Using GIS to Integrate Data from Incongruent Geographic Areas: Examples of Combining School District Data with Census Geography

(Draft Not For Citation)
September 15, 2004

Jana Chavers, Kate Dykgraaf, and Laura Nixon

Center for the Study of Inequality
and
Department of Sociology

College of William and Mary
P.O. Box 8795
Williamsburg, VA 23187-8795
Phone: (757) 221-2604
Email: sjsapo@wm.edu

We would like to acknowledge our faculty sponsors and supervisors Salvatore Saporito and Deenesh Sohoni. This research was funded by the American Educational Research Association which receives funds for its "AERA Grants Program" from the NSF and NCES (U.S. Department of Education) under NSF Grant #REC-0310268.

Combining Data from Incongruent Geographic Areas to Create a Common Unit of Analysis: Examples of Combining School District Data with Census Geography

Abstract

Demographers often work with data that summarize populations living in pre-set geographic areas (e.g., census areas and school districts). In some cases it is necessary to summarize data describing two different types of geography to one common geographic unit. This is not always an easy task as geographic areas often have incongruent boundaries. For example, it is not possible to summarize block-group population characteristics to school districts by simply summing population characteristics of block-groups that lie partially inside those school districts. To overcome this problem, we developed a technique (using Geographic Information Systems) to assign population weights to the characteristics of block-groups that lie partially within a second type of geographic area. These weighted block-group characteristics can then be summarized to school districts reliably. We demonstrate how this is accomplished and use a variety of data to assess the accuracy of the our geographic weighting method.

In this paper, we describe a method of assigning data describing characteristics of school districts with information describing census areas. In many cases, assigning information describing the characteristics of one geographic area to a second geographic area is straightforward. A simple example would be summarizing population information describing census blocks to larger geographic areas (e.g., zip codes) in which they are nested. One can easily locate census blocks neatly nested within larger areas and summarize block-level population figures to those areas.

Yet, there are many other examples in which the boundaries of one type of geography are not nested neatly within the boundaries of a second type of geography. In cases where the boundaries of one geographic type cross over the boundaries of a second geography type it is not possible to simply sum the population characteristics of one geography to another. A simple example of such incongruent geography illustrates why this is not possible: assume that we wanted to determine the poverty rate within a zip code using data describing census tracts. Portions of many census tracts lie only partially within a given zip code area. It would be problematic to simply assign all of the poverty information of overlapping census tracts to the zip code. However, it is reasonable to assign portions of the tract-level poverty data to a zip code but this requires two pieces of basic information: 1) identifying the portion of the census tract that is within a given zip code; 2) the number of people within the census tract/zip code overlap.

In this paper, we describe a method of joining incongruent data using Geographic Information Software (hereafter, GIS). We illustrate the method using two examples that have immediate applications for basic demographic research. In the first example, we assign information describing school districts (e.g., per-pupil expenditures in a school district) to Public Use Micro Areas (PUMAs) used by the Census Bureau. In the second analysis, we determine the

poverty rates of school attendance boundaries given the census block-groups that lie partially or totally within those attendance zones. For this second analyses we describe in detail the technical steps used with GIS. After we demonstrate our assignment technique with these two examples, we suggest a way of testing its accuracy by comparing known characteristics of geographic areas with characteristics that are assigned using our method. For example, we will correlate the known racial composition of school attendance zones (based upon complete count census data at the block-level) with racial composition of school attendance zones derived from our technique.

EXAMPLE ONE: ASSIGNING DATA FROM SCHOOL DISTRICTS TO PUMAS

The techniques we describe have immediate uses in two research projects we are now undertaking. In one of these studies, we estimate the probability that black, white, and Hispanic children attend private school given the racial composition of the areas in which these children live. Our data are derived from the 2000 Public Use Micro Data Sample (or PUMS Data) which describe social characteristics for individual children, including their age, race, and whether or not they are enrolled in private school. These data also identify the Public-Use Micro Areas (or PUMAs) in which these children live. This makes it possible to summarize racial data to PUMAs and use these aggregate data to assess whether the racial composition of PUMAs is correlated with the probability that a child will attend a private school. Our hypothesis is that the racial composition of a child's neighborhood is causally related to private school attendance.

This analysis calls for the inclusion of control variables that could theoretically mitigate any correlation between neighborhood racial composition and private school attendance. One critical control variable would be per-pupil expenditures in a school district. However, school district characteristics would need to be tabulated for PUMAs and this is not straightforward because there is no geographic congruence between them.

The Technique

In the United States there are 12,475 school districts with set geographic boundaries. Some districts encompass one or more towns or municipalities, and in some southern states they serve an entire county. In contrast to school district boundaries created by local and county governments, PUMA boundaries are created by the federal government based upon population figures. The Census Bureau created 2,071 PUMAs for the 2000 Census. Each covers an area that contains roughly 100,000 people and they do not necessarily correspond to any political geography. Although there are many more school districts than PUMAs they do not always nest neatly within PUMAs and, in some urban areas, several PUMAs may fit with a single school district. This makes linking school district boundaries with PUMAs a challenging process because the two sets of geography overlap in many different ways, as shown in Figure 1.

Figure 1 here

The left-most illustration in Figure 1 represents School District areas and the middle illustration represents PUMAs. When the two geographic areas are layered (shown in the right-most illustration) the result indicates school districts and PUMA are not congruent. In some cases an entire school district lies completely within a PUMA (as shown by Intersection I_{3A}). In most other cases, portions of school districts overlap with portions of PUMAs. Such incongruencies result in intersections such as I_{1A}.

