Received: 29 August 2017 | Revised and accepted: 16 November 2017

DOI: 10.1002/wics.1426

FOCUS ARTICLE

Harnessing the power of data to support community-based research

Sallie Keller[1] | Stephanie Shipp[1] | Gizem Korkmaz[1] | Emily Molfino[2] | Joshua Goldstein[1] | Vicki Lancaster[1] | Bianica Pires[1] | David Higdon[1] | Daniel Chen[1] | Aaron Schroeder[1]

[1] Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, Blacksburg, Virginia

[2] Performance Analyst, City of Alexandria, Virginia

Correspondence
Sallie Keller, Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, Blacksburg, Virginia.
Email: sallie41@vt.edu

Funding information
National Science Foundation, Grant/Award number: 1338491; U.S. Census Bureau

Volumes of data are generated at every moment as we go through the paces of our daily lives. Many of these data flows are routinely captured through administrative records, social media, and surveys. Historically, agencies at different levels of government have been responsible for curating and reporting statistics about our social, economic, and health conditions associated with these data flows. Recently, the U.S. government has proposed the use of data derived from administrative records at the federal level to support social policy and program evaluation. Why not consider parallel activities at state and local levels? Harnessing local data sources and integrating them with state and federal sources will provide timelier and more geographically specific analyses to support local insights and policy development. Leveraging community-based participatory research, researchers and civic leaders can work together to identify the questions and execute rigorous, yet flexible, processes for building local sustainable community learning cultures based on data-driven discovery. In the process of conducting research with local civic leaders, we have observed that issues can be classified into 3 categories: locating and describing a population within a community; estimating a statistic and a measure of variability; and evaluating a program, policy, or procedure. Through a series of case studies, this paper demonstrates the unexpected value in liberating and repurposing local data.

This article is categorized under:
Applications of Computational Statistics > Organizations and Publications

1 | INTRODUCTION

Natural flows of data about people are generated at every moment of every day as we go through the paces of our daily lives. These data are generated through our daily economic and social activities such as shopping online, posting a message on Facebook, visiting a doctor, paying our taxes, cashing a check, enrolling in school, making an emergency 911 call, swiping our transit card, and captured through government administrative records and social media. Data are also generated at the level of the environment and captured through embedded sensors, air quality measures such as ozone level and particle pollution, weather conditions, traffic patterns, road conditions, and archived in state and federal agencies.

Historically, different agencies at different levels of government have been responsible for curating and reporting statistics associated with these data flows (Office of Management and Budget (OMB), 2017). This has been through national surveys, frequently collected by the Federal Statistical System. Recently, the U.S. government has proposed the use of data derived from administrative records at the federal level to support social policy and program evaluation. The Commission on Evidence-Based Policymaking (CEP) was established to develop a strategy for increasing the availability and use of data to build evidence about government programs (https://www.cep.gov/). Why not consider parallel activities at state and local levels?

Today, there is an opportunity to access and repurpose many sources of data at all levels of government to better understand the social condition and to improve decision-making (National Research Council (NRC), 2013). In particular, significant untapped sources of data exist within local communities. Harnessing these data sources and integrating them with state and federal sources will provide timelier and more geographically specific analyses to support local insights and policy development. Through a series of case studies, this paper describes the benefits of liberating and repurposing local data sources.

2 | DATA SOURCES AND DATA DISCOVERY

Recognition of the wide variety of social data is not new. Data sources have been classified in similar but distinct ways by different researchers, for example, designed versus organic (Groves, 2011); survey versus big data (United Nations Economic Commission for Europe (UNECE), 2014); and survey versus secondary (United Nations Economic Commission for Europe (UNECE), 2015). What is new seems to be the determination to actually discover and acquire organic or big data to support community-based research.

Data discovery should become a formal part of the research process. Data discovery is the identification of potential data sources that could be related to the specific topic of interest, for example, affordable housing or equitable education. The ability to scrape data from the web and wrangle local government administrative data records allows what used to be impossible, writing the story of a topic through the lens of multiple data sources, possible. Therefore, data discovery should start with brainstorming about the possible data sources that exist, thinking as broadly and imaginatively as possible to assemble a list of potential sources. The data discovery process is followed by a formal inventory and data screening process to assess the utility and accessibility of the data sources for the particular research questions.

The concept of data-driven governance has become the buzz among organizations and at all levels of government. However, it is not always well understood that data-driven governance or evidence-based policy development needs to involve data! If we are able to effectively communicate this concept, data can become the new language for communicating within and among departments and offices at all levels of government.

Part of the problem is that what is meant by data is not always well defined. When thinking about the types of data that might be useful to give insight into a community issue or question, it is useful to categorize the data types as this can help drive the data discovery process. The categories we find useful are:

- *Designed data* are statistically designed and intentional observational data collections such as from surveys and experiments.
- *Administrative data* include data collected for the administration of an organization or program.
- *Opportunity data* are derived from internet-based information and social media.
- *Procedural data* focus on how processes and policies shape our lives, such as the algorithm that runs a stoplight or a policy such as the Department of Defense's "Don't Ask, Don't Tell."

