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, PD&R 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.*

A Beginner's Guide To Creating Small-Area Cross-Tabulations

Haydar Kurban
Howard University

Ryan Gallagher
Northeastern Illinois University

Gulriz Aytekin Kurban
University of Chicago

Joseph Persky
University of Illinois at Chicago

Abstract

This short article introduces two techniques of generating cross-tabulations in small areas (for example, block groups or tracts) for which only univariate distributions are available. These techniques require either a microsample or a cross-tabulation from a larger geographic area (for example, a Public Use Microdata Area). One technique uses hill-climbing algorithms, and the other is based on iterated proportional fitting. In this article, we identify the general characteristics of both techniques. We present and evaluate an example (generating cross-tabulations of households by housing value and number of children enrolled in public school), briefly discuss extensions of both techniques to synthetic population construction, and test the synthetic populations by comparing the estimated microsamples with the actual population.

Introduction

A common problem in small-area data analysis is the lack of cross-tabulations for minor geographic areas. For large areas, microdata are often available, from which one can construct cross-tabulations in a straightforward manner. For example, using the Census Public Use Microdata Sample (PUMS), one can construct cross-tabulations for all geographies at or above the level of Public Use Microdata Areas (PUMAs). A PUMA has at least 100,000 residents. For any smaller area, only select cross-tabulations are readily available from the census. Thus, for a census tract in the 2000 Census, American Fact Finder provided information on house value by household income, but these data were not reported for block groups. Even at the tract level, American Fact Finder provided no cross-tabulation for household value by number of household children enrolled in public schools. Recently, for a research project (Kurban, Gallagher, and Persky, 2011) on in-kind transfers, including those through public school systems, this value was one of several key cross-tabulations needed at the block-group level. An interest in imputing such small-area tables is widespread and can arise in any number of projects involving microgeography. The challenge is to build small-area tables in a reliable fashion. This article describes two relatively straightforward, and now quite accessible, techniques for generating synthetic cross-tabulations in small-area data analysis. The first technique, largely developed in Great Britain and favored by geographers (Huang and Williamson, 2002; Ryan, Maoh, and Kanaroglou, 2009; Voas and Williamson, 2000; Williamson, Birkin, and Rees, 1998), is based on hill-climbing algorithms from computer science.[1] The second technique, which is better known in the United States, is iterated proportional fitting (IPF).[2] IPF has been widely used by transportation analysts (Baggerly et al., 1998; Beckman, Baggerly, and McKay, 1996).

The common starting point for both techniques is information on small-area univariate distributions or marginals. For example, the census provides us with the distribution of households across categories of housing values within block groups. For each block group, we also have the distribution of households by number of children in the public schools. What we do not have (and what we want) is the cross-tabulation of households by housing values *and* number of children in the public schools. Both IPF and hill climbing are heuristic methods that start with a real cross-tabulation at a higher level geography and alter it in an effort to reproduce the known marginals for the lower level geography IPF works directly on the higher level table with a number of sequential adjustments aimed at bringing that table into conformity with the small-area marginals. By contrast, hill climbing begins with the raw microdata for the higher level geography and assigns the individual observations to each of the small areas, which compose a larger area. We make these assignments to match the marginals available for each small area.

In the next section, we briefly introduce the conceptual foundations of each of these techniques. We then make some suggestions for introductory software tools that are easily available on the

[1] The term "hill climbing" is used broadly in computer science to cover a range of heuristics based on random search. Given an objective function, hill-climbing methods search randomly around an initial point in an attempt to maximize that function locally (that is, to find a hilltop). To avoid being trapped at a local maximum, hill-climbing algorithms randomly restart their search at more distant points, keeping track of their global performance. See Michalewicz and Fogel (2004) and Russell and Norvig (2003).

[2] Iterated proportional fitting was first introduced by Deming and Stephan (1940).

Internet. Although we do not provide step-by-step instructions, we do present an extended example, which applies both techniques to the same problem. The last section introduces extensions of these two methods that can be used to create full-scale synthetic populations for small areas.

Two Techniques

Perhaps the easiest way to fill in a small-area cross-tabulation is simply to take a larger area cross-tabulation and scale it down to the size of the small area in question. Alternatively, if microdata are available for the larger area, a simple random assignment of household observations to the various component small areas, achieving those areas' total populations, would be expected to generate a similar result. The two techniques presented in this article offer substantial improvements on these naive approaches by incorporating iterative procedures that account for univariate marginals of the small areas. Hill climbing starts with a random assignment of microdata, while IPF starts with scaled-down cross-tabulations.

