Data Shop

Data Shop, a department of Cityscape, *presents short articles or notes on the uses of data in housing and urban research. Through this department, the Office of Policy Development and Research introduces readers to new and overlooked data sources and to improved techniques in using well-known data. The emphasis is on sources and methods that analysts can use in their own work. Researchers often run into knotty data problems involving data interpretation or manipulation that must be solved before a project can proceed, but they seldom get to focus in detail on the solutions to such problems. If you have an idea for an applied, data-centric note of no more than 3,000 words, please send a one-paragraph abstract to david.a.vandenbroucke@hud.gov for consideration.*

Using Administrative Data for Spatial and Longitudinal Analysis of the Housing Choice Voucher Program

Eric Schultheis
Massachusetts Institute of Technology

Gregory Russ
Carolina Lucey
Cambridge Housing Authority

Abstract

Place and time are important dimensions of the administration of and policy behind the Housing Choice Voucher Program (HCVP). Spatial and longitudinal analyses of the HCVP are rare, however. In part, this scarcity is because of the lack of widely available, microscale spatial and temporal HCVP data. This article introduces a process that researchers and public housing authorities (PHAs) can use to generate a spatially located, person-period data set of participant households in the HCVP, using off-the-shelf software and administrative data that HUD requires PHAs to collect via Form HUD-50058. This spatially located, person-period data set enables researchers and PHAs to conduct a variety of longitudinal and microscale spatial analyses not possible using untransformed 50058 data or other widely available data sources.

Introduction

Where do Housing Choice Voucher Program (HCVP) households live? How has the spatial pattern of where they live changed over time? Using currently available HCVP data, these two seemingly simple questions are surprisingly difficult to answer, at least at the subcity scale. These two questions are directly relevant, however, to ongoing research and administration questions about the HCVP, such as whether HCVP participants live in "high-opportunity" neighborhoods.

Public housing authorities (PHAs) capture and store a wealth of data about client demographics and spatial location. In most cases, these data are collected in administrative databases designed to support program operations and that comply with various HUD reporting requirements. One such data source, Form HUD-50058, provides comprehensive secondary, household-level, program-participant data. The 50058 data include household-level information for all participants in the public housing program, HCVP, and Section 8 Moderate Rehabilitation program. In exhibit 1, we summarize the information available from the 50058 data.

This article outlines a process to transform the 50058 data into a spatially located, person-period data set using off-the-shelf software and free tools.[1] The resulting data set enables researchers to

Exhibit 1

A Summary of the Data Captured by Form HUD-50058

Section (selected)	Contents (selected)
Agency	Agency name; PHA code; program
Action	Action type; effective date; action correction (y/n)
Household	Head of household; household size; demographic information about all household members: name, age, sex, relation, disability, race, ethnicity, and citizenship
Background at admission	Date entered waiting list; ZIP Code before admission; homeless before admission (y/n)
Unit to be occupied on effective date of action	Unit address; number of bedrooms; date of last HQS inspection; structure type; year structure built
Assets	Owner of asset; asset type; asset cash value; income from asset; total household assets
Income	Income by household member; income after exclusions; annual household income
Expected income per year	Amount of permissible deductions by type and household member; household unreimbursed medical expenses; dependent allowance; unreimbursed childcare costs; adjusted annual household income
Total tenant payment	Total tenant payment; most recent total tenant payment; qualify for minimum rent hardship (y/n)
Tenant-based vouchers	Number of bedrooms on voucher; qualify as hard-to-house family (y/n); utility allowance; housing assistance payment to owner

HQS = Housing Quality Standards. PHA = public housing authority. y/n = Binary response of "yes" or "no."
Source: Form HUD-50058

[1] A person-period format has one record for each household for each temporal unit. This format is particularly well suited for a variety of longitudinal analyses (Singer and Willett, 2003).

conduct a variety of longitudinal, microscale spatial analyses that they can use to explore administration- and policy-relevant questions about how the HCVP functions across space and time.[2]

In this article, we first outline the process for transforming the 50058 data into a spatially located, person-period data set. Second, we propose a method to determine an HCVP household's program status from its certification event history and to identify several reliability issues associated with the 50058 data. Third, we briefly describe two analyses of the HCVP, an exploratory spatiotemporal data analysis project and a stock-and-flow mapping project[3] that the Cambridge Housing Authority (CHA), in Cambridge, Massachusetts, performed with these data.