To overcome this problem we assign a school district's per-pupil expenditure to a PUMA by taking the mean of per-pupil expenditures for any school district that is partially or totally contained within a PUMA boundary. (That is, we take the mean per-pupil expenditures of all intersections within a PUMA.) But, as we describe in detail below, we weight each school district's per-pupil expenditures by the number of elementary school-aged children residing

within each intersection. We then sum these weighted figures and divide by the total number of children living in a PUMA.

The first step in the process of taking a weighted average of school district information is to create a map representing the intersections of overlapping school districts and PUMAs. This is shown in the left-most illustration in Figure 2. The new map of intersections associates each area with the original school district and PUMA identification numbers. After creating a map of intersections, we match per-pupil expenditures to each intersection that is identified with a specific school district. Thus, any intersection that contains school district 1 is assigned per-pupil expenditures of $1,000 (as is the case with Intersections 1A and 1B). This is depicted in the right-most illustration in Figure 2.

Figure 2 here

Once per-pupil expenditures are assigned to each intersection, we weight expenditures by the number of school-aged children within each intersection. To determine the number of school-aged children in each intersection, we "overlay" the map of intersections on top of a map of block-groups. This process is depicted in the right-most illustration in Figure 2. Block groups are relatively small, usually comprising roughly five to eight city blocks.[1] As shown in the right-hand portion of Figure 2, there are 11 block groups within intersection "1A."Because the block group data from the 2000 Census summarizes all persons by age we are able to determine that there are 100 children aged 5 to 17 living in intersection 1A.

[1] Because block groups are relatively small geographic areas they fit neatly within our map of census area/school district intersections.

Once we determine the number of children in each intersection, we multiply per-pupil expenditures by the number of children in each intersection thereby weighting per-pupil expenditures by the number of school-aged children. In the example of Intersection 1A we multiply $1,000 by 100 resulting in a weighted per-pupil expenditure so $100,000. We sum the weighted per-pupil expenditures for each Intersection within a given PUMA area and, finally, divide this weighted sum by the total number of children residing within the entire PUMA. Using the example presented in Figure 2, we would sum the weighted totals of intersections 1A through 6A (which equals $1,080,000) and divide it by the sum of children living in all block groups within PUMA A (which equals 450). This gives a weighted average of per-pupil expenditures of $2,400 for PUMA A.

EXAMPLE 2: SCHOOL ATTENDANCE BOUNDARIES AND BLOCK GROUPS

We use the technique described above to accomplish a similar integration of poverty data to school attendance boundaries. In this instance, we have census data describing poverty rates in block groups and we want to summarize these data to school attendance boundaries. As with the example above, some block groups are not nested neatly within school attendance boundaries making it problematic to integrate poverty information.

In this example, we extend our demonstration by specifying how we created and integrated our maps and data with a GIS mapping software package called ArcView. (This demonstration assumes some basic knowledge of ArcView.)[2] We start with a simple map of some school attendance boundaries and block groups located in a section of Chicago, as shown in Figure 3. The map shows two "themes"–one for block groups (that contain our poverty data)

---

[2] We will show the process in detail in our poster session and will have a laptop available to show interested researchers how we complete every step.

and one of school attendance boundaries for elementary schools. We highlight one block group (with the identification number "1611002") that does not lie completely within any school attendance zone. As the figure shows, portions of the highlighted block group lie with Belding and Scammon school attendance zones.

Figure 3 here

Our next step is to create a single map that represents each unique school attendance zone/block group intersection. This is completed using ArcView with an extension called "xtools,"[3] as depicted in Figure 4. The figure shows the "intersect themes" function in the "xtools" extension. This function essentially creates a unique polygon shape for each individual overlap between school attendance boundary and block group.

Figure 4 here

The resulting map of the "intersect themes" function is shown in figure 5. This process produces a new map with unique intersections that contain the school identification name and the block group number for every intersection. For example, in Figure 5 the block "1611002" is part of two unique intersections–the first intersection with Belding and the second with Scammon (this can be seen in the inset of Figure 5).

Figure 5 here

After creating intersections, it is necessary to determine the number of children that reside within each of the them. At this stage, we add a map of blocks to the map of intersections, as depicted in Figure 6. As shown, every block lies almost entirely within an intersection. This allows us to geographically associate blocks with intersections, thereby permitting us to

---

[3] This extension can be downloaded from the Internet for free at: (http://www.odf.state.or.us/divisions/management/State_forests/XTools.asp)

determine the number children living within each intersection. The block-level geographic data (which now include intersections identified with school boundary and block group identifiers) are then matched in the number of children for each block (as displayed in Figure 7). We then aggregate block-level data to each intersection which produces us the number of children per intersection. The block-group poverty rate is multiplied by the number of children within each intersection, giving us a weighted value for the intersection. Finally, we sum these weighted figures and divide by the total number of children living in the school attendance boundary thereby giving us a poverty rate for each school attendance boundary.

Figure 6 and 7 here

Proposed Evaluation of Technique

We propose a straightforward method of evaluating data produced with our technique by comparing it with data that actually exist for the same geography.[4] We do this for both examples shown in this paper by comparing variables that we know to describe a given geography accurately with variables that are derived from our technique. (Of course, the variables we use come from "congruent geography," thus allowing us to assess the validity of our technique of combining incongruent geography.) In the first evaluation, we compare actual racial percentages for school attendance boundaries (based upon block-level information available from the 2000 Census) with racial percentages produced with our technique in which we combined block-group data with school attendance zones. In the second example, we compare actual racial percentages of PUMAs with racial percentages produced by using our technique to integrate school district level data with PUMAs.

---

[4] Preliminary results demonstrate the strength of our technique and we will show complete results at our poster session.

