Automating Tabulation and Reporting of ACS Data for Emergency Management

Matthew Graham[1,2,3]

Abstract

OnTheMap for Emergency Management[4] is a online tool created by the U.S. Census Bureau to help emergency planners and managers access population and employment data for specific emergency events. Early versions of this application presented LEHD Origin-Destination Employment Statistics (LODES) and 2010 Census SF1 data, both of which are released at the census block level, making tabulation to arbitrary geographical areas relatively simple. In 2014, ACS 5-year estimates for 2008-2012 were included in OnTheMap for Emergency Management for the first time. Because ACS estimates are not published for census blocks, a number of unique issues had to be resolved.

Two of the most difficult methodological challenges were: creating relatively accurate approximations arbitrary geographies, such as wildfire and flood areas, using ACS publication geographies and minimizing the derived margin of error (MOE) for each tabulation. As part of the solution, we allowed multiple summary levels of estimates to be combined in order to meet specific geographic requirements and to create the most accurate estimates (lowest proportional MOE) within each event boundary.

This novel implementation is flexible and can be extended to other uses/requirements. It also points the way to a more general method for handling the tabulation of ACS estimates to arbitrary areas of interest (neighborhoods, economic analysis zones, CBDs, etc.).

This paper will describe the specific needs and requirements of this development process, some of the specific implementation challenges and solutions for adding ACS estimates to the tool, useful lessons for ACS data users, and future development plans.

Background

The Longitudinal Employer-Household Dynamics (LEHD) program[5] is based upon a partnership model with each of the state Labor Market Information (LMI) agencies. Through this partnership, LEHD gains access to administrative data sources with information on firms, workers, and jobs. Also part of this

[1] Geographer, LEHD/Center for Economic Studies, U.S. Census Bureau. matthew.graham@census.gov
[2] Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.
[3] Special thanks to Jody Hoon-Starr for his assistance in preparing the ACS data for analysis.
[4] http://onthemap.ces.census.gov/em/
[5] http://lehd.ces.census.gov/

partnership is a commitment by LEHD to return value-added to the state partners as well as to the public in general. Initially this value-added took the form of newly developed public-use data products on the states' individual and combined labor markets, specifically the Quarterly Workforce Indicators (QWI).

As the program continued to grown, the value-added provided to the state partners and the larger public expanded to include further data products[6] as well as dissemination and analytical tools meant to expand the use community by making the data easier to access and understand and to ease the application of the data to specific analytical questions. The first tool developed for these purposes was OnTheMap, which allows users to select standard legal or statistical geographies or import/develop nonstandard geographies for which the LODES data can be tabulated. These geographies were all based upon the census tabulation blocks (initially those blocks defined after the 2000 Decennial Census, but more recently those defined after the 2010 Decennial Census).

OnTheMap for Emergency Management

As OnTheMap grew in popularity, an increasing number of requests came to LEHD to perform custom tabulations of public-use data for emergency events such as tornadoes, hurricanes, wildfires, and other events (such as the Deepwater Horizon oil spill in 2010). While most of these requests could be performed by anyone with the OnTheMap application, the roadblock for many users was the general lack of knowledge about where to find spatial information on these events and how to bring it into the application.

In order to make this process more efficient and again expand the user community for these data products, LEHD developed OnTheMap for Emergency Management which attempted to solve the problem of users not knowing either how to find or what to do with the spatial information on emergency events. OnTheMap for Emergency Management takes advantage of the live data feeds provided by NOAA, FEMA, and the Department of Interior to automate the collection of event boundary information and the tabulation of LODES data to these event areas.[7] The application provides a portal in which a user needs only know the name or place/time of an event to access pre-tabulated reports and charts on jobs and workers within the event area. The events currently covered by the application are: wildfires, Federal Disaster Declarations, winter storms, hurricanes, and floods.

Block Selection Methodology

With the exception Federal Disaster Declarations, which are county-based, all of the supported events have boundaries that do not conform to standard legal/statistical geography. As such the boundaries must be approximated from the geographies for which we have data. In the case of both LODES and the

[6] LEHD Origin-Destination Employment Statistics (LODES) and more recently the Job-to-Job Flows (J2J) data. More information on all of LEHD's data products can be found at http://lehd.ces.census.gov/data/

[7] A limited set of variables from the 2010 Decennial Census were added to application on a trial basis in 2013.

2010 Decennial Census, the data are published at the census block-level, which is small enough to approximate most event areas.[8]

Specifically, the approach for selecting census blocks to be included in the data tabulation for an event boundary is a simple point-in-polygon method. Each census block carries with it an "internal point," about which the only guarantee is that the point is topologically within the census block. When a fire or other event boundary is brought into OnTheMap for Emergency Management, the application identifies all census blocks whose internal points are within the boundary, and it is the data for these census blocks which are tabulated. Figure 1 shows an example of this kind of selection methodology.