Transforming the 50058 Data Into a Spatially Located, Person-Period Data Set

The 50058 data can support a broad range of heretofore unexplored policy research questions that require temporal and spatial data. The 50058 data must be transformed into a spatially located, person-period data set to support such research, however. In this section, we outline how to spatially locate the 50058 data and how to reshape the 50058 data into a person-period data structure.

Spatially Locating the 50058 Data

The 50058 data can be spatially located because it includes the addresses of program participants. The process of converting address data to geographic coordinates (typically latitude and longitude) is called geocoding (for example, Longley et al., 2010).[4]

Despite advances in geocoding methods, inaccurate or improperly formatted address data pose a substantial challenge in the case of the 50058 data. As with most address data captured for administrative purposes, the 50058 data's structure does not facilitate geocoding.[5] The likelihood that PHAs did not establish or enforce guidelines aimed at normalizing the address capture-and-storage processes magnifies this difficulty. The likelihood that the 50058 address data, like most administrative data, are rife with typographical errors is a further complication.[6]

Commercial geocoding engines deploy various methods to geocode addresses that (1) are in nonstandardized and nonnormalized formats, (2) contain common typographical errors, or (3) both.

[2] Use of the 50058 data for research often raises issues of PHA client confidentiality. These issues range from not inadvertently disclosing client addresses in map visualizations to the handling and use of PHA clients' personal data. If a PHA partners with university-affiliated researchers, the universities' institutional review boards can help the PHA and researchers design protocols to ensure that PHA clients' personal information is sufficiently protected.

[3] Johnson and Nelson (1998) provide a useful introduction to stock-and-flow mapping for those unfamiliar with this form of cartographic representation.

[4] Alternatively, this process is sometimes referred to as "address matching" (Demers, 2008).

[5] The 50058 data likely store addresses in a single attribute field. Address data stored in multiple fields are frequently easier to geocode. For instance, street address data might be stored in separate fields such as street number, street name, and street type.

[6] Such typographical errors could be minimized if agencies implemented data input interfaces that used "lookup" tables or autocomplete functionality. In the authors' experience, PHA data entry systems rarely implement such functionalities.

Although such geocoding engines offer ease of use, they have several drawbacks. First, they can be cost prohibitive. Second, the methods behind commercial geocoding engines are typically a "black box," to protect proprietary algorithms. Thus, although commercial engines likely could geocode the 50058 data addresses as is, they are less than ideal. We sought an alternate engine to geocode the 50058 data addresses and settled on a minimal-cost geocoding service that the Geographic Information System (GIS) Research Laboratory at the University of Southern California (USC WebGIS) developed (Goldberg and Wilson, 2012).[7]

Using the USC WebGIS service necessitated conducting some preprocessing of the 50058 data addresses before geocoding. We determined that the minimal cost and overall transparency of the USC WebGIS service compared with those of commercial services outweighed this preprocessing burden. Before geocoding, we normalized the 50058 data addresses using a combination of lookup tables in Microsoft Access and the "find" and "replace" functions in Microsoft Excel.[8, 9]

Although tedious, the address normalization process can be implemented relatively quickly. We devoted approximately 10 hours of staff time to normalize the addresses of HCVP households, spanning a 7-year period.[10] In the 50058 data, we observed a 15-percent increase in the geocoding success rate after implementing basic address normalization techniques. Using the addresses, our normalization process, the USC WebGIS geocoding engine, and minimal manual geocoding, we achieved good-quality geocodes for more than 98 percent of the 50058 data addresses. In exhibit 2, we summarize our geocoding success rates.

Exhibit 2

Results of Geocoding HCVP Household Residence Addresses From the 50058 Data

Year	HCVP Households	HCVP Household Addresses	Addresses Geocoded	Match Rate (Percent)	Matched Using USC WebGIS Geocoding Engine	Matched Using Manual Geocode
2004	2,876	3,111	3,074	98.8	3,000	74
2005	2,832	3,078	3,036	98.6	2,957	79
2006	2,809	3,038	2,985	98.2	2,885	100
2007	2,952	3,140	3,082	98.1	2,963	119
2008	2,894	3,079	3,033	98.5	2,886	147

HCVP = Housing Choice Voucher Program. USC WebGIS = Geographic Information System Research Laboratory at the University of Southern California.