Summary

It is often the case that social scientists wish to create contextual variables that describe some areal unit such as school zones, zip codes, census tracts and other types of geography. Frequently, the variables that scholars want to use to describe one level of geography exist only on another level of geography. This creates the need to devise a technique that can accurately integrate data from incongruent areas. We have used to two illustrative examples to demonstrate how to combine data from one set of geography to another using Geographic Information Systems such as ArcView. This allows scholars to build data sets with variables derived from multiple sources in a reliable way.

Figure 1

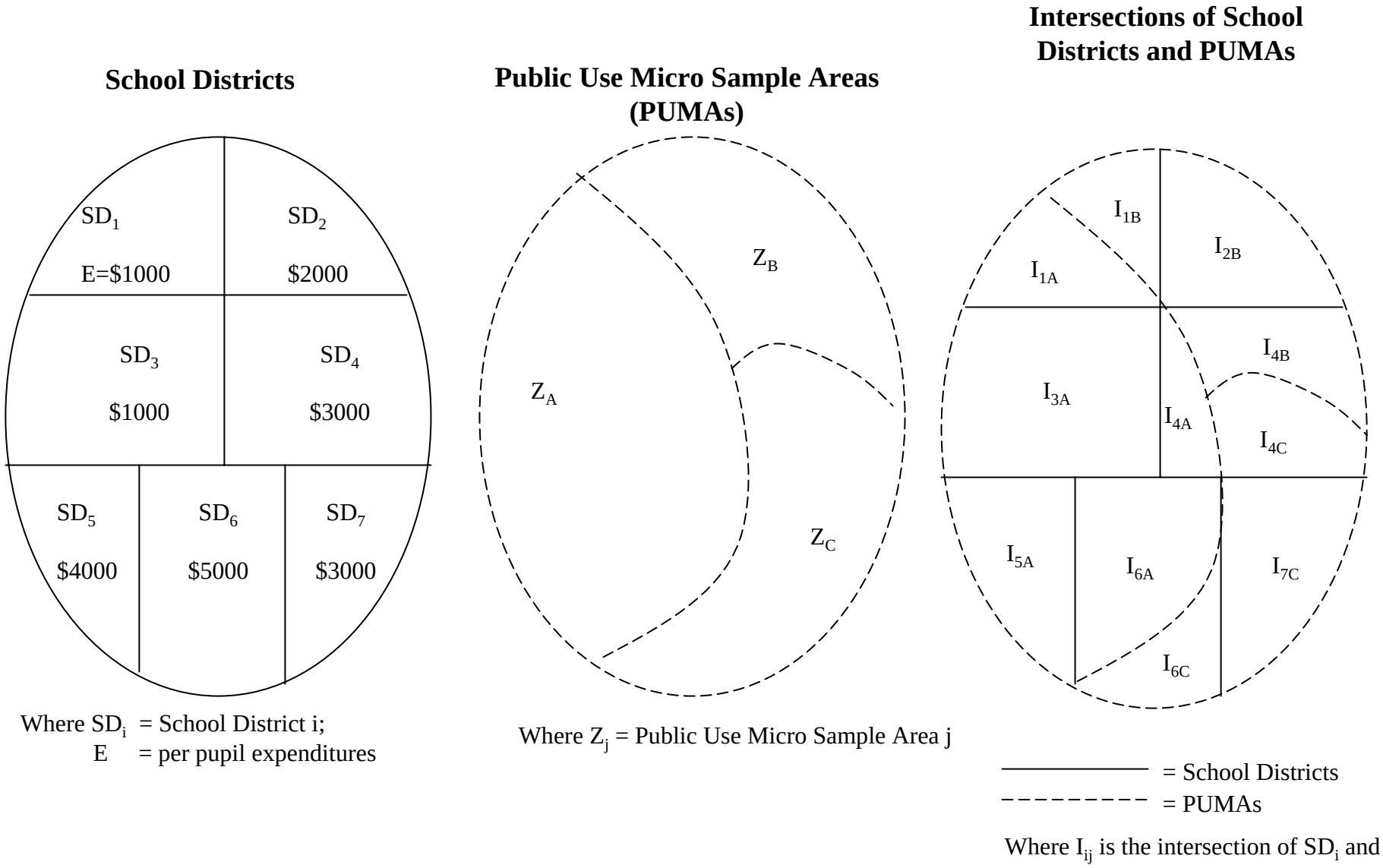

School Districts
SD1
E=$1000
SD2
$2000
SD3
$1000
SD4
$3000
SD5
$4000
SD6
$5000
SD7
$3000
Where SDi = School District i;
E = per pupil expenditures
Public Use Micro Sample Areas (PUMAs)
ZA
ZB
ZC
Where Zj = Public Use Micro Sample Area j
Intersections of School Districts and PUMAs
I1A
I1B
I2B
I3A
I4A
I4B
I4C
I5A
I6A
I6C
I7C
= School Districts
= PUMAs
Where Iij is the intersection of SDi and Zj