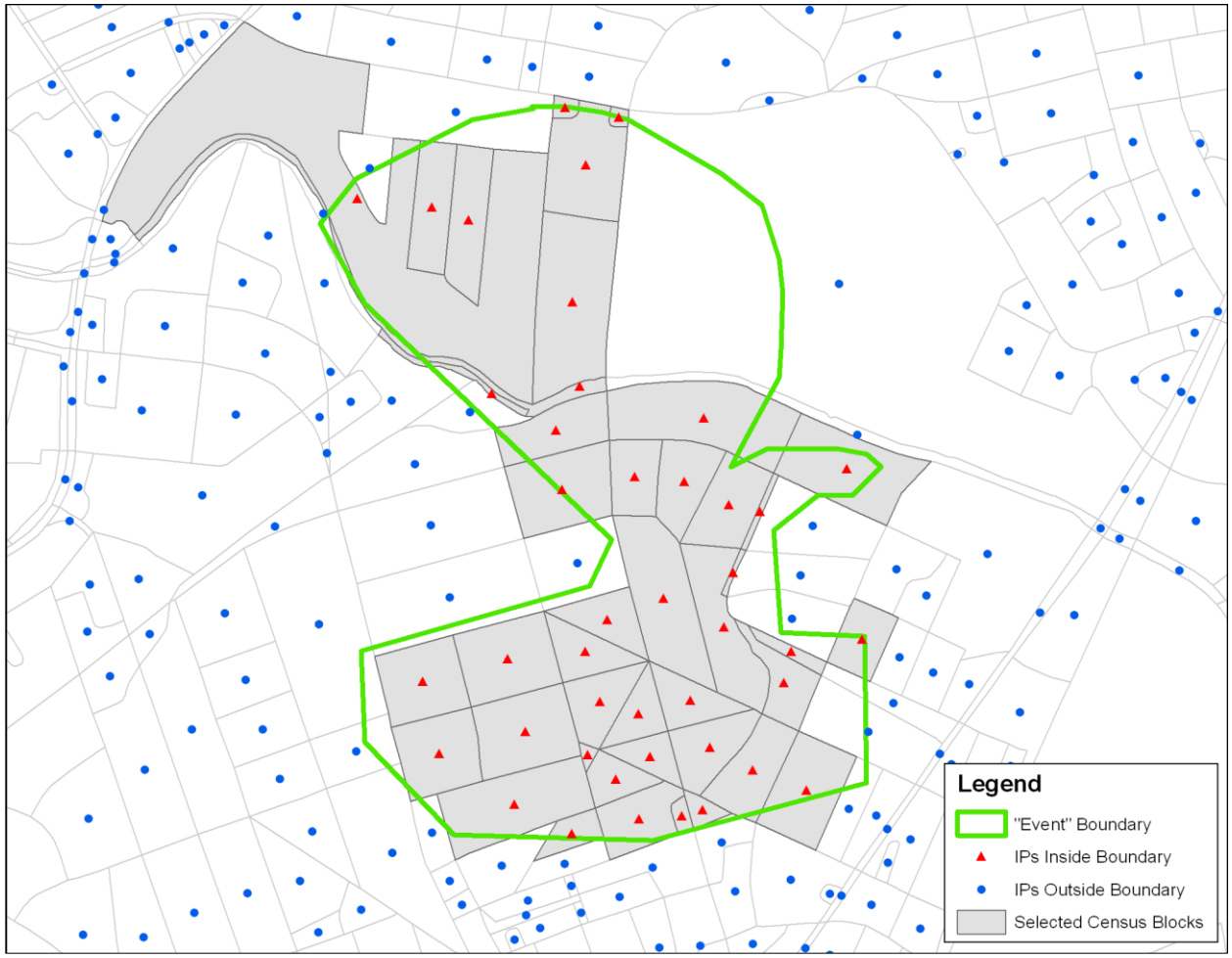

Figure 1: Point-in-Polygon Selection of Census Blocks

Integrating ACS 5-year Estimates

With the 2014 development cycle of OnTheMap for Emergency Management, it was decided to add ACS data into the application due to user interest in a number of demographic and household variables that

[8] See Chapter 2 of https://www.census.gov/ces/pdf/2013_CES_Research_Report.pdf for further discussion of the requirements for implementation of event types within OnTheMap for Emergency Management.

only appear on the ACS. The particular variables of interest include questions related to how the public will be able to respond to a natural hazard or emergency event or what kind of emergency assistance might be necessary.[9]

Data-Driven Challenges

The addition of ACS estimates created two major challenges that were driven by key aspects of the public-use data. Up to this point the application had only used census blocks to approximate event areas. However, even with 5-year estimates the best geographic resolution available via ACS is census blockgroup. This led to the question of whether event boundaries could be approximated by census blockgroups, and if so, what was the best methodology for creating those approximations.

The second data-driven challenge was rooted in the representation of derived margins of error (MOEs). LEHD was committed to appropriately representing the MOEs within the application in a way that was both informative to users and as correct as the data allowed. However, the MOEs released with the ACS estimates become proportionally larger as the cells (or geography in this case) become smaller. As such, the aggregate MOE for all census tracts in a county will be greater than or equal to the MOE for the county estimate itself, even though the total estimates for the county in both cases are the same. As such, the core issue became how to minimize the derived MOE for an arbitrary boundary's estimate.