[7] To geocode more than 2,500 addresses, a user must register as a partner with USC WebGIS. Additional information about USC WebGIS usage rules is available at https://webgis.usc.edu/About/UsageCosts.aspx.

[8] We used lookup tables to list the variable spellings of road names in the 50058 data to create a new data set, wherein we assigned a given address component (that is, street name, city, county, or state) a single spelling. Using this process, we were able to correct most typographical errors in the 50058 data addresses. In addition, we used Microsoft Excel's "find" and "replace" functions to normalize the format of the 50058 data addresses by removing address data irrelevant to the geocoding process (such as apartment numbers).

[9] Detailed technical notes about our address normalization and reshaping process are available on request from the authors. The technical documentation includes a detailed listing of the Excel functions and Structured Query Language queries used.

[10] Because of reliability issues at the ends of the data set, we were able to use only 5 years of the data.

Reshaping the 50058 Data Into a Person-Period Format

We reshaped the 50058 data into a person-period format.[11] In exhibit 3, we present, for one HCVP household, the 50058 data in the original event-level structure and also in a person-period structure. The person-period structure supports various longitudinal analyses, and we argue that it is, in general, a more useable format.

Many statistical software packages (for example, R and STATA) can reshape data from an event-level to a person-period structure. We used Microsoft Access to reshape the 50058 data, which had several advantages. First, our approach enabled collaboration between CHA database administrators fluent in Structured Query Language and policy staff fluent in Access' graphical user interface. Second, we designed the reshaping process iteratively and tested the logic of our design by viewing the results of intermediate queries.[12] Third, the iterative design and implementation afforded us the opportunity to identify, investigate, and correct issues in the 50058 data during the reshaping process.

Exhibit 3

Sample HCVP Household Data in Event-Level (a) and Person-Period (b) Structures

(a) Event-Level Structure

Tenant ID	Certification Effective Date (2003)	Income
1	January 1	150
1	April 28	200
1	December 1	120

(b) Person-Period Structure

Tenant ID	Period (2003)	Last Certification Effective Date (2003)	Income
1	January 1	January 1	150
1	February 1	January 1	150
1	March 1	January 1	150
1	April 1	January 1	150
1	May 1	April 28	200
1	June 1	April 28	200
1	July 1	April 28	200
1	August 1	April 28	200
1	September 1	April 28	200
1	October 1	April 28	200
1	November 1	April 28	200
1	December 1	December 1	120

HCVP = Housing Choice Voucher Program.

Issues Related to the Transformed 50058

Determining an HCVP household's program status for each period in the transformed 50058 data is one of the more difficult aspects of our process. In all likelihood, an HCVP household's program certification type (from which one can determine HCVP status) will be recorded inconsistently in the 50058 data. We assume that inconsistent recording of certification type is widespread across PHAs and thus describe, in some detail, our method of interpolating HCVP status.

[11] See footnote 1 for the definition of a person-period data structure.

[12] Inspecting intermediate steps of the reshaping process would be substantially more difficult if we used R or STATA.

When certification event type is missing or inconsistently recorded, determining an HCVP household's program status is difficult. For instance, looking at one certification event that lacks certification type information, one cannot tell if the certification event is a program start certification, an annual recertification, or a program termination certification. Most analyses using the 50058 data will require that an HCVP household be assigned a program status for every period in the transformed 50058 data. For instance, mapping HCVP households at a given point in time requires knowing whether a particular HCVP household was a program participant at that given time.

We interpolated an HCVP household's program status using a set of assumptions grounded in HCVP requirements and CHA staff expertise. In particular, we assumed that, absent contrary evidence in the 50058 data, an HCVP household would have a certification event every 12 months, in accordance with federal law.[13] In addition, we assumed that an HCVP household reported major changes to its composition or income via an interim recertification, as required by the same federal regulation. Applying these rules produced unreliable results in the first and last years of the 50058 data analyzed. Because of this unreliability, despite having the 50058 data covering the period from January 1, 2003, through December 31, 2009, we excluded data from 2003 and 2009 (that is, the data set's edges).