These administrative and opportunity data sources are ubiquitous in our daily lives. However, the intent of their collection is not for research or statistical analyses. Although not the focus of this paper, there are challenges in the repurposing of the various data sources to support community research questions. Developing disciplined approaches for this data repurposing is critical (Keller, Korkmaz, Orr, Schroeder, & Shipp, 2017; Wickham, 2014). Also, anticipating the time that will be needed to wrangle the data is important (Dasu & Johnson, 2003).

3 | COMMUNITY-BASED RESEARCH: COMMUNITY LEARNING THROUGH DATA-DRIVEN DISCOVERY

Community-based research, also described as community-based participatory research, typically starts with research questions from the community to ensure relevancy to their issues. This involves a systematic and iterative process in which findings and recommendations are presented to all partners (Israel et al., 2005). The partners are equitably involved in all parts of the planning, research, presentation of results, identification and implementation of policy changes, and evaluation of these changes (Cacari-Stone, Wallerstein, Garcia, & Minkler, 2014). One of the challenges of community-based research involves a commitment to work with communities to sustain the partnership (Minkler, 2005).

We have leveraged the roots of community-based research to develop a community learning process with a focus on working with civic leaders who provide the questions and collaborate in the research. Working with these civic leaders, we execute rigorous, yet flexible, processes for building local sustainable community learning cultures based on data-driven discovery.

The community learning process cycles through discovering data sources to answer these questions, acquiring and repurposing the data sources, conducting exploratory data analysis, generating hypotheses, creating statistical learning models, and collaborating with these civic leaders who provide insights about the findings and contribute to interpretation of the statistical analyses (Keller, Lancaster, & Shipp, 2017). Our goal is to foster learning and to build capacity for data-driven governance. Based on the statistical analyses, the stakeholders can propose policies and interventions. The learning cycle continues through implementation of interventions and their evaluations, leading to policy updates. This creates a dynamic and adaptive data-driven feedback loop.

In the process of conducting research with local civic leaders, we have observed that issues can be classified into three (not necessarily mutually exclusive) categories:

1. Locating and describing a population (human, animal, and inanimate) within a community,
2. Estimating a statistic and a measure of its variability to evaluate its usefulness for the purpose at hand,
3. Evaluating a program, policy, or standard operating procedure.

The case studies to follow demonstrate the unexpected value of local and state administrative, opportunity, and procedural data sources to support community-based research. They also highlight how these data can augment federal data collections.

4 | CASE STUDY I: WHY DO STUDENTS DROP OUT OF HIGH SCHOOL?

High school dropouts earn considerably less than high school graduates. In 2015, annual median income for those with less than a high school degree was $10,000 less than those who graduated. Understanding why students drop out of high school has been limited by lack of data to study this issue. The availability of Statewide Longitudinal Data System (SLDS) education information has changed that (https://nces.ed.gov/programs/slds/about_SLDS.asp). Using grants provided by the Department of Education, many states now make available longitudinal student data to improve research and decision-making.

The SLDS includes administrative data for all children in public schools, such as enrollment information, student characteristics, and dropout status. These data are collected by schools and school districts in the fall, spring, and summer and are reported to the state. The state provides aggregate information to policymakers regarding the health of the public schools. The goal of this case study is to use the detailed student-level data to develop relevant population descriptions.

The case study draws on the student-level data (administrative data) in the Kentucky SLDS system to describe the student population in two geographic regions in Kentucky—Appalachia and non-Appalachia counties. The boundaries of the two geographic regions are defined by the Appalachian Regional Commission (procedural data). Using the geographic-based student SLDS data, we constructed an explanatory model to explore dropout reasons in the two regions (Ziemer et al., 2017). Student-level data from the SLDS included gender, race and ethnicity, grade, dropout status, dropout reasons, and county. Each student's dropout reason was classified into the constructs "pullout" or "pushout" based on the literature (Fine, 1986; Rumberger, 1987):

- Pushout reasons: boredom, course selections, expulsion, failing grade, and teacher conflict.
- Pullout reasons: employment, family problems, illness, marriage, and pregnancy.

These reasons for dropping out are an example of the richness of the SLDS administrative data that is not available in national federal survey data collections.

An extensive data repurposing, profiling, and cleaning process resulted in 64 combinations of frequency counts by region, grade, gender, race, and dropout reasons. Log-linear models and correspondence analysis are two complementary statistical methods that were used to identify and display the interactions among many categorical variables (Agresti, 2013).

The findings show that overall, both pushout and pullout reasons were common in both Appalachia and non-Appalachia. However, the reasons within each group varied in the regions. White males in the non-Appalachia dropped out because there were few courses that met their needs, they were failing classes, did not get along with their teachers, or got married (see top left of Figure 1). In contrast, white males in Appalachia dropped out because they were expelled, had to go to work, were bored, were sick, or had family problems (see bottom left of Figure 1).

