sources, as well as be useful for imputing missing values in other longitudinal and cross-sectional surveys, such as the Current Population Survey, conducted by the Census Bureau.

C. Concluding Thoughts

This project produced an initial data framework for assessing the quality and fitness-for-use for using sources of data that are external to the federal statistical system to enhance and complement ACS data. The origins of the sources of external data differ from federal statistical data in that the Census Bureau has no control and limited knowledge over the measurement and collection processes of the data. To assess their quality and fitness-for-use requires new approaches and methods that are flexible and adaptable to the data and types of issues being addressed. This project provides an important milestone towards developing a disciplined approach to wrangling external data.

Bibliography

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? Quarterly Journal of Economics 106, 979–1014. http://qje. oxfordjournals.org/content/106/4/979.short.

Arlington County (2015). https://propertysearch.arlingtonva.us/Home/Search.

Australian Bureau of Statistics (2009, May). ABS data quality framework. http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Quality:+The+ABS+Data+ Quality+Framework

Batini, C., C. Cappiello, C. Francalanci, and A. Maurino (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR) 41(3), 16.

Bennici, F. J. and E. W. Strang (2015). An analysis of language minority and limited English proficient students from NELS 1988. Report to the Office of Bilingual Education and Minority Languages Affairs, US Department of Education.

Biemer, P., D. Trewin, H. Bergdahl, and L. Japec (2014). A system for managing the quality of official statistics. Journal of Official Statistics 30(3), 381–415.

Braaksma, B. and K. Zeelenberg (2015). "Re-make/Re-model" Should big data change the modelling paradigm in official statistics? Statistical Journal of the IAOS 31(193-202).

Brackstone, G. (1999). Managing data quality in a statistical agency. Survey Methodology 25(2), 139–150.

Brodie, M. L. (1980). Data quality in information systems. Information & Management 3(6), 245–258.

Census Bureau (2013). American FactFinder 2009-2013. http://factfinder.census.gov/ faces/nav/jsf/pages/index.xhtml.

Census Bureau (2014). American Community Survey - Design and Methodology. Version 2.0. http://www.census.gov/programs-surveys/acs/methodology.html.

Chin, A., N. M. Daysal, and S. A. Imberman (2013). Impact of bilingual education programs on limited English proficient students and their peers: Regression discontinuity evidence from texas. Journal of Public Economics 107, 63–78. Christian, D. (2006). Introduction. New York, NY: Cambridge University Press.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology 40(2), 137–161.

Collins, M. and W. Sykes (1999). Extending the definition of survey quality. Journal of Official Statistics 15(1), 57.

Court, A. (1939). Hedonic price indexes with automobile examples. The Dynamics of the Automobile Demand.

Daas, P. J., J. Arends-Tóth, B. Schouten, and L. Kuijvenhoven (2008). Proposal for a quality framework for the evaluation of administrative and survey data. In Paper for the Workshop on the Combination of Surveys and Administrative Data, pp. 29–30. Citeseer.

Daas, P. J., S. J. Ossen, and J. Arends-Tóth (2009). Framework of quality assurance for administrative data sources. Paper for the 57th Session of the International Statistical Institute, 16–22.

Davis, J. W. and K. Bauman (2011). School enrollment in the united states: 2011. population characteristics. https://www.census.gov/prod/2013pubs/p20-571.pdf

Deming, W. (1999). On errors in surveys. American Sociological Review 9(4), 359–369.

Dippo, C. (1997). Survey Measurement and Process Improvement: Concepts and Integration, pp. 455–474. John Wiley & Sons, Inc. http://dx.doi.org/10.1002/9781118490013. ch20

ESS (2015, May). Quality assurance framework of the European statistical system, Version 1.2. European Statistical System.

FTC (2014 March). FTC puts conditions on CoreLogic, Inc.'s proposed Acquisition of Data Quick Information Systems. Federal Trade Commission. https://www.ftc.gov/news-events/press-releases/2014/03/ ftc-puts-conditions-corelogic-incs-proposed-acquisition-dataquick

Goodman, J. S. (2012). The labor of division: Returns to compulsory math coursework. Harvard Kennedy School of Government Faculty Research Working Paper Series RWP12-032.