For every month in the period analyzed, we assigned HCVP households one of six program statuses: *Not Yet in Program, Start of Program Participation, Program Participant, Final Certification, Termination Ghost, or No Longer Program Participant*.[14] We used the following decision rules to interpolate an HCVP household's program status from its certification event history.

- *Not Yet in Program*: All periods before an HCVP household's first certification.
- *Start of Program Participation*: The first full period after an HCVP household's first certification became effective.
- *Program Participant*: All periods between an HCVP household's first and last certifications.
- *Final Certification*: The first full period after an HCVP household's last certification.
- *Termination Ghost*: The 12 months (periods) after an HCVP household's final certification. If PHAs consistently recorded program terminations, this status would be unnecessary. Given the quality of the 50058 data, however, we knew only that an HCVP household had its last certification on a particular date. We had no knowledge of when its program participation terminated. If an HCVP household had no subsequent certifications, based on the assumption that an HCVP household should have an annual recertification every 12 months, we assigned that HCVP household Termination Ghost status for 12 months after its last certification. This status indicates that we were unsure of the HCVP household's status in the HCVP, because the program termination date was not properly recorded.

[13] "Family Income and Composition: Regular and Interim Examinations," 24 CFR Part 982.516. 59 FR 36682. July 18, 1994.

[14] Before we excluded records from 2003 and 2009, we included two additional household statuses to indicate instances in which we were unable to ascertain an HCVP household's program participation status because of edge effects. For instance, if an HCVP household's first certification in the 50058 data occurred in 2003, we were unable to determine whether that certification was the HCVP household's first or whether its first certification actually occurred at some previous time outside the 50058 data's temporal span.

- *No Longer Program Participant*: All periods after the Termination Ghost status ended. This status indicates our relative certainty that an HCVP household was no longer participating in the HCVP.

We provide the following example to clarify the application of these decision rules. Imagine an HCVP household that had three certifications. Its first certification occurred on January 1, 2006, its second on June 1, 2006, and its third and final on June 1, 2007.

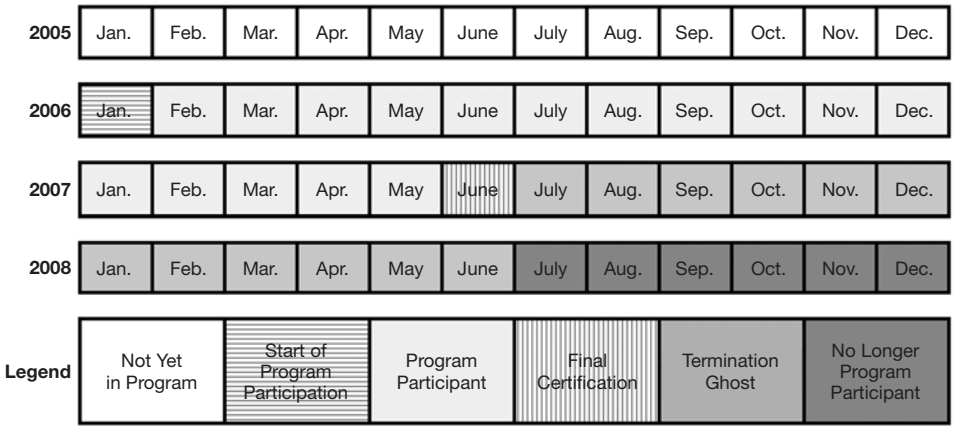
- For all periods before January 1, 2006, we assign the status *Not Yet in Program*.
- For the period of January 1, 2006, we assign the status *Start of Program Participation*.
- For all periods after January 1, 2006, but before June 1, 2007, we assign the status *Program Participant*.
- For the period of June 1, 2007, we assign the status *Final Certification*.
- For the 12 months (periods) after June 1, 2007, we assign the status *Termination Ghost*.
- Beginning on the period of July 1, 2008, and for all subsequent periods, we assign the status *No Longer Program Participant*.

In exhibit 4, we provide a visual representation of the assignment of program participation status to the example HCVP household.

Although our process results in a more useable data set, the resulting product is only as reliable as the underlying data. Researchers have explored the reliability of administrative data in other contexts, but the reliability of administrative housing data, such as the 50058 data, has been explored insufficiently (see, for example, Boehmer et al., 2002; Lee et al., 2005). As such, researchers and PHAs should be cautious about relying solely on administrative data to test hypotheses or evaluate agency policy.