Figure 2

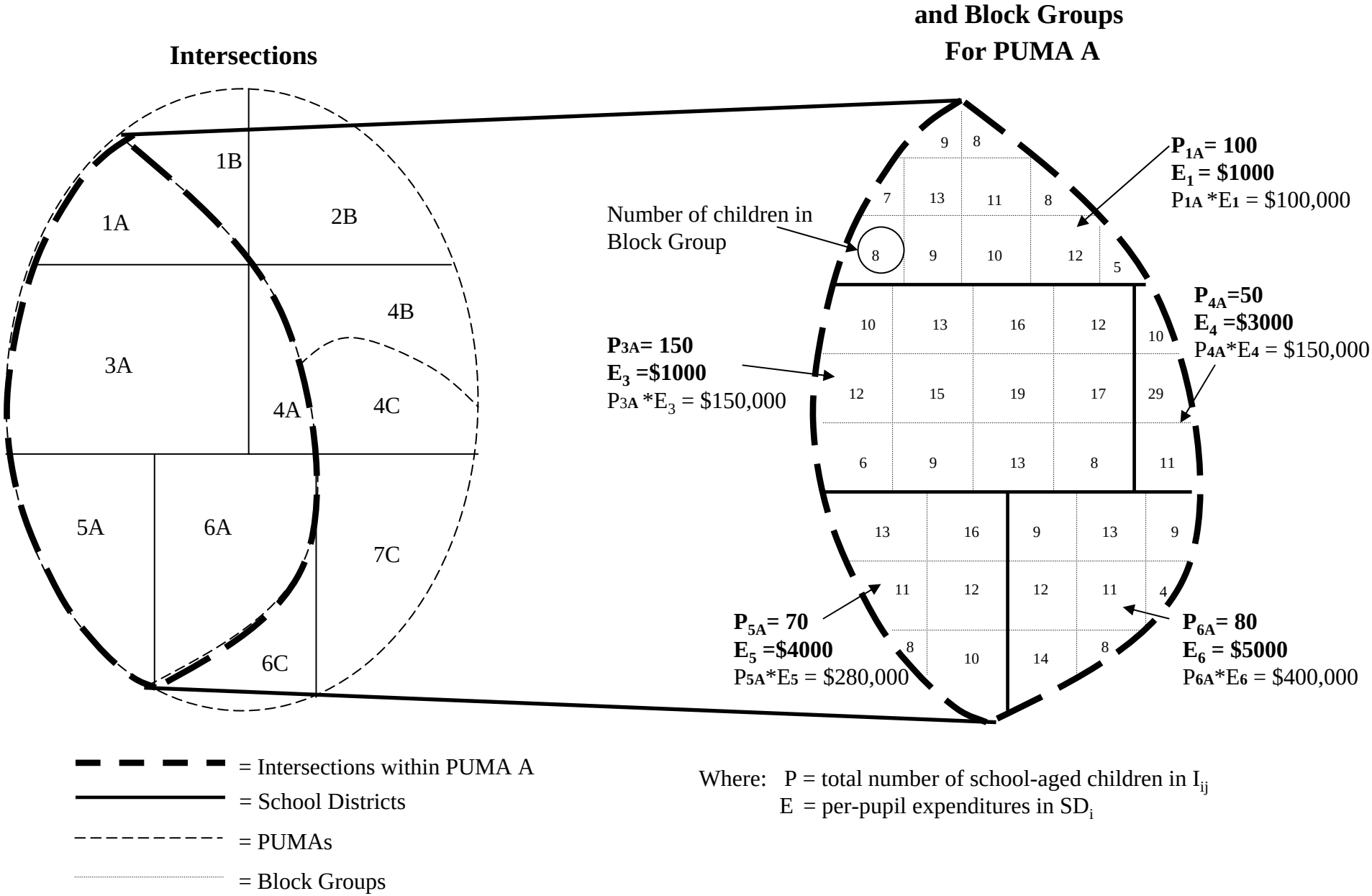
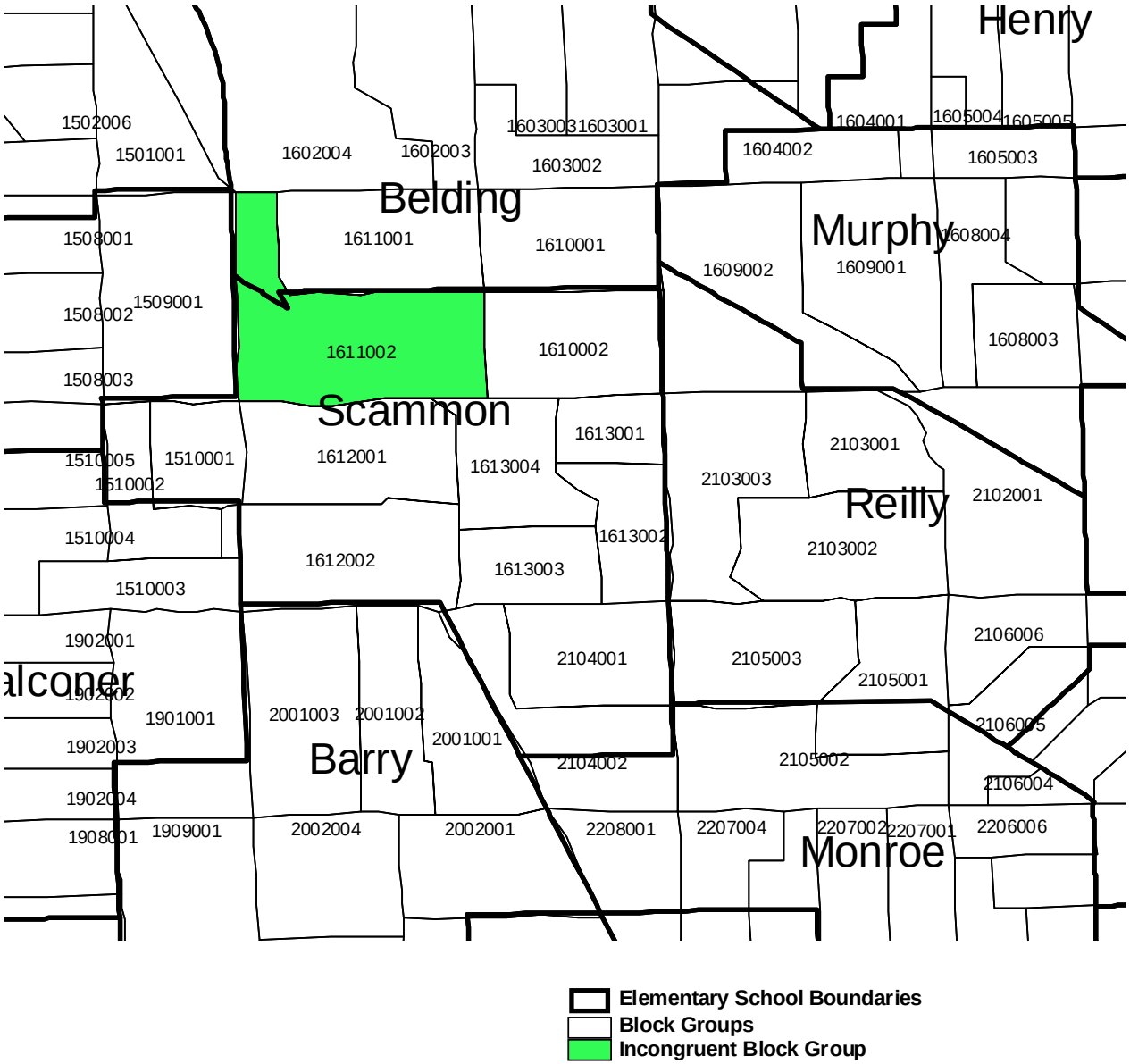