Approximating Arbitrary Boundaries

Many different methods for approximating arbitrary boundaries with a standard set of geography exist. During the design of OnTheMap for Emergency Management, one of the explicit goals was to make any spatial methodology as simple and as easy to understand as possible. Three different methodologies will be discussed here. These are the methodologies that were discussed during the design process to add ACS data into the application. However, these are not the only possible methodologies.

In order to approximate an arbitrary boundary, we must make some basic assumptions:

1. Because we are using ACS 5-year estimates, we assume that census blockgroups provide enough spatial resolution to adequately approximate an arbitrary input boundary for all supported event types.
2. For one of the methods, we also assume that we have internal points for census blocks and that the census blocks are nested within blockgroups.

Method 1: "Minimum" Approximation

In this method, we select only those blockgroups that are entirely within the input boundary. If even a small part of a blockgroup is not inside the boundary, then that blockgroup is *excluded* from the tabulation. This is a very restrictive approximation (thus the "minimum" appellation) which could be desirable if it is important to the use case that absolutely no population outside the boundary be included in the combined estimate.

[9] A list of data items from ACS that are included in OnTheMap for Emergency Management can be found at http://lehd.ces.census.gov/doc/help/onthemap_em/OTMEM_DataSources.pdf.

Method 2: "Maximum" Approximation

In this method, we select only those blockgroups that intersect the input boundary.[10] If even a small part of a block group is inside the boundary, then that blockgroup is *included* in the tabulation. This is a very inclusive approximation, which could be useful if it is important to the use case that no population inside the boundary be left out of the tabulation.

Method 3: "Middle" Approximation

As their names imply, the Minimum and Maximum approximate can be viewed as bounds on other approximations. As an in-between approximation we started with the Minimum (all those blockgroups totally inside the boundary). Then we found all census blocks whose internal point was inside the boundary and added any additional blockgroups that contained one of these blocks. Effectively this filters out many cases in which very small slivers of a blockgroup intersect with the input boundary. Other methods exist for performing this kind of refinement, but this method could be implemented using only data published in the TIGER shapefiles and without setting an arbitrary minimum area requirement or some similar limit. The is Middle approximation method is the one that is actually implemented in OnTheMap for Emergency Management.

In Table 1, Table 2, and Table 3, we compare these three different selection methods by the count of blockgroups, the land area of those blockgroups, and the ACS 2009-2013 5-year estimates for several actual event boundaries.[11]

Selection Method	Minimum	Middle (IPs)	Maximum
Blockgroup Count	96	155	158
Land Area [sq. mi.]	922	3,001	3,089
ACS Pop Estimate	114,173	181,573	185,981

Table 1: Boundary Approximation for Event 1 (Flood on 3/25/2015)

Selection Method	Minimum	Middle (IPs)	Maximum
Blockgroup Count	11,922	12,272	12,296
Land Area [sq. mi.]	28,344	34,274	34,435
ACS Pop Estimate	17,444,239	18,013,344	18,047,590

Table 2: Boundary Approximation for Event 2 (Snow on 3/5/2015)

Selection Method	Minimum	Middle (IPs)	Maximum
Blockgroup Count	0	3	4
Land Area [sq. mi.]	0	469	477
ACS Pop Estimate	0	1,997	4,359

[10] There are a number of different ways that an intersection test can be implemented. In this case we mean any areal overlap between two polygons, but *not* cases in which two polygons only share a boundary. As is the case with any calculation of this type, numerical precision of the data and the database/system on which it is being calculated can impact the outcome.

[11] Further details about the events used here are presented in the Appendix.

Table 3: Boundary Approximation for Event 3 (Fire on 8/14/2012)

First, these events are of very different size and geographic location. The flood area straddles a state border and include a small city; the snow event crosses many states and includes several large cities; and the fire event includes a small rural/mountainous area that only touches populated places at the edges. Generally, we see from the results that the Middle method is much closer in outcome values (blockgroup count, land area, or total population estimate) to the Maximum method than it is to the Minimum method. For both the Flood and Snow events, the Maximum method outcomes are anywhere from 0.2% to 2.9% larger than the Middle method outcomes, whereas the Minimum method outcomes are 2.9% to 69% lower than the Middle method outcomes.

Another basic observation is that the largest event area (the snow event) has the least proportional difference between the Minimum and Maximum method outcomes. In general, larger areas are less likely to be strongly impacted by the inclusion or exclusion of a single or small number of small geographic areas (blockgroups), a point that is discussed further in the final section.

On the small event side, we see that there is some additional risk for the Minimum method of not capturing any data when the event is small. In the case of the Fire, the event area was not large enough to encompass any whole blockgroups, but it did intersect with four of them. Given the potentially large non-intersecting parts of these blockgroups that contributed several thousand individuals to the population count, raising the question of estimate validity would not be unwarranted. Only data at a more detailed geographic level would allow us to make a conclusive determination.