Exhibit 4

Visual Representation of Status Assignment for an Example HCVP Household



HCVP = Housing Choice Voucher Program.

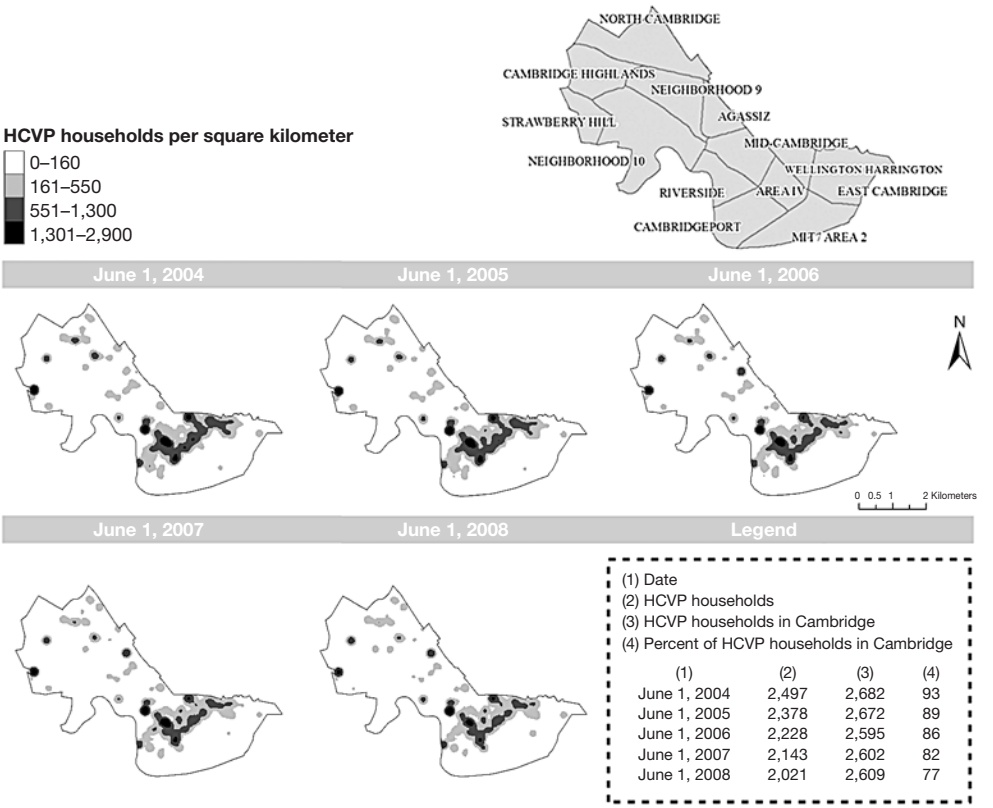
Examples of Analyses Enabled by the Transformed 50058 Data

CHA used the transformed 50058 data to explore several issues related to changing spatial patterns of HCVP household residences. Exhibits 5 and 6 are maps excerpted from an exploratory spatial data analysis (ESDA) project that CHA conducted. Exhibit 5 is a kernel density map of where HCVP households reside in Cambridge, Massachusetts, at five different temporal cross sections. The overall spatial distribution of HCVP households in Cambridge does not change. We observe a progressive decrease in the count and density of HCVP households in Cambridge, however, as we move from earlier to later periods. This phenomenon is easier to read from the tabular element of exhibit 5, indicated on the map with a dashed callout line.

Exhibit 6 is an overlay map of rental-unit density in Cambridge and the residences of HCVP households on June 1, 2008. We observe that the spatial distribution of HCVP households roughly conforms to the density of Cambridge rental units. The exception to this observation is that few HCVP households reside in the more affluent northeastern portions of Cambridge.

Exhibit 5

Kernel Density Map of HCVP Households That Reside in Cambridge



(1)	(2)	(3)	(4)
June 1, 2004	2,497	2,682	93
June 1, 2005	2,378	2,672	89
June 1, 2006	2,228	2,595	86
June 1, 2007	2,143	2,602	82
June 1, 2008	2,021	2,609	77

HCVP = Housing Choice Voucher Program.