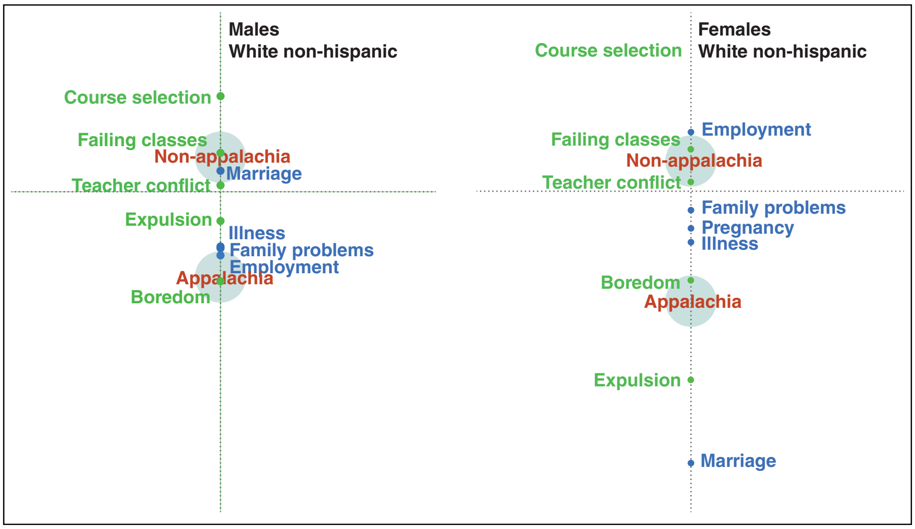

FIGURE 1 Correspondence analysis plot of dropout reasons for ninth grade male (left panel) and female (right panel) students in Appalachia (top) versus non-Appalachia (bottom) regions of Kentucky. Pushout reasons (green) and pullout reasons (blue) are displayed. The correspondence analysis' principal coordinate points (large gray circles) for the geographic regions are enlarged to aid in interpretation. The further apart the gray circles are on the vertical axis the more varied the dropout profiles are for Appalachia versus non-Appalachia (i.e., if both gray circles were centered at the intersection of the horizontal and vertical lines the dropout profiles for the two geographic regions would the same). The principal coordinate points for the dropout reasons lie in the neighborhood of the geographic region (gray circles) they are most prevalent in. For example, for both male and female students, there are relatively more dropouts due to boredom in Appalachia than non-Appalachia; whereas failing classes and teacher conflict are more prevalent in non-Appalachia. *Data sources:* Kentucky Statewide Longitudinal Data Systems (school years: 2008–2014), and geographic boundaries of Appalachia, defined by the Appalachian Regional Commission (https://www.arc.gov/appalachian_region/TheAppalachianRegion.asp). Reprinted with permission from Ziemer et al. (2017)

By using SLDS data and geographic classification from local policy definitions, it was possible to develop a statistical description of pushout and pullout reasons for students dropping out of high school. The community learning gained from understanding why students leave high school before graduation provide policymakers and high school staff with the findings to create programs that will motivate students to graduate.

5 | CASE STUDY II: MEASURING NEIGHBORHOOD DIVERSITY USING LOCAL HOUSING PROPERTY DATA

The second case study aims to develop a measure of neighborhood diversity to help Arlington County demographers support questions about school zoning and equity. Studies have shown that diverse neighborhoods are more likely to thrive even under stress (Jacobs, 1992).

This case study uses and compares house values (as a proxy for wealth) to characterize the diversity in Arlington communities at various levels of geography. One source of data is the federal American Community Survey (ACS) which gives the

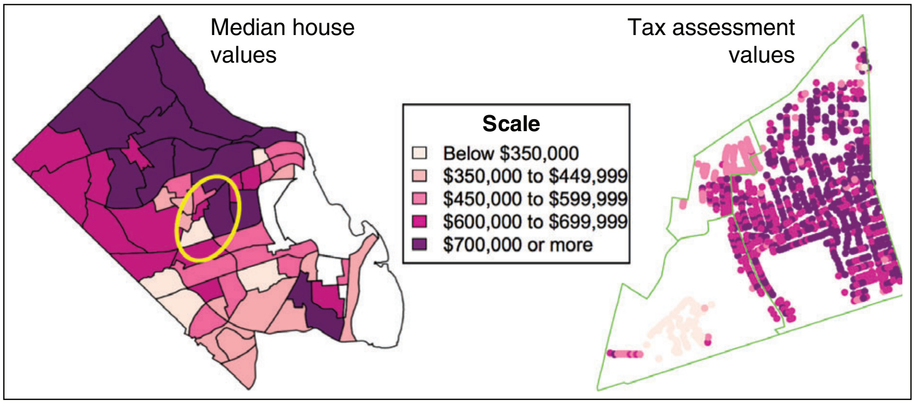

FIGURE 2 House values in Arlington, Virginia. The left map shows median house values based on American Community Survey, and individual local property tax assessment values for the area highlighted by the yellow circle is illustrated on the right map. Data sources: American Community Survey, 2013 and Arlington County tax assessments, 2013

distribution of home values at the census tract level—the left map in Figure 2 shows the median values at the census tract level aggregated into five categories for illustrative purposes. The area highlighted on the map by the yellow circle is shown in more detail on the right map using the actual local county administrative property data for housing values based on 2013 property tax assessments. Since the county property tax assessment data provide the actual values at the house level, it can demonstrate that there is more heterogeneity within a census tract that could not be seen using the ACS data.

To quantify the geographical distribution of the house values within the county, we use Simpson's index of diversity (Simpson, 1949). The specific formulation used is given by

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

where R is the number of housing value categories, and p_i is the proportion of houses in the i^{th} category. 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.