Google (2015). https://developers.google.com/maps/documentation/geocoding/ intro

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric of quality change. The Price Statistics of the Federal Government, pp. 173–196. NBER.

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly 75(5), 861–871.

Hazen, B. T., C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics 154, 72–80.

Hill, R. J. (2013). Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. Journal of Economic Surveys 27(5), 879–914.

Holt, D. T. (2007). The official statistics Olympic challenge: Wider, deeper, quicker, better, cheaper. The American Statistician 61(1), 1–8.

Humes, K., N. A. Jones, and R. R. Ramirez (2011). Overview of race and Hispanic origin, 2010. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Iacoviello, M. M. (2011). Housing wealth and consumption. FRB International Finance Discussion Paper (1027).

ISO (1992). ISO9000 Int'l standards for quality management. International Organization for Standardization.

Iwig, W., M. Berning, P. Marck, and M. Prell (2013). Data Quality Assessment tool for Ad- ministrative Data. Prepared for a Subcommittee of the Federal Committee on Statistical Methodology, Washington, DC (February).

Jemal, A., E. Ward, R. N. Anderson, T. Murray, and M. J. Thun (2008). Widening of socioeconomic inequalities in U.S. death rates, 1993-2001. PLoS ONE 3.

Juran, J. and A. B. Godfrey (1999). Juran's Quality handbook. Republished McGraw-Hill.

Kang, H. S., E. Haddad, C. Chen, and E. Greenberger (2014). Limited English proficiency and socioemotional well-being among Asian and Hispanic children from immigrant families. Early Education and Development 25(6), 915–931.

Keller, S., Shipp, S., Orr, M., Higdon, D., Korkmaz, G., Schroeder, A., Molfino, E., Pires, B., Ziemer, K., Weinberg, D. (2016). "Leveraging External Data Sources to Enhance Official Statistics and Products." Proceedings of the Biocomplexity Institute, Wiki for Technical Report. TR# 2021-065. University of Virginia. https://uva-bi-sdad.github.io/census2016_wiki/ (https://doi.org/10.18130/kn89-xm38)

Kindler, A. L. (2002). Survey of the States: limited English proficient students and available educational programs and services: 2000–2001 summary report. Washington, DC: National Clearinghouse for English Language Acquisition 8.

Lyberg, L. E., P. P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (1997). Survey measurement and process quality. John Wiley & Sons.

Malpezzi, S. et al. (2003). Hedonic pricing models: a selective and applied review. Section in Housing Economics and Public Policy: Essays in Honor of Duncan Maclennan.

Migration Policy Institute (2011). Limited English proficient individuals in the united states: Number, share, growth, and linguistic diversity. National Center on Immigrant Integration Policy.

Moretti, E. (2007). Crime and the costs of criminal justice, pp. 142–159. Washington, D.C.: Brookings Institution Press.

Moss, M. and M. Puma (1995). Prospects: The congressionally mandated study of educational growth and opportunity. first year report on language minority and limited English proficient students. The National Clearinghouse for Bilingual Education at George Washington University.

Muenning, P. (2007). Consequences in health status and costs, pp. 125–141. Washington, D.C.: Brookings Institution Press.

Narwold, A. and J. Sandy (2010). Valuing housing stock diversity. International Journal of Housing Markets and Analysis 3(1), 53–59.

National Clearinghouse for English Language Acquisition (NCELA) (2011, February). The growing numbers of English learner students 1998/99-2008/09. U.S. Department of Education, Office of English Acquisition mini-poster, Washington, D.C.

Ossen, S. J., P. J. Daas, and M. Tennekes (2011). Overall assessment of the quality of administrative data sources. Paper accompanying the poster at the 58th Session of the International Statistical Institute. Dublin, Ireland.