Minimizing Derived Margin of Error

If the ACS margins of error (MOEs) added linearly as do the estimates, then it would be enough to select a set of blockgroups and tabulate the estimates and derived MOEs[12] associated with those blockgroups. Instead, given the opportunity we look to minimize the derived MOE based on the selection of blockgroups made using the "middle" method described above. The method used here relies upon the hierarchical nature of the State-County-Census Tract-Census Blockgroup geographies along with some approved statistical techniques.[13]

First, higher levels of geography have proportionally lower MOEs and as such, we prefer to replace a set of selected blockgroups with the equivalent tract, a set of selected tracts with the equivalent count, and a set of selected counties with the equivalent state when at all possible. This allows us to combine a set of cells with the lowest MOEs to create the smallest derived MOE. The basic rules for creating derived MOEs between variables or between geographies are outlined in ACS Technical documentation.[14] All combinations of cells performed within OnTheMap for Emergency Management are sums. Also, It is

[12] When an margin of error is produced from the combination of more than one published cell and does not have an exact equivalent in the published ACS tables, the MOE is referred to as "derived."

[13] These additional steps described in this document were implemented based on direct communication with the American Community Survey Office (ACSO), the Decennial Statistical Studies Division (DSSD), and the Social, Economic, and Housing Statistics Division (SEHSD).

[14] http://www.census.gov/acs/www/Downloads/data_documentation/Statistical_Testing/2013StatisticalTesting3and5.pdf

important to note that because covariances are not published for ACS data, they are assumed to be zero.

This section will describe the method that was used to minimize the derived MOEs for each boundary based on the approximate boundary methods. A set of experiments is given below run to show the effect of this aggregation on the derived MOE.

In addition to this basic aggregation, two other steps are also implemented within OnTheMap for Emergency Management:

- When the derived estimate contains more than one cell with a zero-value estimate, the derived MOE is constructed by including in the derived MOE calculation only the maximum MOE for all of the zero-value estimate cells at each geographic level.[15]
- In cases where cell estimates and MOEs have been suppressed for confidentiality purposes,[16] both the estimate and the MOE are treated as zero.

In order to communicate these dynamics clearly to users when an estimate and MOE have been derived from multiple published cell and has no exact published equivalent, OnTheMap for Emergency Management flags the value as being derived. In addition, when an estimate and MOE contain one or more suppressed cells, OnTheMap for Emergency Management marks the values as containing suppressed cells.

In the tables below, we show examples from several events and the difference in derived MOEs when using all blockgroups or a fully combined set of geographies that should minimize the derived MOE.

	All Blockgroups	Combined Areas
Total Pop Estimate	181,573	181,573
Derived MOE	3,381	1,647

Table 4: MOE Improvement Comparison for Event 1 (Flood on 3/25/2015) – Total Population

	All Blockgroups	Combined Areas
Total Pop Estimate	1,8013,344	1,8013,344
Derived MOE	38,248	10,291

Table 5: MOE Improvement Comparison for Event 2 (Snow on 3/5/2015) – Total Population

[15] To make the last phrase explicit: If there are zero-value estimates in some tracts and zero-value estimates in some blockgroups, then the calculation include exactly one zero (and its MOE) for tract level and one zero for the blockgroup level.

[16] See http://www.census.gov/acs/www/Downloads/data_documentation/data_suppression/ACSO_Data_Suppression.pdf for more information on suppression.

	All Blockgroups	Combined Areas
Total Pop Estimate	1,997	1,997
Derived MOE	422	422

Table 6: MOE Improvement Comparison for Event 3 (Fire on 8/14/2012) – Total Population

In the first two events – the Flood and the Snow – we can see that combining the various blockgroups to the highest levels of geography produces significant gains for the derived margin of error. In the case of the flood event (Table 4), the MOE improves from 1.86% of the estimate to 0.907% of the estimate. And in the case of the snow event (Table 5), the MOE improves from 0.212% of the estimate to 0.0571% of the estimate.

In Table 6, we see that for a very small area there is no opportunity to switch out multiple small geographies for larger geographies with improved margins of error, and so the derived MOE for the fire event is unchanged between the two methods.

For cases in which there are data (nonzero cells) in a significant share of blockgroups, most or all of the derived MOE improvement comes from swapping out the small geographies for larger ones. However, for a data item that is very sparse (e.g. skill of speaking English given a different language spoken at home for a specific age cohort), the MOE improvement comes instead from the elimination of multiple zeros within each summary level (geographic level), as described above.

	All Blockgroups	Combined Areas
Estimate	157	157
Derived MOE	445	110

Table 7: MOE Improvement for Event 2 (Snow on 3/5/2015) - Language Variable

Table 7 shows such a case in which such a sparse data item is tabulated.[17] In this case, most of the improvement in MOE – in fact, going from a situation in which the estimate is not significantly different from zero to one in which it is – comes from the dynamic of removing duplicate zeros at each geography level.