Using the property tax assessment data, Simpson's indices can be calculated at the census tract level and census block group levels. Figure 3 illustrates the diversity scores at the census tract level (left map) and at the block group level (right map) based on the 24 categories used in the ACS housing tables (e.g., Table B25075: Owner Occupied Housing Units, 2013 American Community Survey, 1-Year Estimates [Source: https://factfinder.census.gov]). We observe that the diversity index ranges from .03 to .88. Using local property tax assessment data provide a more geographically accurate estimate of county diversity than the estimate provided by the ACS (Molfino et al., 2017).

The ACS public use microsurvey (PUMS) data also provide information about house value and characteristics, but only at two geographic levels—north and south Arlington—called public use microdata area or PUMAs. There are 498 owner-occupied housing units in the 2013 ACS PUMS files, which are weighted to represent the population. This paints a picture of Arlington that shows north Arlington being wealthier and less diverse than south Arlington. Based on the 498 PUMS data, the diversity score (based on house value) at the county level is calculated as 0.78 (see Table 1). Using Arlington County local property tax assessment data which includes 60,343 single-family homes, the county diversity score is 0.81. The property tax data show more diversity in housing at the county level and a higher difference in diversity between the north and the south of Arlington.

Arlington demographers found the mapping of home values and diversity indices shown in Figure 3, to be very informative for assessing the availability of a variety of housing opportunities for families. Combining these data with Census and ACS demographics provided useful planning information for school enrollment by age group. The comparisons also present opportunities to use local data to supplement or potentially replace ACS collection for selected housing variables such as home value (Keller et al., 2016; Molfino et al., 2017).

6 | CASE STUDY III: LINKING DATA SILOS—EXAMINING UTILIZATION OF FIRE AND MEDIC UNITS

This case study looks at the complexity of local 911 fire and emergency management data to address a question that was posed to us by the Arlington County Fire Chief. He asked if it was possible to recreate fire and emergency medical service

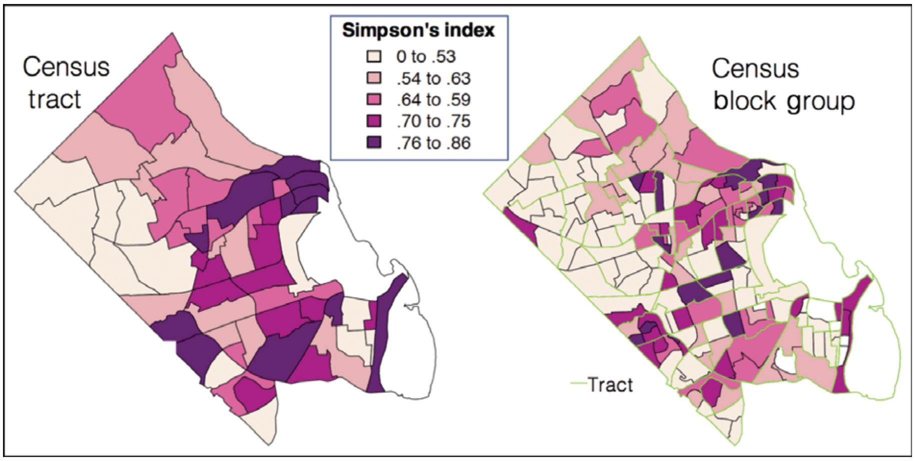

FIGURE 3 Comparisons of neighborhood diversity in Arlington, Virginia. The bottom panel shows the Simpson's diversity indices for Arlington, Virginia using local property tax assessments at the census tract (left) and census block group (right) levels. Data source: Arlington County tax assessments, 2013

TABLE 1 Simpson's indices: Comparison using American Community Survey and Arlington Real Estate Tax Assessment Data for 2013. Reprinted with permission from Keller et al. (2016)

Location/Simpson's index	American Community Survey (ACS)	Arlington Real Estate Tax Assessment Data
Arlington	0.78	0.81
North Arlington	0.75	0.74
South Arlington	0.79	0.84

(EMS) incidents end-to-end through the data that are generated during course of the incidents. This would provide a corpus of information that could be used to improve the fire department's situational awareness. Situational awareness is the "ability to have a high level of attentiveness to the environment in a dynamic situation, or making the proper decision and acting on it in the most appropriate fashion" (Jenkins, 2011).

The question sounded simple until we realized each type of the fire and EMS data is collected and maintained in separate, disconnected data systems. These data silos include the computer-aided dispatch (CAD) call data, the fire engine and medic units dispatched, response times, after-action reports, and the EMS medical records. Data need to be linked across these systems to recreate the sequence of events. They are each collected and stored in different systems, by different individuals, and without a common linking variable.

The data linkage process was complicated. It was accomplished by using multiple variables, date, CAD call time, units deployed, addresses we geocoded through Google APIs, and response times (various available times to locations). Some of the linking was straightforward and other linking required probabilistic model-based approaches to account for time and distance gaps. The after-action reports and the EMS medical records were then attached based on estimated times and locations. When things did not align, call type was matched to information in the after-action reports and the EMS medical records. Text processing was required to extract relevant data for both after action reports and the medical records.