Hill Climbing

Hill climbing begins by randomly populating small areas with household observations taken from the larger area. One draws household observations from the larger area with replacement and assigns each small area only the number of households it actually holds. At this point, the simulated univariate distributions will not generally match the real distributions. Next, one randomly swaps households from one small area to another so as to improve the match between the real and simulated marginals while holding the small-area populations unchanged. Hill climbing implements pair-wise swaps only if the swaps improve the fit of the allocation. These swaps are repeated several times to improve the fit between the marginals gradually. To avoid becoming trapped in local optimums at the expense of reaching a global optimum, Huang and Williamson (2002), Voas and Williamson (2000), and Williamson, Birkin, and Rees (1998) implemented a flexible annealing procedure that allows the swapping algorithm to accept some swaps that produce poorer performance. This procedure helps improve the overall fit by allowing the algorithm to search across local optimums to get closer to a global optimum. The goodness-of-fit for each allocation is continuously recorded, and a prespecified stopping procedure determines when the swapping will come to a halt.

Iterated Proportional Fitting

A second technique for small-area table construction is iterated proportional fitting (IPF). The basic idea of IPF is straightforward. For a set of small areas that comprise a larger unit such as a PUMA, seed each small-area table with a copy of the PUMA-level table scaled to the small-area population. At this point, any cell in the large-area table just barely equals the sum of corresponding entries in the small-area seed tables; however, neither row marginals nor column marginals from the seed tables will add up to the actual marginals for the small areas. Next, multiply each row in each small-area seed table by a unique constant so that the cells in that row sum to the known corresponding row marginal for the actual small area. After these operations, column sums in the adjusted seed tables will generally not equal the corresponding true column marginals, nor will the cells in the

adjusted seed tables sum to the corresponding cells in the large-area table. The second adjustment takes the new small-area seed tables and multiplies each column by a constant so that its elements add up to corresponding column marginal for the actual small-area data. Finally, a third set of multiplicative adjustments guarantees that entries in the small-area seed tables sum to the corresponding entry of the actual large-area table.

The IPF technique consists of repeated iterations of these row, column, and stack adjustments. The overall process is brought to a halt by specifying an appropriate stopping rule based on the absolute magnitude of cell adjustments. IPF has a strong intuitive appeal. It is well known to converge (Fienberg, 1970; Ireland and Kullback, 1968). In the general case, in which small-area cell entries may be generated from very different processes, the quality of IPF estimates is not guaranteed.

An Example

To illustrate the two techniques for generating small-area cross-tabulations, we now turn to an example suggested by our own work. But, as opposed to a scenario where small area cross-tabulations are not known, we use geographies in this example for which complete data are available. Therefore, for this case, the accuracy of each technique can be fully assessed and compared.

Basic Data for Both Techniques

As suggested previously, our project involved estimating cross-tabulations of households with given housing values and specific numbers of children enrolled in public schools. Although we were interested in generating data for block groups and school districts using known cross-tabulations for PUMAs, for this example, we frame the problem in terms of generating tables for PUMAs using known cross-tabulations for their super-PUMA. By doing so, we can compare the results with known values for those PUMAs.[3]

The data are from the 2000 Integrated Public Use Microdata Series (IPUMS) data file (http://www.ipums.org/)[4] for a suburban Chicago super-PUMA (17100) consisting of four PUMAs (numbers 3101, 3102, 3103, and 3104). The 2 variables of interest are housing values aggregated into 13 categories and number of children in public schools aggregated into 5 categories. The IPUMs provides these data at the household level. Exhibits 1a and 1b present the housing value and school children distributions (that is, marginals) for each PUMA. Exhibit 2 presents the relevant cross-tabulation for the super-PUMA.[5]

[3] PUMAs are defined by the Census Bureau: each PUMA must be contiguous and have at least 100,000 people. PUMAs do not cross state boundaries. Super-PUMAs have at least 400,000 people and are made up of contiguous PUMAs. Like PUMAs, super-PUMAs do not cross state lines. PUMS data allow tabulations at both the PUMA and super-PUMA levels.

[4] University of Minnesota, Minnesota Population Center. 2008. "Integrated Public Use Microdata Series: Version 4.0."

[5] The exhibits presented here are constructed from unweighted micro-observations. They could easily be weighted for use in actual practice.