Putting It All Together

The following example case combines the Middle geographic approximation and the MOE minimization described above. A short version of this example was initially described in the "OnTheMap for Emergency Management: Geographic Selection Methodology" document[18] on the application's help pages. It is also the algorithm that is implemented within OnTheMap for Emergency Management to tabulate ACS 5-year estimates.

[17] The data item used in this case is B16004_023. This is the table called, "AGE BY LANGUAGE SPOKEN AT HOME BY ABILITY TO SPEAK ENGLISH FOR THE POPULATION 5 YEARS AND OVER." And the item is "5-17 years: Speak other languages: Speak English 'not at all.'" See http://www2.census.gov/acs2013_3yr/summaryfile/ACS2013_TableShells.xlsx for more information.

[18] See http://lehd.ces.census.gov/doc/help/onthemap_em/OTMEM_SelectionMethodology.pdf.

Algorithm

1. Select all states that are wholly within the event boundary.
2. Select all counties that are wholly within the event boundary minus the states selected in #1.
3. Select all census tracts that are wholly within the event boundary minus the states selected in #1 minus the counties selected in #2.
4. Select all blockgroups that have at least one constituent census block with an IP inside the remainder of the boundary (event boundary minus the states selected in #1 minus the counties selected in #2 minus the census tracts selected in #3).
5. Check whether, through the addition of blockgroups, a whole tract, county, or state has been created from its parts. If so, then substitute the larger area for its parts.

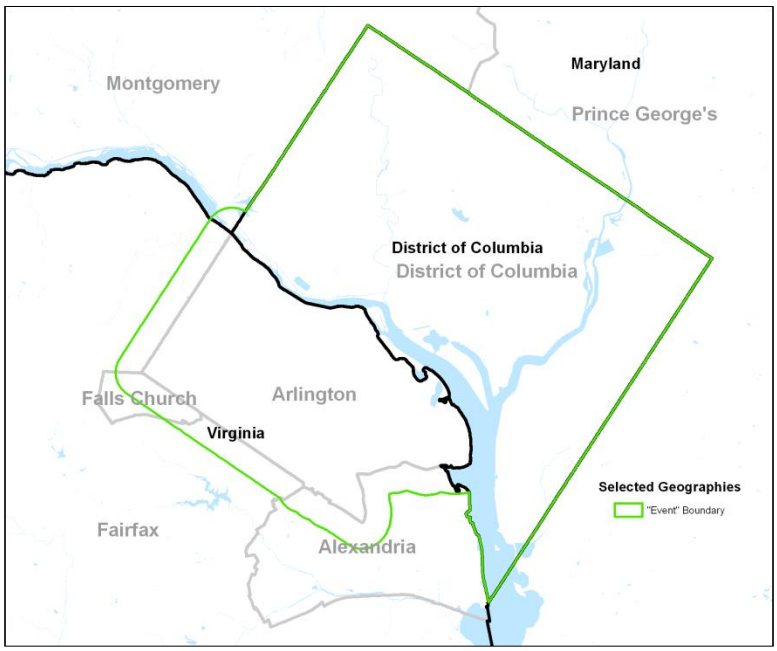

Figure 2

Figure 2 shows the example event boundary outline in green. It is the union of the District of Columbia and Arlington County, VA, where Arlington has been buffered by 1000m. While some parts of the boundary align with the boundary between the District of Columbia and Maryland, the buffer of Arlington County ensures that the boundary does not conform to any legal or statistical areas.

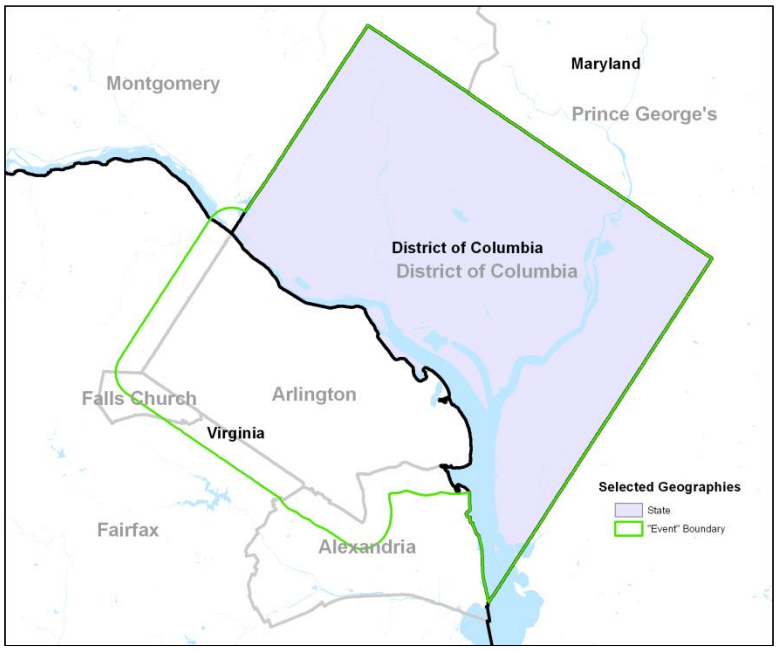

Figure 3

Figure 3 shows the District of Columbia shaded because it is the only state (equivalent) fully within the boundary.