The top display in Figure 4 shows the unit-level linked data by time of day for January 25, 2014. Linking the data in this way was valuable for both incident recreation and for answering the fire chief's first question about the utilization of medic units. He said that "We feel that we are always out of medic units but we do not know how to show or measure this." By linking the unit-level data over 3 years, we were able to create a dynamic visualization (the bottom display in Figure 4 is plotted for the hours of the same day in the visualization) of medic and fire unit use over time. This visualization aggregates when each of the fire engine units (the gold lines) and medic units (the blue lines) were out on calls, showing the volume of units in use and the comparing this to total units available. From this, the peaks in unit utilization, as well as when the utilization hit the limit of units available (dotted golden and blue lines represent maximum number of units in the fleet) become apparent.

Once these data were linked at the unit level, it was then possible to create boxplots that showed the median time on call and the time on call for the two types of units (Figure 5). The boxplots show fire engine units were used for median times of about 10 min. Many of the units were used for less time due to false alarms and a few units were used for longer time when

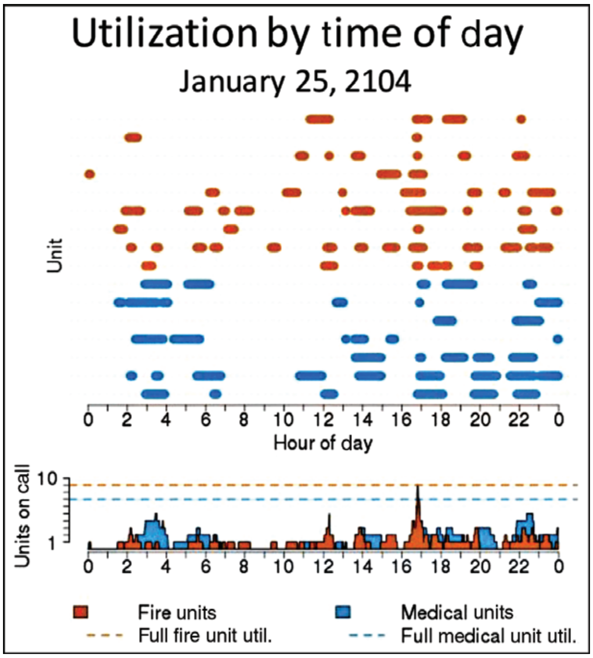

FIGURE 4 Fire and emergency medical service ((EMS) unit utilization in Arlington County, Virginia. The top chart shows the unit-level linked data by time of day for January 25, 2014. The panel below aggregates the units over the hours of the same day, showing the volume of units in-use and then comparing this to total units available. Data source: Fire/EMS data for Arlington, Virginia (2011–2013). Reprinted with permission from Keller, Lancaster et al. (2017)

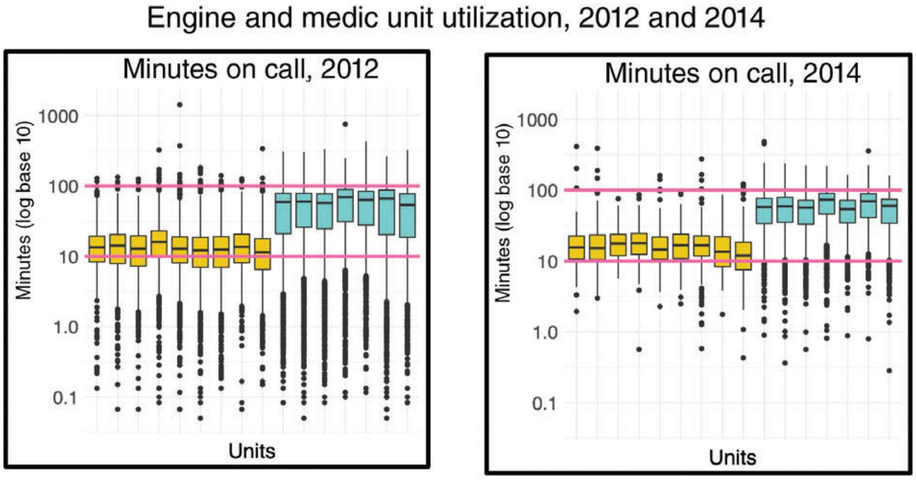

FIGURE 5 The charts illustrate the distribution of minutes on call (in log base 10 scale) for fire (gold) and medic (green) units for 2012 and 2014. The two horizontal lines mark 10 and 100 min. The variation in time on call is much larger in 2012 than in 2014. There appears to be fewer false alarms and short calls in 2014. Data source: Fire/emergency medical service (EMS) data for Arlington, Virginia (2011–2013). Reprinted with permission from Keller, Lancaster et al. (2017)

there was an actual fire event. Medic units are out for a median time of 100 min—much longer than engine units. These findings allowed the Fire Chief to better allocate his medic units and justify requests for more resources.

7 | CASE STUDY IV: IDENTIFYING VULNERABLE POPULATIONS THROUGH 911 DATA

This case study demonstrates how naturally occurring data (as opposed to periodic surveys) can be used to characterize populations. The example focuses on the 911 fire and emergency management data discussed in the previous case study.