Sources: City of Cambridge, Cambridge Housing Authority, July 14, 2011

Exhibit 6

Map of 2008 HCVP Household Residences and Cambridge Rental Unit Density

HCVP = Housing Choice Voucher Program.

Notes: Point symbols of HCVP household residences dispersed in circle pattern to prevent overlapping symbols. HCVP households residing in Cambridge (total HCVP households) = 2,021 (2,609). Number of Cambridge rental units (2000 census) = 29,616.

Sources: 2009 census, Summary File 1; City of Cambridge, Cambridge Housing Authority, July 21, 2011

One working hypothesis that emerged from the ESDA was that HCVP households were increasingly moving out of Cambridge because of relatively higher rental costs compared with those of neighboring cities. CHA tested the salience of this hypothesis by conducting stock-and-flow mapping of HCVP households between 2004 and 2008.[15] This stock-and-flow mapping highlighted the "regionalization" of CHA's HCVP activities and the existence of a substantial number of HCVP household moves out of Cambridge.[16] CHA's stock-and-flow mapping of the 50058 data provided the agency with empirically based and persuasive analysis that challenged the parochial stance of limiting PHA policy inquiry to intra-PHA matters and an agency's formal jurisdictional bounds.

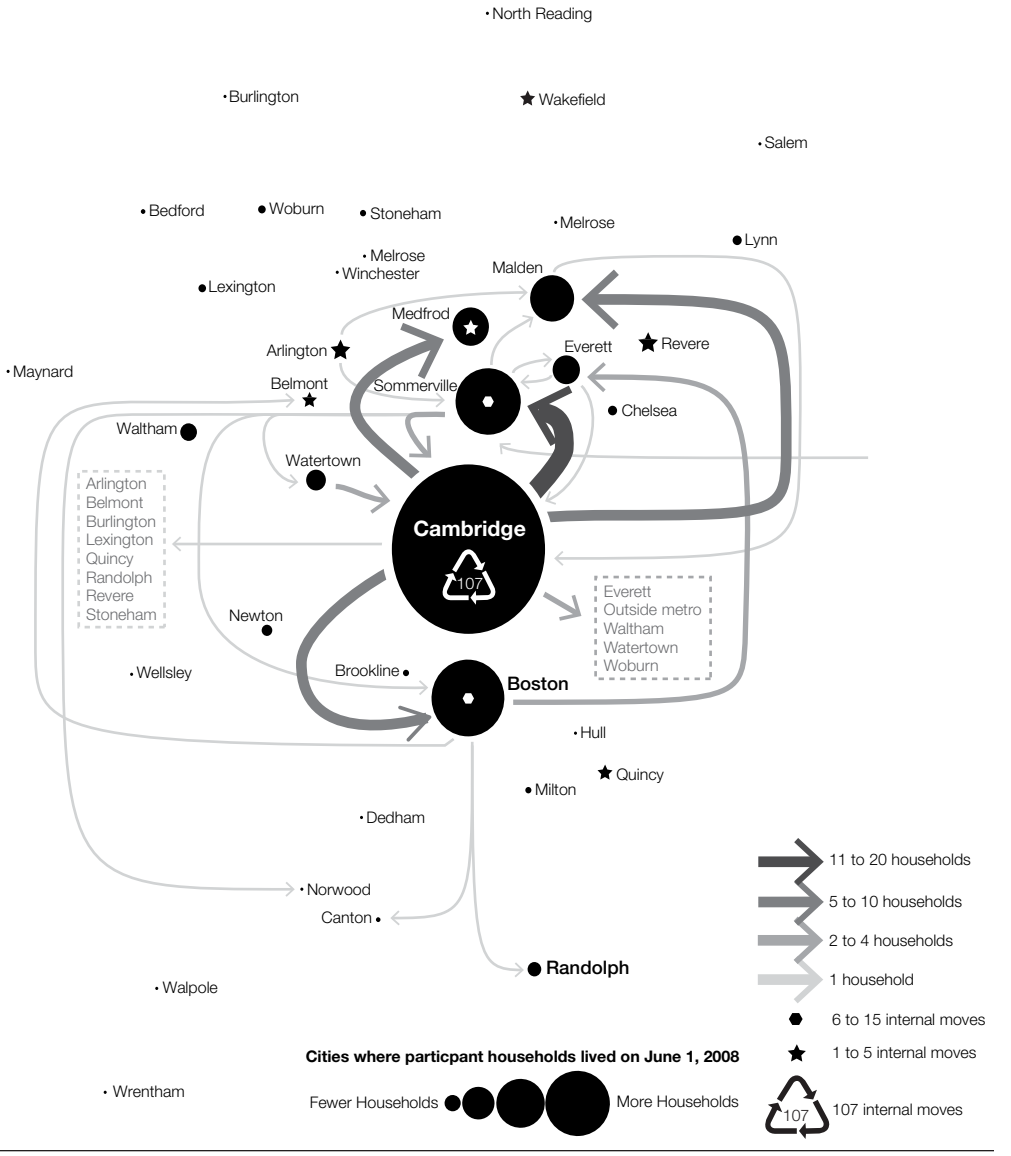
[15] The result of this analysis was a finding that most of the decline in the number of HCVP households residing in Cambridge resulted from the net effect of program starts and ends in neighboring cities. This finding focused CHA's research on understanding why new HCVP households are increasingly leasing up outside Cambridge.