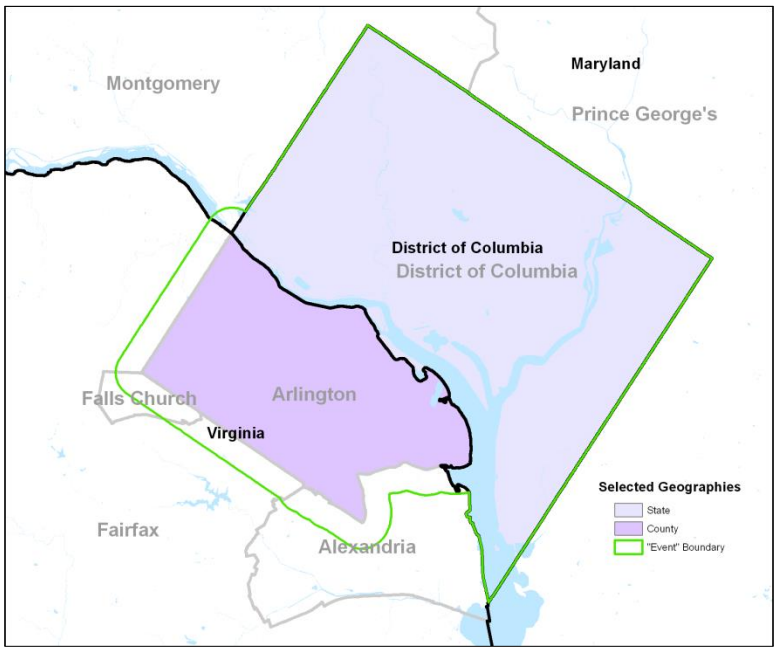

Figure 4

In Figure 4, Arlington County is shaded because it is fully within the remainder of the boundary once the District of Columbia has been removed.

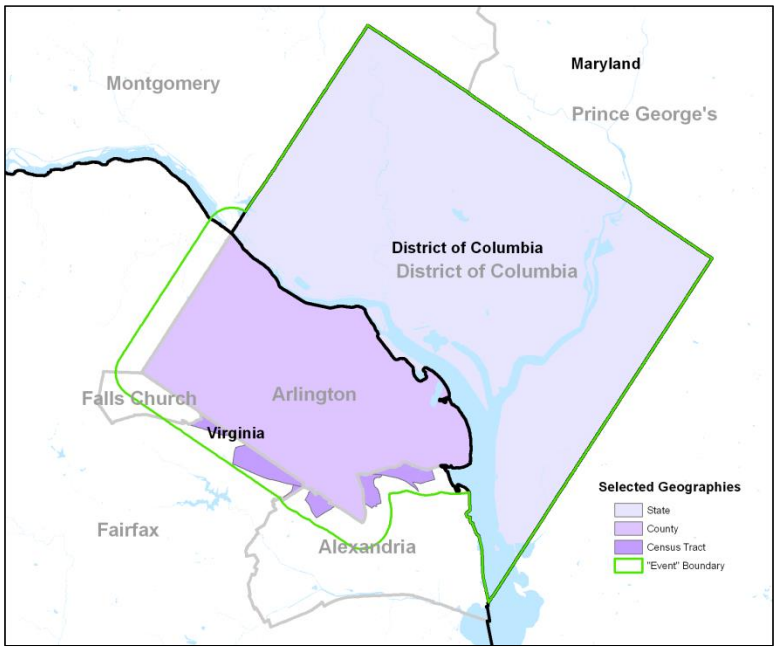

Figure 5

Figure 5 continues the selection sequence by choosing the census tracts that are wholly within the boundary after the District of Columbia and Arlington County have been removed.