Many people call 911 even though there is not a true emergency. This increases the stress and anxiety among firefighters and results in an inefficient use of resources (Lamplugh, 2017). By analyzing the 911 call volumes for the incident data across the Arlington County, frequent callers (those who call three or more times in a month) can be identified (see Figure 6). Using the after-action reports to identify calls that were nonemergencies and then mapping the geographic locations of these calls resulted in regions across the county with large call volumes. Examples of nonemergency calls are when

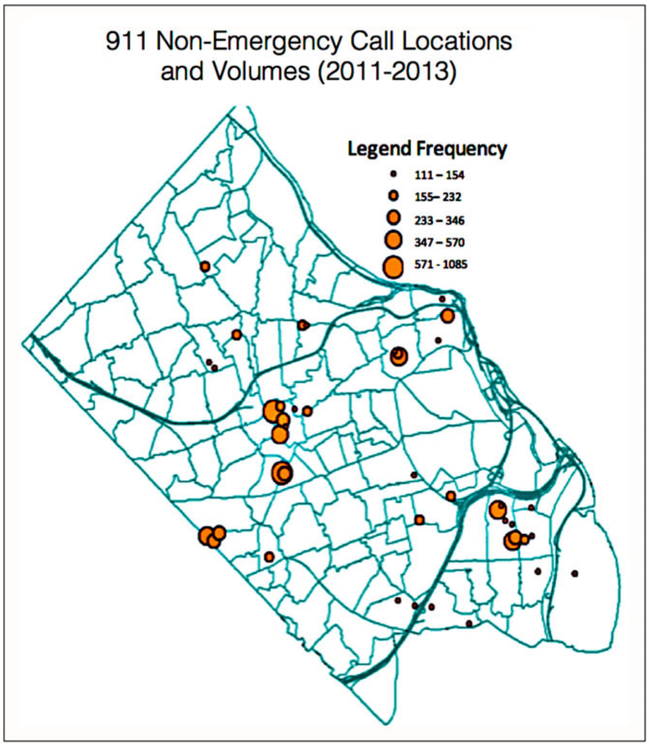

FIGURE 6 Identification of nonemergency call locations of 911 fire and EMS callers. Loyal locations are those that calls come from three or more times in 1 month. Loyal locations indicated by larger circles (centrally located) are where vulnerable populations, especially senior populations, seek support. Data sources: Fire/emergency medical service (EMS) data for Arlington, Virginia (2011–2013). Reprinted with permission from Keller, Lancaster et al. (2017)

a patient or staff call 911 to take a patient for a regularly scheduled doctor's appointment or to pick up a prescription or to ask a question about the hospital's hours of operations.

Looking at these locations, we observed that these calls were coming from areas where senior homes and rehabilitation centers are located. Matching up the census block groups where these calls were coming from with corresponding ACS data confirmed these were areas with high senior and disabled populations. Using the local 911 data allows us to locate vulnerable populations and to track the migration of these populations over time. Matching these data up with other call conditions and police 911 incidents could create an interesting and accurate representation of the various populations across the county such as at-risk youth or areas with poor housing quality.

For this particular case study, the Arlington County Fire Chief worked with the head of the human services agency to arrange visits to senior homes and rehabilitation centers to identify where and when alternative services are needed and to provide training about the purpose of 911 calls. This intervention resulted in a decline in calls implying fewer fire and medic units on the road and better allocation of resources in the Fire and Human Services departments. However, it is anticipated that the 911 call volumes will begin to rise in the future, pointing to the need to revisit these locations and repeat the outreach education.