Exhibit 1a

Univariate House Value Tabulations by PUMA

House Value ($)	PUMA			
	3101	3102	3103	3104
0–49,999	147	80	15	10
50,000–79,999	176	228	41	51
80,000–89,999	137	183	55	41
90,000–99,999	161	179	84	71
100,000–124,999	351	220	196	267
125,000–149,999	356	212	264	283
150,000–174,999	248	111	282	322
175,000–199,999	182	54	247	205
200,000–249,999	164	35	295	207
250,000–299,999	99	10	201	120
300,000–399,999	58	10	96	163
400,000–499,999	13	4	24	56
500,000+	22	7	19	51

PUMA = Public Use Microdata Area.

Exhibit 1b

Univariate Public School Children Tabulations by PUMA

PUMA	Number of Children				
	0	1	2	3	4+
3101	1,381	344	259	88	42
3102	901	197	152	59	24
3103	1,053	322	286	132	26
3104	972	391	306	125	53

PUMA = Public Use Microdata Area.

Exhibit 2

Super-PUMA 17100 Cross-Tabulations

House Value ($)	Number of Children				
	0	1	2	3	4+
0–49,999	181	41	18	7	5
50,000–79,999	342	80	36	28	10
80,000–89,999	280	65	53	13	5
90,000–99,999	344	70	57	17	7
100,000–124,999	658	185	135	42	14
125,000–149,999	707	184	142	56	26
150,000–174,999	572	171	144	55	21
175,000–199,999	394	130	110	44	10
200,000–249,999	362	136	135	50	18
250,000–299,999	229	85	69	35	12
300,000–399,999	146	67	65	38	11
400,000–499,999	44	23	17	9	4
500,000+	48	17	22	10	2

PUMA = Public Use Microdata Area.

The Hill-Climbing Technique

Paul Williamson (2007) designed a readily available application of hill climbing that is suitable for beginners. Williamson calls his technique combinatorial optimization (CO) and a description of his program can be downloaded from his website.[6] Users can quickly adapt the CO application to their particular needs.

The CO program uses the household microdata for super-PUMA 17100 and all PUMA marginals. Starting with PUMA 3101, the program assigns household observations by randomly drawing (with replacement) a subset of the 2,114 households from the super-PUMA microsample (exhibit 2) to match the total population in this PUMA. After each PUMA has been randomly populated, the CO program begins the swapping and simulated annealing procedures. The fit of the swapping procedures is continuously assessed using a goodness-of-fit function proposed by Huang and Williamson (2002) and Voas and Williamson (2001).

The final output is simply a list of households allocated to each PUMA. From this list of households, it is relatively easy to construct any desired cross-tabulation, including the cross-tabulation of house value and number of children of interest here. The final estimate for PUMA 3101 is presented in exhibit 3b. Exhibit 3a contains the actual cross-tabulations for this PUMA. Across all four PUMAs, the program does quite well. The mean absolute error per household suggests that reallocating 6 percent of the households in the super-PUMA would allow an exact match to all four actual PUMA cross-tabulations.

Exhibit 3a

PUMA 3101 Real Cross-Tabulations

House Value ($)	Number of Children				
	0	1	2	3	4+
0–49,999	111	23	8	2	3
50,000–79,999	132	28	6	7	3
80,000–89,999	82	29	19	4	3
90,000–99,999	115	18	21	5	2
100,000–124,999	223	60	47	15	6
125,000–149,999	247	55	35	15	4
150,000–174,999	146	44	39	11	8
175,000–199,999	113	34	24	10	1
200,000–249,999	90	26	30	12	6
250,000–299,999	60	16	13	5	5
300,000–399,999	37	9	11	0	1
400,000–499,999	9	0	3	1	0
500,000+	16	2	3	1	0

PUMA = Public Use Microdata Area.

[6] Go to http://pcwww.liv.ac.uk/~william/microdata/CO%20070615/CO_software.html.

Exhibit 3b

Hill-Climbing Cross-Tabulations for PUMA 3101*

House Value ($)	Number of Children				
	0	1	2	3	4+
0–49,999	110	24	7	2	4
50,000–79,999	128	29	8	8	3
80,000–89,999	94	20	21	2	0
90,000–99,999	116	21	15	8	1
100,000–124,999	236	63	35	12	5
125,000–149,999	243	48	43	14	8
150,000–174,999	153	39	40	11	5
175,000–199,999	109	35	30	5	3
200,000–249,999	98	24	25	11	6
250,000–299,999	55	21	13	6	4
300,000–399,999	24	13	12	6	3
400,000–499,999	3	4	4	2	0
500,000+	12	3	6	1	0

PUMA = Public Use Microdata Area.