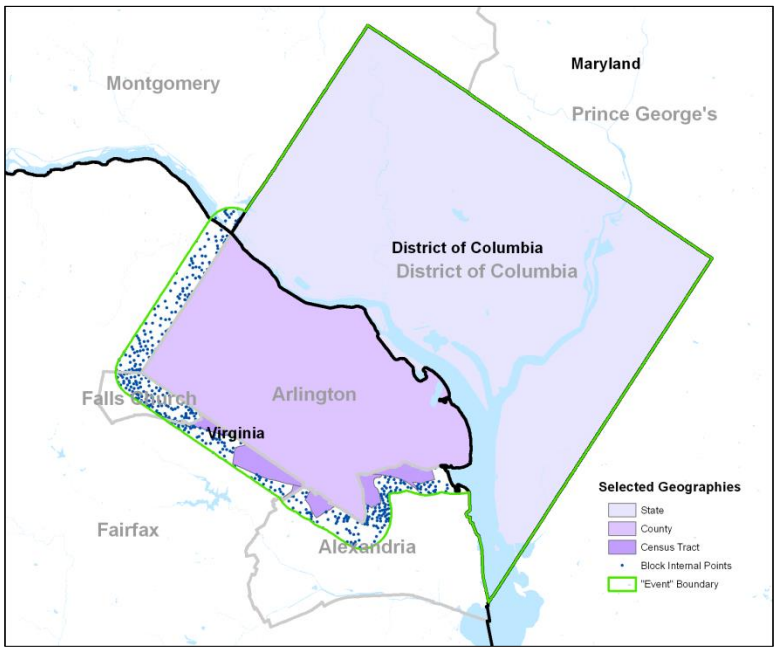

Figure 6

Figure 6 shows the census block internal points (as blue dots) that fall within the boundary once the District of Columbia, Arlington County, and the selected tracts have been removed.

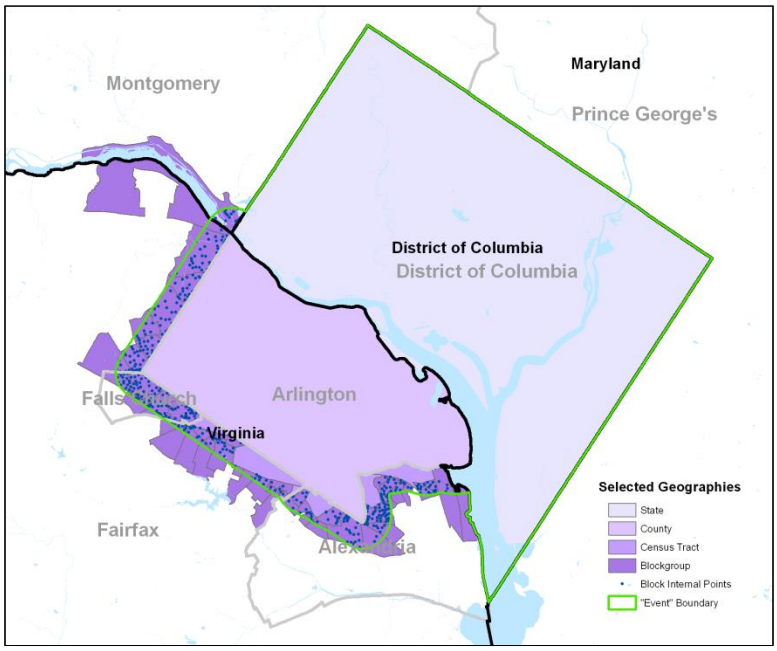

Figure 7

Then all the blockgroups in which these census blockgroups fall are selected and shown in Figure 7. Note that the nature of this method means that some blockgroups extend beyond the input boundary.

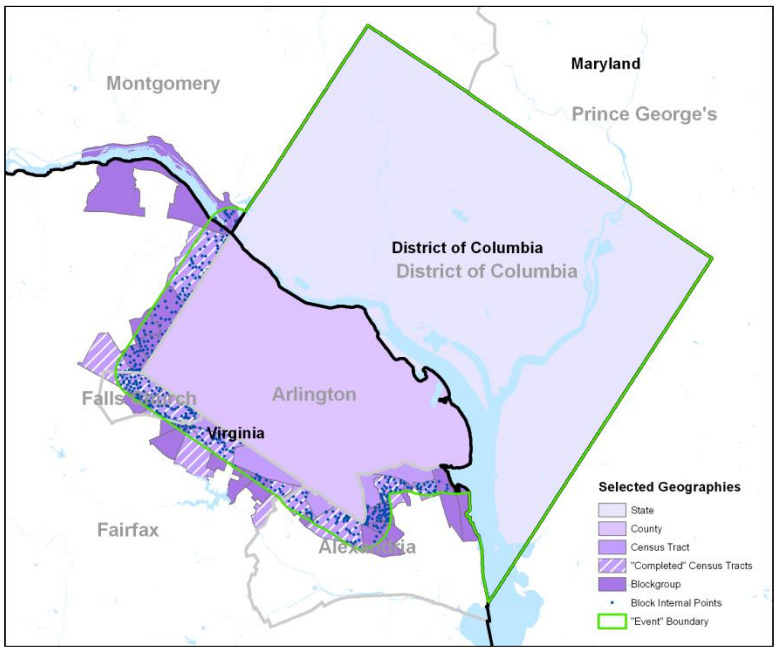

Figure 8

Finally, a cleanup pass is run in which the set of selected blockgroups make complete tracts. These are substituted for the blockgroups. Additionally, the set is checked to determine whether any counties or states have been completed with the addition of these last blockgroups. The set of finalized, largest

possible geography is shown in Figure 8 and this is the set that OnTheMap for Emergency Management would use to tabulate data if this were a real emergency event.

Table 8 shows a comparison between the different selection methods for this event area and Table 9 shows a comparison between the Blockgroup-only, mixed, and mixed-completed MOE calculations for this area. The full, final list of selected geographies is given in the Appendix.