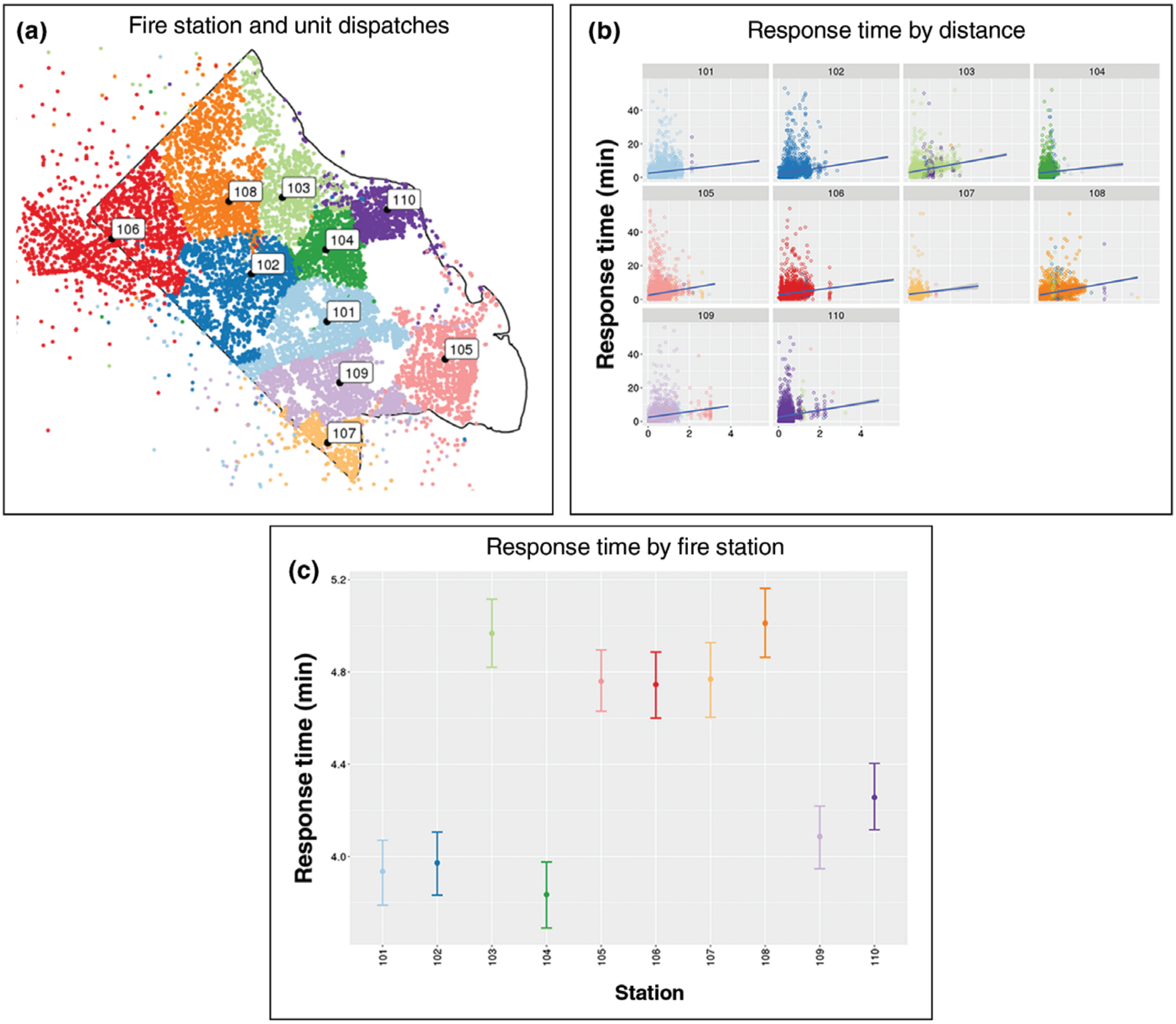

FIGURE 7 Analysis of response times for 10 fire stations in Arlington, Virginia. (a) Top left panel shows the geographic location of the calls that units were dispatched to. The colors identify the responding stations. (b) Top right panel plots the response times in minutes by distance for each station. (c) The bottom panel graphs the 95% credible intervals for response times in seconds controlling for hour of the day, computer-aided dispatch (CAD) call type, station, year, month, and apparatus type, and limited to incidents within a mile of the station and with response times less than 20 min. Reprinted with permission from Arnsbarger et al. (2016)

8 | CASE STUDY V: IDENTIFYING THE FACTORS THAT AFFECT FIRE AND EMERGENCY RESPONSE TIMES

The previous two case studies lead to this final study that aims to understand the variation in emergency response times through statistical modeling. The target incident response time for fire and EMS 911 calls is 320 s from the time the call arrives at the fire station (dispatch time) to the time the units arrive on the scene (National Fire Protection Association (NFPA), 2016). The incident response time by distance for the 10 Arlington fire stations is calculated from the fire/EMS data as described in case study III. Distance is calculated "as the crow flies," from the responding station location to the incident location. In the top left panel of Figure 7, each dot indicates the location of the calls that fire/EMS units were dispatched to, and the colors indicate the responding stations. The top right panel shows the response times by distance for the 10 stations in Arlington. Some of the outlying (longer response times) data is indicative of the fact that when a fire station has all its units on a call, another fire station will respond although it is farther away. Nonetheless, one thing to note is that all of the stations look very similar in their response times by distance (Figure 7, top right panel). But, are they?

Using the 870,906 observations from Arlington fire/EMS data for 2010–2015, we developed a Bayesian linear model for response time to analyze the effect of various predictors (Arnsbarger et al., 2016). The variables include hour of the day, CAD call type, station, year, month, and apparatus type, on response time to an incident within one mile of the station. Response times are computed as the difference between arrival time and dispatch time. The model excludes (a) the CAD calls coded as "Alarm" as these are monitored or activated fire alarms, with no physical signs of smoke or fire reported and (b) observations with response times longer than 20 min (less than 1% of the data).

The findings from the Bayesian analysis are displayed in bottom panel of Figure 7. The notable result is that controlling for the predictors, one is able to see that the variability in response time is a function of the station and that some stations respond significantly faster than others, far below the 320 s target. This analysis provides information to Arlington County Fire Department that can help them look into the main causes of these differences.

9 | CONCLUSIONS AND FUTURE SCOPES

Civic leaders face many challenges in maintaining the quality of life of their residents. They must deliver critically important services, forecast and meet the demand for future services, and identify and provide outreach to vulnerable populations. Local governments must be innovators to face these challenges. Our community learning process involves researchers working with civic leaders to repurpose their administrative data flows and to combine administrative data with other data sources (geospatially generated opportunity data, local procedures and policies, as well as state and federal data, including federal surveys) to provide more accurate and timely information for decision making. Through this process, civic leaders can accelerate the process of achieving a more efficient and equitable quality of life for their residents.