Pitney Bowes (2015). https://www.pitneybowes.com/pr/ location-intelligence-software/geocoding/geostan.html

Plunk, A. D., W. F. Tate, L. J. Bierut, and R. A. Grucza (2014). Intended and unintended effects of state-mandated high school science and mathematics course graduation requirements on educational attainment. Educational Researcher 43, 230–241.

R Core Team (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Ramani, A., K. and A. Noel (2014). Evaluating American Community survey data on school-age children who speak English with difficulty. American Community Survey Users Conference.

Redman, T. C. (1992). Data quality: management and technology. Bantam Books, Inc.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. The Journal of Political Economy, 34–55.

Rouse, C. E. (2007). Consequences for the labor market, pp. 99–124. Washington, D.C.: Brookings Institution Press.

Ruiz-de Velasco, J. and M. Fix (2000). Overlooked & underserved: Immigrant students in US secondary schools. Urban Institute.

Sentell, T. and K. Braun (2012). Low health literacy limited English proficiency, and health status in Asians, Latinos, and other racial/ethnic groups in California. Journal of Health Communication 17, 82–99.

Simpson, E. H. (1949). Measurement of diversity. Nature.

SN-MIAD (2013, June). Methodologies for an Integrated Use of Administrative Data in the Statistical Process (MIAD). Statistical Network Responsible for Developing Methodologies for an Integrated Use of Administrative Data in the Statistical Process.

Statistics Canada (2009, October). Statistics Canada Quality Guidelines, 5th Ed.

Statistics Netherlands (2012). 49 Factors that Influence the Quality of Secondary Data Sources.

Swanson, C. B. (2009). Cities in crisis 2009: Closing the graduation gap: Educational and economic conditions in America's largest cities. Bethesda, MD: Editorial Projects in Education, Inc.

Tayi, G. K. and D. P. Ballou (1998). Examining data quality. Communications of the ACM 41(2), 54–57.

TEA (2015). Enrollment Trend. Texas Education Agency. http://tea.texas.gov/acctres/enroll_index.html

UK Office of National Statistics (2013, September). Guidelines for Measuring Statistical Output Quality, Version 4.1.

UNECE (2013, December). Generic Statistical Business Process Model (GSBPM), Version 5.0.

United Nations Economic Commission for Europe.

UNECE (2014, December). A Suggested Framework for the Quality of Big Data. United Nations Economic Commission for Europe.

UNECE (2015). Using Administrative and Secondary Sources for Official Statistics: A Hand- book of Principles and Practices. United Nations Economic Commission for Europe.

US Census Bureau (2015, January). Review of Administrative Data Sources Relevant to the American Community Survey.

Verschaeren, F. (2012). Checking the usefulness and initial quality of administrative data. 2012 Meeting of the American Statistical Association.

Virginia State (1996). Virginia State Building Code. http://www.dhcd.virginia. gov/index.php/va-building-codes/building-and-fire-codes/regulations/ virginia-state-building-codes-and-regulations-1996-present.html

Waldfogel, J., I. Garfinkel, and B. Kelly (2007). Welfare and the costs of public assistance, pp. 16—174. Washington, D.C.: Brookings Institution Press.

Wang, R. Y., V. C. Storey, and C. P. Firth (1995). A framework for analysis of data quality research. Knowledge and Data Engineering, IEEE Transactions on 7(4), 623–640.

Wang, R. Y. and D. M. Strong (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 5–33.

Weinberg, D. H. (2011). US neighborhood income inequality in the 2005–2009 period. American Community Survey report. Washington, DC: US Census Bureau.

Weinberg, D. H. (2014). Data sources for US housing research, part 1: Public sector data sources. Cityscape 16(3), 131 1936–007X.

Weinberg, D. H. (2015). Data sources for US housing research, part 2: Private sources, administrative records, and future directions. Cityscape: A Journal of Policy Development and Research 17(1).

Zabel, J. (2015). The hedonic model and the housing cycle. Regional Science and Urban Economics 54, 74–86.