Selection Method	Minimum	Middle (IPs)	Maximum
Blockgroup Count	660	695	706
Land Area [sq. mi.]	91	104	106
ACS Pop Estimate	878,878	931,160	944,785

Table 8: Selection Method Comparison for Example Boundary

MOE Aggregation	All Blockgroups	Combined Areas
Estimate	931,160	931,160
MOE	8,647.77	2,396.80

Table 9: MOE Improvement Comparison for Example Event

As with the results from real events, this complete example shows that the Middle selection method is much closer in blockgroup count, land area, and estimate to the Maximum method than to the Minimum method. Additionally, in Table 9, we see significant improvement in the estimate of the total population by combining smaller geographies to larger ones with improved MOEs.

Further Work

Sensitivity

One dynamic, not otherwise discussed here is the sensitivity of the estimates and derived MOEs to slight alterations in the event boundaries. Put another way, how much does it matter in the output statistics that an event was located in one place versus 500 meters to the East? Or how much does it matter that a snow event area was not defined 1 mile larger in diameter?

This is a question that is tied to both the event size and location as well as the size and distribution of the geographical units that carry the statistical data. Potential exists to calculate some form of "event sensitivity" and present it along with the derived MOE as an additional metric of how "good" the estimates are for a given event boundary. Further research and user testing would be required prior to any implementation.

A General Web Service

Finally, we propose a general purpose "geostatistical transformation service" that would provide automated approximation services in a formal web service. It would also provide wide access to recommended methodologies for approximating arbitrary boundaries and correctly aggregating standard errors for estimates to those approximations.

The minimum requirements for such a service are:

1. The technical ability to accept arbitrary input boundaries;
2. An explicit target dataset from which query results are expected to be drawn;
3. A standard set of reference geography that matches the target dataset; and
4. One or more approved transformation methodologies.

In the case of OnTheMap for Emergency Management, these specific requirements are met as:

1. The application gathers disaster event boundaries in standard GIS formats from trusted live data feeds;
2. ACS 2009-2013 5-year estimates is the current target dataset;
3. TIGER 2013 is the matching reference geography for this set of 5-year estimates; and
4. The Middle approximation method is the approved method in use for these ACS estimates in this context.

This proposed service could be implemented into new data dissemination and data analysis tools for both ACS data and other datasets produced by the Census Bureau.

The methodologies used in the tools described here along with the proposals for further work expand the usage of both ACS data and better statistical methods for handling derived margins of error for nonstandard geographies. It is also hoped that the explicit documentation of these methods will encourage data users to consider more closely the tradeoffs between spatial accuracy and data quality.

Appendix

Example Events

The events used in the examples were actual events displayed by OnTheMap for Emergency Management. These events were chose to provide a range of different query sizes and features. The specific events used were:

Flood area from March 25, 2015 centered on Paducah, KY. The event can be accessed in OnTheMap for Emergency Management at
http://onthemap.ces.census.gov/em/#d00ad0a16e77551ea5662bbe25bae2eb

Fire area from August 14, 2012 in California. "Wye Fire." The event can be accessed in OnTheMap for Emergency Management at
http://onthemap.ces.census.gov/em/#253bc3a0088c9a60a9a670446e09f2fd

Forecasted Snow from March 5, 2015 in the mid-Atlantic. The event can be accessed in OnTheMap for Emergency Management at
http://onthemap.ces.census.gov/em/#b224ec8816588cf80ecb052f373b6993

Geography List for Final Example

Table 10 lists the complete list of geography that is used for the middle method approximation of the example area including Washington, DC and Arlington, VA.

Geography Type	Unique GEOID
State (or equivalent)	11
County (or equivalent)	51013
Census Tract	51059451300
Census Tract	51059451400
Census Tract	51059451501
Census Tract	51059451502
Census Tract	51059452801
Census Tract	51059452802
Census Tract	51059470300
Census Tract	51059471000
Census Tract	51510200106
Census Tract	51510200107
Census Tract	51510200201
Census Tract	51510201000
Census Tract	51510201100
Census Tract	51510201203
Census Tract	51510201204
Census Tract	51610500100
Census Tract	51610500300
Census Blockgroup	240317058003

Census Blockgroup	510594503004
Census Blockgroup	510594504001
Census Blockgroup	510594516011
Census Blockgroup	510594516013
Census Blockgroup	510594516022
Census Blockgroup	510594527001
Census Blockgroup	510594701002
Census Blockgroup	510594709001
Census Blockgroup	510594709002
Census Blockgroup	510594709003
Census Blockgroup	510594709005
Census Blockgroup	515102001052
Census Blockgroup	515102008011
Census Blockgroup	515102009001
Census Blockgroup	515102009002
Census Blockgroup	515102009004
Census Blockgroup	515102012021
Census Blockgroup	515102018012
Census Blockgroup	515102018013
Census Blockgroup	516105002003

Table 10: Final Geography Selection for Example Boundary