[16] Unlike some PHAs, CHA has the option to continue to administer vouchers outside of its jurisdictional bounds. This option stems from Massachusetts law and CHA's participation in the Moving to Work demonstration project.

Exhibit 7 is a stock-and-flow map of HCVP households for the 2008 calendar year. In exhibit 7, a circle represents a city where at least one CHA HCVP household resides. We graduated each circle's size to reflect differences in the number of HCVP households that reside in the city. The various lines connecting city symbols represent the number of HCVP households that moved from one city to another during the period analyzed. We graduated each line's thickness to represent differences in the number of HCVP households that moved between two cities. The directional arrow points to the city to which HCVP households moved. The line's origin is at the city of departure.

Exhibit 7

HCVP Household Flows, June 1, 2007 to June 1, 2008



Conclusion

The transformed 50058 data enable researchers and PHAs to conduct a variety of microscale spatial analyses of the HCVP. In addition to supporting spatial analyses, the transformed 50058 data support a variety of household-level, longitudinal statistical analyses. Both of these types of analyses offer the potential to deepen our understanding of the HCVP and to better evaluate how it is administered.

Acknowledgments

The authors thank the Doctoral Public Policy Fellowship Program of the Rappaport Institute for Greater Boston, whose support made possible the project on which we based this article. They also thank the staff at the Cambridge Housing Authority, in particular Tito Evora, for their ongoing support and willingness to collaborate on research endeavors.

Authors

Eric Schultheis is a doctoral student in the Department of Urban Studies and Planning at the Massachusetts Institute of Technology.

Gregory Russ is executive director of the Cambridge Housing Authority.

Carolina Lucey is senior program manager for administration and policy at the Cambridge Housing Authority.

References

Boehmer, Ulrike, Nancy R. Kressing, Dan R. Berlowitz, Cindy L. Christiansen, Lewis E. Kazis, and Judith A. Jones. 2002. "Self-Reported vs. Administrative Race/Ethnicity Data and Study Results," *Research and Practice* 92 (9): 1471–1473.

Demers, Michael N. 2008. *Fundamentals of Geographic Information Systems*. Hoboken, NJ: Wiley.

Goldberg, D.W., and J.P. Wilson. 2012. "USC WebGIS Services." https://webgis.usc.edu (accessed April 4).

Johnson, Harry, and Elizabeth S. Nelson. 1998. "Using Flow Maps To Visualize Time-Series Data: Comparing the Effectiveness of a Paper Map Series, a Computer Map Series, and Animation," *Cartographic Perspectives* 30: 47–64.

Lee, Douglas S., Linda Donovan, Peter C. Austin, Yanyan Gong, Peter P. Liu, Jean L. Rouleau, and Jack V. Tu. 2005. "Comparison of Coding of Heart Failure and Comorbidities in Administrative and Clinical Data for Use in Outcomes Research," *Medical Care* 43 (2): 182–188.

Longley, Paul A., Mike Goodchild, David J. Maguire, and David W. Rhind. 2010. *Geographic Information Systems and Science*. Hoboken, NJ: Wiley.

Singer, Judith D., and John B. Willett. 2003. *Applied Longitudinal Data Analysis*. New York: Oxford University Press.