Intersections
1B
1A
2B
4B
3A
4A
4C
5A
6A
7C
6C
Intersections
and Block Groups
For PUMA A
Number of children in
Block Group
P1A= 100
E1 = $1000
P1A *E1 = $100,000
P4A=50
E4 =$3000
P4A*E4 = $150,000
P3A= 150
E3 =$1000
P3A *E3 = $150,000
P5A= 70
E5 =$4000
P5A*E5 = $280,000
P6A= 80
E6 = $5000
P6A*E6 = $400,000
= Intersections within PUMA A
= School Districts
= PUMAs
= Block Groups
Where: P = total number of school-aged children in Iij
E = per-pupil expenditures in SDi

Figure 3

Downtown Chicago:
Map of School Attendance Boundaries and Block Groups



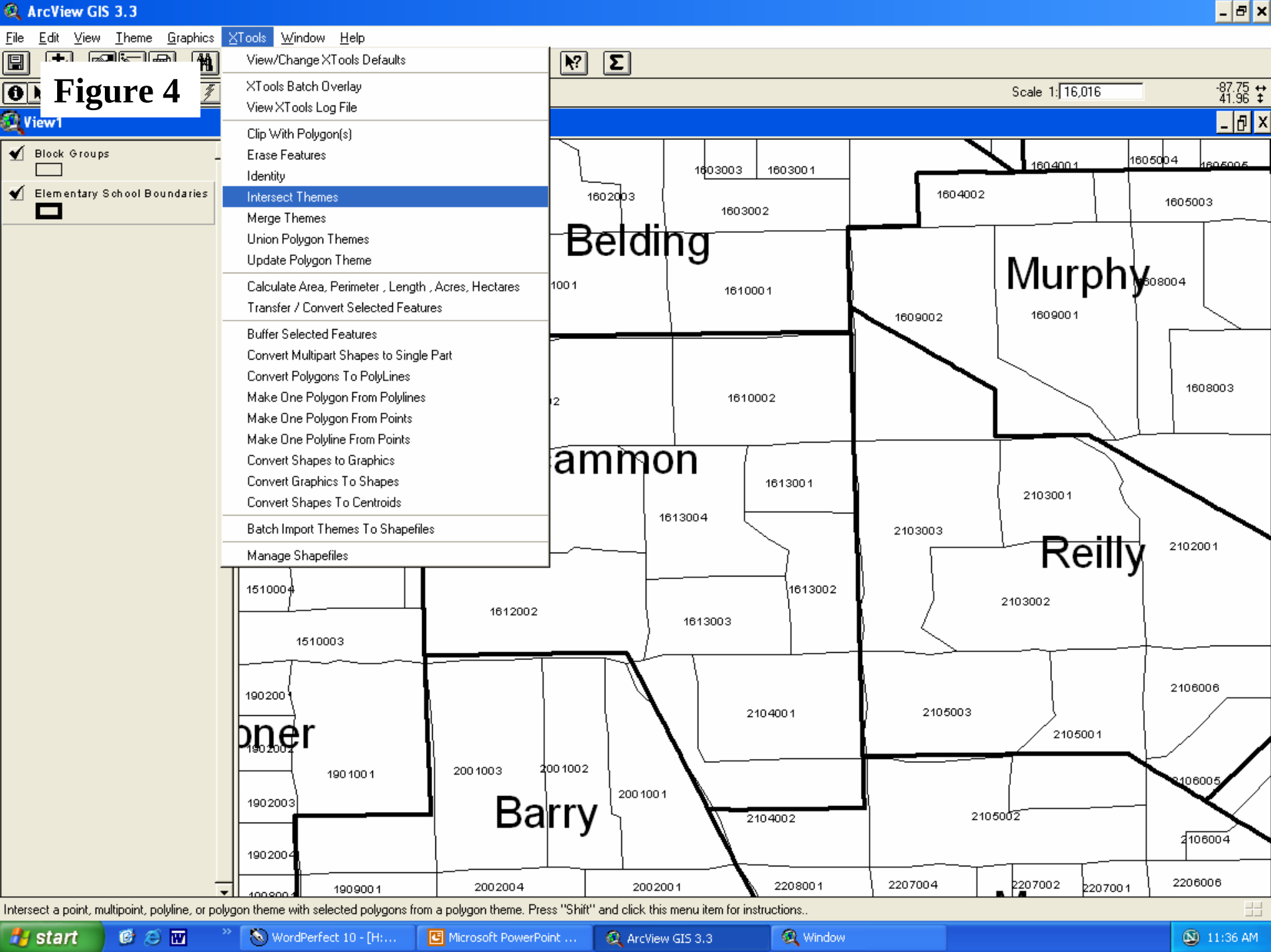

ArcView GIS 3.3
File Edit View Theme Graphics XTools Window Help
Figure 4
View/Change XTools Defaults
XTools Batch Overlay
View XTools Log File
Clip With Polygon(s)
Erase Features
Identity
Intersect Themes
Merge Themes
Union Polygon Themes
Update Polygon Theme
Calculate Area, Perimeter , Length , Acres, Hectares
Transfer / Convert Selected Features