** Values are rounded to the nearest whole number.*

The IPF Technique

A number of programs for doing IPF are downloadable from the Internet.[7] Several of these programs are part of elaborate data processing systems and require considerable investment of energy to learn. For beginners interested in generating only simple cross-tabulations, we found the work of Nels Tomlinson and Eddie Hunsinger (Alaska Department of Labor and Workforce Development, 2011)[8] programmed in R the simplest to customize and apply.

The basic inputs to an IPF program are four simple tables. These include (1) the marginals (in our case, the material of exhibits 1a and 1b), (2) the cross-tabulations for the larger area (our exhibit 2), and (3) and (4) an initial "seed table" for each of the small areas. The seed tables are copies of the larger area table (exhibit 2) scaled to the populations of each small area. These tables are the starting points for the successive row and column adjustments that characterize this procedure.

The IPF program generates cross-tabulations for each small area. In exhibit 3c, we present the resulting table for PUMA 3101. As with the hill-climbing technique described previously, the IPF-estimated marginals for the small-area tables match the actual values almost perfectly. A comparison of the cells of the IPF-estimated tables with the actual table cells demonstrates that IPF successfully approximated the cross-tabulations. The overall mean absolute error suggests that a reallocation of 4 percent of the households in the super-PUMA would allow an exact match to all four actual PUMA cross-tabulations. Thus, for this example, the IPF technique performed somewhat better than the hill-climbing technique described previously.

[7] Two well-known software applications are the U.S. Department of Transportation's TRANSIMS program available at http://tmip.fhwa.dot.gov/community/user_groups/transims and Arizona State University's PopGen program available at http://urbanmodel.asu.edu/popgen.html.

[8] For programs, documentation, and an introduction to the iterated proportional fitting literature, go to http://www.demog.berkeley.edu/~eddieh/datafitting.html.

Exhibit 3c

IPF Cross-Tabulations for PUMA 3101*

House Value ($)	Number of Children				
	0	1	2	3	4+
0–49,999	106	24	10	4	3
50,000–79,999	124	28	12	8	4
80,000–89,999	94	21	17	4	2
90,000–99,999	115	22	17	5	2
100,000–124,999	234	58	42	11	5
125,000–149,999	238	55	41	14	8
150,000–174,999	158	40	33	11	5
175,000–199,999	113	32	26	9	3
200,000–249,999	93	29	29	9	5
250,000–299,999	58	18	14	6	3
300,000–399,999	30	11	10	5	2
400,000–499,999	7	3	2	1	1
500,000+	12	3	4	2	0

IPF = iterated proportional fitting. PUMA = Public Use Microdata Area.

** Values are rounded to the nearest whole number.*

Discussion

Both hill climbing and IPF handle the example problem quite well.[9] It should be clear that both table-generating techniques are now executable without a major investment in training and programming. Although IPF worked somewhat more effectively for our example problem, the two techniques seem quite comparable in terms of results. At this level, IPF is probably somewhat easier to implement.

The hill-climbing technique has one major advantage: The immediate product of hill climbing is a full assignment of households to small areas. After that assignment is made, it can be used directly as a set of synthetic populations.[10] From those synthetic populations, it is relatively easy to estimate not just the initial cross-tabulations of interest but also virtually any cross-tabulations of household characteristics in the original data set. Moreover, the hill-climbing technique can be easily extended to include any number of relevant marginal conditions in the initial assignment.

In contrast, constructing a synthetic population using the IPF technique requires a major second step in which household assignments are carried out using repeated random samples constrained

[9] The results achieved here by these techniques are probably better than what can be expected generally. For one thing, the PUMA marginals we used are drawn from the same basic data set as the microsample for the super-PUMA. When the marginals and sampled microdata are drawn from separate sources, a certain amount of sampling variation would be likely. More importantly, the quality of the fit at the small-area level depends on the extent to which higher level and lower level areas have similar table structures or correlations. If the small areas are quite different from each other and from their sum, both techniques are likely to suffer.

[10] Researchers are showing an increasing interest in studying small-area phenomena. For small areas such as block groups or census tracts, microdata samples are generally unavailable. Generating a synthetic population at this small-area level means formulating a collection of households appropriately selected from a larger geography, such as a PUMA. The problem, of course, is to define what selection mechanisms are "appropriate."

by the estimated table cells (Beckman, Baggerly, and McKay, 1996). A system of this type requires more sophisticated programming. It at least necessitates becoming familiar with one of the packages (for example, TRANSIMS[11] and PopGen[12] described in footnote 7) available to the public. These packages also involve considerably more sophisticated versions of the basic IPF algorithm, because extending the dimensionality of the table to be estimated requires building estimates of lower level tables as well.