As our work with Arlington County expands to new localities, we are in the process of creating and maturing the Science of Data Science to improve our understanding of the statistical underpinnings of repurposing these administrative data flows. This includes estimating a statistic and a measure of its variability to understand the uncertainty of the findings. Our next steps involve proposing, implementing, and measuring the effects that result from changing a policy or program. In this brave new world of using multiple data sources, many of them not designed for statistical analysis, bringing statistical rigor is a mandate that we cannot ignore.

ACKNOWLEDGMENTS

This research was partially supported by the U.S. Census Bureau under a contract with the MITRE Corporation and was partially based on work supported by the National Science Foundation under Grant No. 1338491.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

REFERENCES

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Arnsbarger, M., Flynn, J., Madison, S, T., Korkmaz, G., Goldstein, J., & Lancaster, V. (2016). *911 response times*. Retrieved from https://www.bi.vt.edu/sdal/content/generic_page/Arlington-County-911-Response-Time-Data-Science-Public-Good-2016.pdf

Cacari-Stone, L., Wallerstein, N., Garcia, A. P., & Minkler, M. (2014). The promise of community-based participatory research for health equity: A conceptual model for bridging evidence with policy. *American Journal of Public Health*, *104*(9), 1615–1623. https://doi.org/10.2105/ajph.2014.301961

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. Hoboken, NJ: Wiley.

Fine, M. (1986). Why urban adolescents drop into and out of public high school. *Teachers College Record*, *87*(3), 393–409.

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

Israel, B. A., Parker, E. A., Rowe, Z., Salvatore, A., Minkler, M., López, J., et al. (2005). Community-based participatory research: Lessons learned from the centers for children's environmental health and disease prevention research. *Environmental Health Perspectives*, *113*(10), 1463–1471. https://doi.org/10.1289/ehp.7675

Jacobs, J. (1992). *The death and life of great American cities*. New York, NY: Vintage Books.

Jenkins, W. A. (2011, May). Scene safety situational: Awareness saves lives. *JEMS: A Journal of Emergency Medical Services*, *36*(5), 30, 323. https://doi.org/10.1016/s0197-2510(11)70116-4

Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application*, *4*(1), 85–108. https://doi.org/10.1146/annurev-statistics-060116-054114

Keller, S., Lancaster, V., & Shipp, S. (2017). Building capacity for data-driven governance: Creating a new foundation for democracy. *Statistics and Public Policy*, 1–11. https://doi.org/10.1080/2330443X.2017.1374897

Keller, S., Shipp, S., Orr, M., Higdon, D., Korkmaz, G., Schroeder, A., … Weinberg, D. (2016). *Leveraging external data sources to enhance official statistics and products*. Report prepared by the Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, for the US Census Bureau. Retrieved from http://cdn.vbi.vt.edu/mc/SDAL/leveraging-external-data-sdal-2016.pdf

Lamplugh, M. W. (2017, May). The stress in EMS: Effects of stress on the unsung heroes of the EMS profession. *Journal of Emergency Medical Services* Retrieved from. http://www.jems.com/articles/2017/05/the-stress-in-ems-effects-of-stress-on-the-unsung-heroes-of-the-ems-profession.html

Minkler, M. (2005). Community-based research partnerships: Challenges and opportunities. *Journal of Urban Health*, *82*(2), ii3–ii12. https://doi.org/10.1093/jurban/jti034

Molfino, E., Korkmaz, G., Keller, S. A., Schroeder, A., Shipp, S., & Weinberg, D. H. (2017). Can administrative housing data replace survey data? *City*, *19*(1), 265–292.

National Fire Protection Association (NFPA). (2016). *Standard for the organization and deployment of fire suppression, emergency medical administration operations, and special operations to the public by career fire departments*, 1710. Retrieved from http://www.nfpa.org/codes-and-standards/all-codes-and-standards/list-of-codes-and-standards/detail?code=1710

National Research Council (NRC) (2013). *Frontiers in massive data analysis*. Washington, DC: National Academies Press.

Office of Management and Budget (OMB) (2017). *Statistical programs of the United States government: Fiscal year 2017*. Executive Office of the President, Washington, DC. Retrieved from. https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/information_and_regulatory_affairs/statistical-programs-2017.pdf

Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. *Review of Educational Research*, *57*(2), 101–121. https://doi.org/10.3102/00346543057002101

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*, 688.

United Nations Economic Commission for Europe (UNECE). (2014). *A suggested framework for the quality of big data*. Retrieved from https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2

United Nations Economic Commission for Europe (UNECE). (2015). *Using administrative and secondary sources for official statistics: A handbook of principles and practices*. Retrieved from https://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10349.aspx

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. Retrieved from. https://www.jstatsoft.org/v059/i10, https://doi.org/10.18637/jss.v059.i10

Ziemer, K. S., Pires, B., Lancaster, V., Keller, S., Orr, M., & Shipp, S. (2017). A new lens on high school dropout: Use of correspondence analysis and the statewide longitudinal data system. *The American Statistician*. https://doi.org/10.1080/00031305.2017.1322002

How to cite this article: Keller S, Shipp S, Korkmaz G, et al. Harnessing the power of data to support community-based research. *WIREs Comput Stat*. 2018;e1426. https://doi.org/10.1002/wics.1426