Buffer Selected Features
Convert Multipart Shapes to Single Part
Convert Polygons To PolyLines
Make One Polygon From Polylines
Make One Polygon From Points
Make One Polyline From Points
Convert Shapes to Graphics
Convert Graphics To Shapes
Convert Shapes To Centroids
Batch Import Themes To Shapefiles
Manage Shapefiles
Scale 1: 16,016
View1
Block Groups
Elementary School Boundaries
Belding
Murphy
Reilly
Barry
Intersect a point, multipoint, polyline, or polygon theme with selected polygons from a polygon theme. Press "Shift" and click this menu item for instructions..
start
WordPerfect 10 - [H:...
Microsoft PowerPoint ...
ArcView GIS 3.3
Window
11:36 AM

Figure 5

Downtown Chicago:
Map of Intersections for School Attendance Boundaries and Block Groups

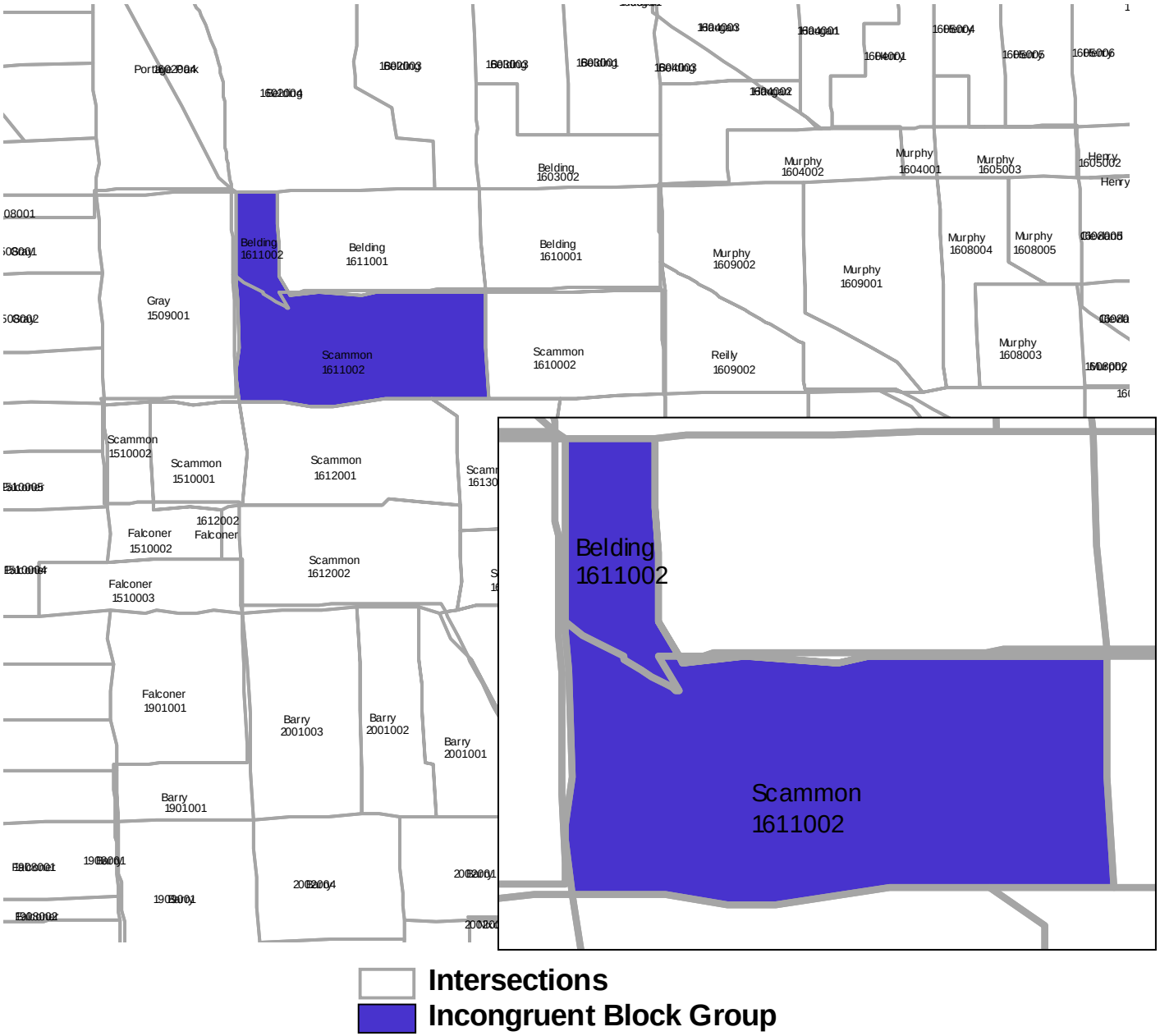


Figure 6

Downtown Chicago: Map of Intersections for School Attendance Boundaries and Block Groups with Census Block Layer Added