IPF and hill climbing are promising techniques for creating synthetic populations, but a number of issues remain unresolved. Generating synthetic populations from IPF results involves the synthetic reconstruction of microsamples based on stochastic approaches that are often subject to sampling error. This error is likely to be more significant for small sample areas. Even if the model estimates are unbiased, there is no guarantee that the variance will not be too large. The second problem with IPF is that it uses a sequential procedure. Some error is introduced in each stage as a result of random sampling, modeling assumptions, and data consistency (Huang and Williamson, 2002). Similar to the IPF procedure, the hill-climbing procedure is a stochastic process. Variations in the sample seed values will alter the baseline household selections and estimated distributions. Two previous studies (Huang and Williamson, 2002; Ryan, Maoh, and Kanaroglou, 2009) have compared the performance of IPF and hill-climbing methods for construction of synthetic populations. Both studies concluded that hill-climbing methods outperformed IPF.

Our somewhat tentative recommendation, therefore, is to continue using IPF for simple table construction. If you are generating more complete synthetic populations, however, hill climbing is more intuitive and perhaps more accurate.

Acknowledgments

The authors thank the U.S. Department of Housing and Urban Development's Office of Policy Development and Research for hosting a seminar where an earlier draft of this paper was first presented.

Authors

Haydar Kurban is an associate professor in the Department of Economics at Howard University.

Ryan Gallagher is an assistant professor in the Department of Economics at Northeastern Illinois University.

Gulriz Aytekin Kurban received a Ph.D. in computer science from the University of Chicago.

Joseph Persky is a professor in the Department of Economics at the University of Illinois at Chicago.

[11] TRansportation ANalysis and SIMulation System (TRANSIMS). U.S. Department of Transportation, Federal Highway Administration website: http://tmip.fhwa.dot.gov/community/user_groups/transims (accessed May 2011).

[12] PopGen 1.1: Population Generator, Arizona State University. http://urbanmodel.asu.edu/popgen.html (accessed May 2011).

References

Baggerly, Keith, Richard Beckman, Michael McKay, and Douglass Roberts. 1998. "TRANSIMS Synthetic Population System User Guide." LA–UR–98–9693. Los Alamos, NM: Los Alamos National Laboratory.

Beckman, Richard J., Keith A. Baggerly, and Michael D. McKay. 1996. "Creating Synthetic Baseline Populations," *Transportation Research A* 30 (6): 415–429.

Deming, W. Edwards, and Frederick F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Tables Are Known," *Annals of Mathematical Statistics* 11: 427–444.

Fienberg, Stephen E. 1970. "An Iterative Procedure for Estimation in Contingency Tables," *Annals of Mathematical Statistics* 41: 349–366.

Huang, Zengyi, and Paul Williamson. 2002. A Comparison of Synthetic Reconstruction and Combinatorial Optimization Approaches to the Creation of Small-Area Microdata. Working paper. Liverpool, United Kingdom: University of Liverpool, Department of Geography.

Ireland, C.T., and S. Kullback. 1968. "Contingency Tables With Given Marginals," *Biometrika* 55 (1): 179–188.

Kurban, Haydar, Ryan Gallagher, and Joseph Persky. 2011. Estimating Local Suburban Redistribution in Property-Tax-Funded School Systems. Working paper. Washington, DC: Howard University.

Michalewicz, Zbigniew, and David Fogel. 2004. *How To Solve It: Modern Heuristics*. New York: Springer.

Russell, Stuart J., and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach,* 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Ryan, Justin, Hanna Maoh, and Pavlos Kanaroglou. 2009. "Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms," *Geographical Analysis* 41 (2): 181–203.

Voas, David, and Paul Williamson. 2001. "Evaluating Goodness-of-Fit Measures for Synthetic Microdata," *Geographical and Environmental Modeling* 5 (2): 177–200.

———. 2000. "An Evaluation of the Combinatorial Optimization Approach to the Creation of Synthetic Microdata," *International Journal of Population Geography* 6 (6): 349–366.

Williamson, Paul. 2007. CO Instruction Manuel. Working paper 2007/1. Liverpool, United Kingdom: University of Liverpool, Department of Geography, Population Microdata Unit.

Williamson, Paul, Mark Birkin, and Phil H. Rees. 1998. "The Estimation of Population Microdata by Using Data From Small Area Statistics and Samples of Anonymised Records," *Environment and Planning A* 30 (5): 785–816.

Additional Reading

Wong, David. 1992. "The Reliability of Using the Iterative Proportional Fitting Procedure," *Professional Geographer* 44 (3): 340–348.