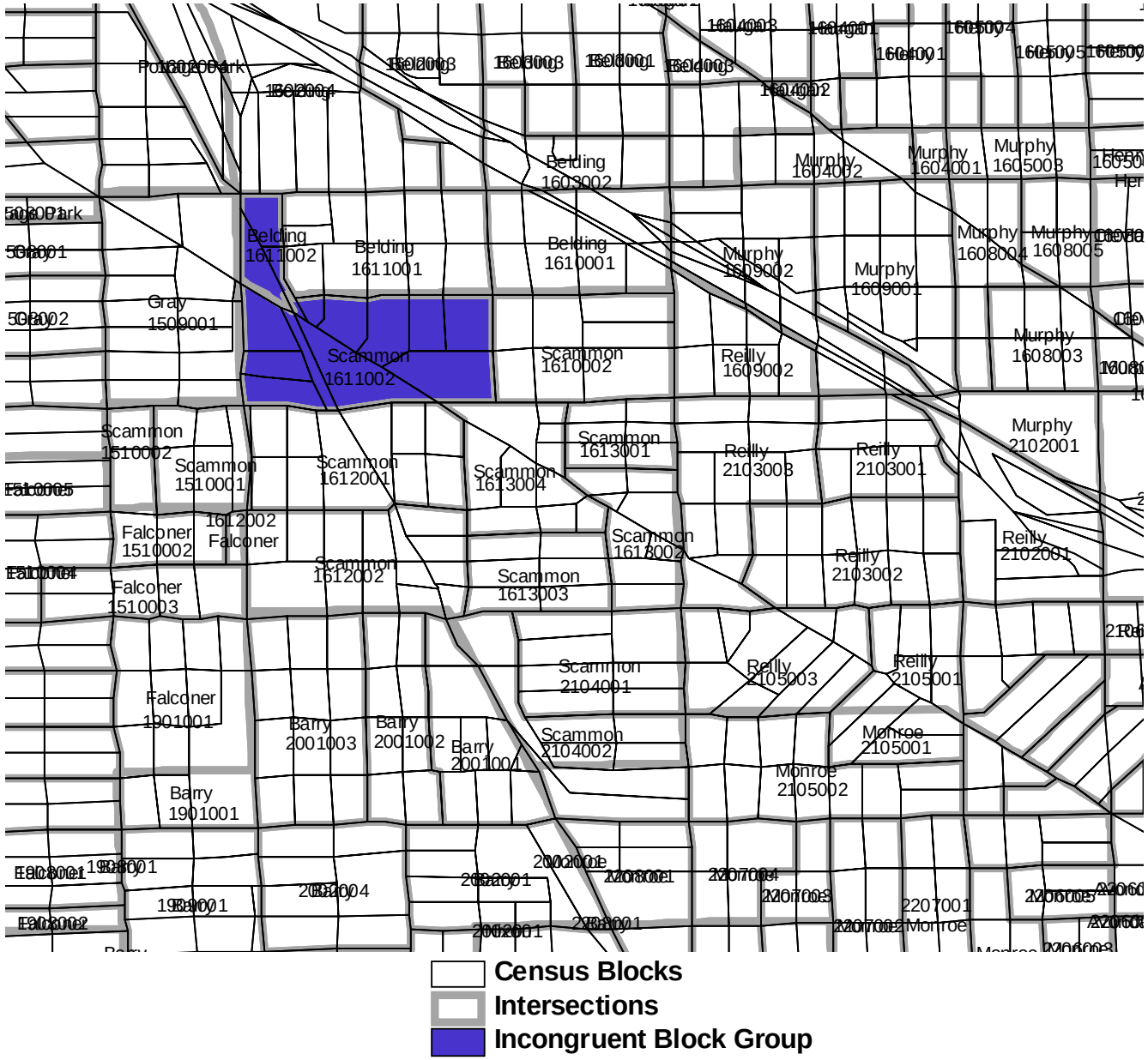
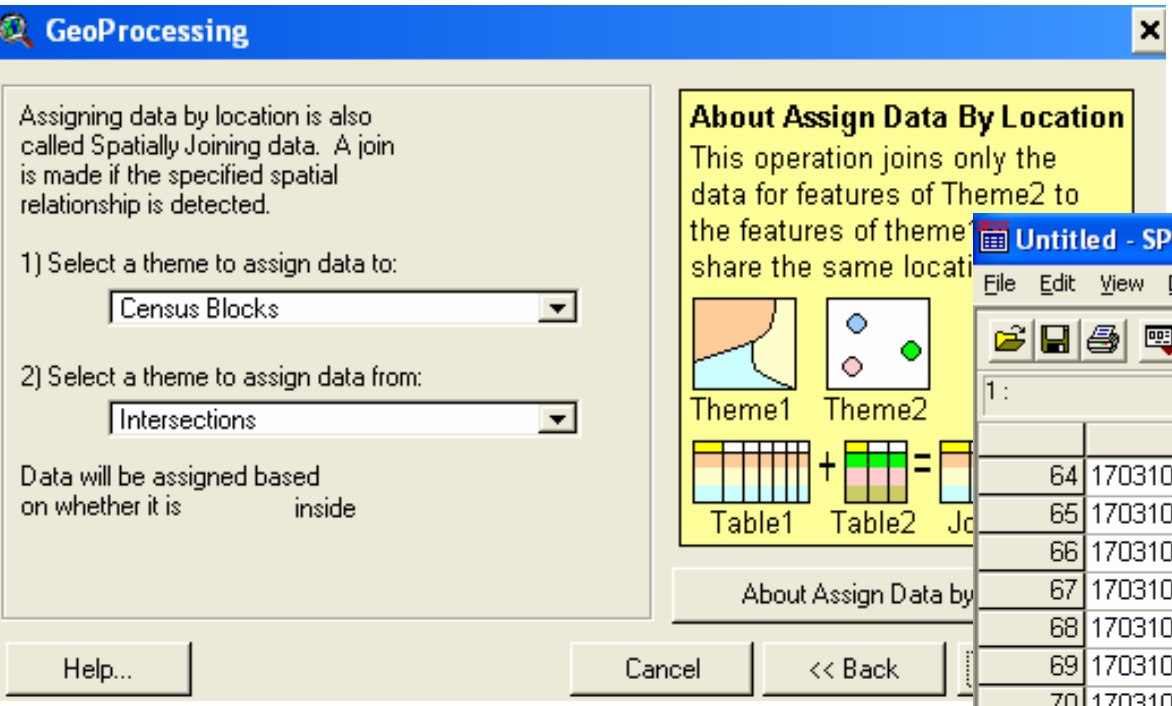
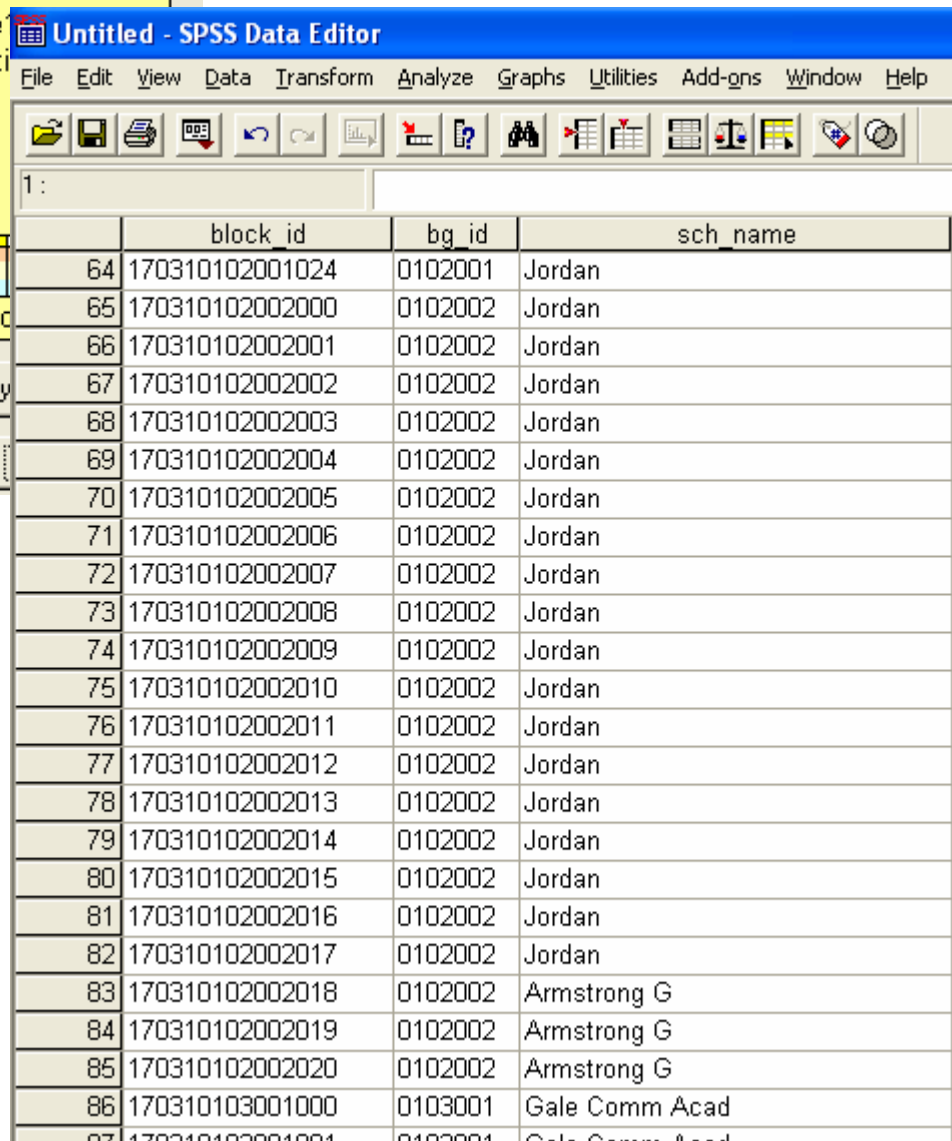


Figure 7





	block_id	bg_id	sch_name
64	170310102001024	0102001	Jordan
65	170310102002000	0102002	Jordan
66	170310102002001	0102002	Jordan
67	170310102002002	0102002	Jordan
68	170310102002003	0102002	Jordan
69	170310102002004	0102002	Jordan
70	170310102002005	0102002	Jordan
71	170310102002006	0102002	Jordan
72	170310102002007	0102002	Jordan
73	170310102002008	0102002	Jordan
74	170310102002009	0102002	Jordan
75	170310102002010	0102002	Jordan
76	170310102002011	0102002	Jordan
77	170310102002012	0102002	Jordan
78	170310102002013	0102002	Jordan
79	170310102002014	0102002	Jordan
80	170310102002015	0102002	Jordan
81	170310102002016	0102002	Jordan
82	170310102002017	0102002	Jordan
83	170310102002018	0102002	Armstrong G
84	170310102002019	0102002	Armstrong G
85	170310102002020	0102002	Armstrong G
86	170310103001000	0103001	Gale Comm Acad