Chapter prepared for Don Lafreniere, Ian Gregory, and Don Debats (editors),
The Routledge Handbook of Spatial History. Routledge UK

Developing GIS Maps for U.S. Cities in 1930 and 1940

John Logan (Brown University)
Weiwei Zhang (South Dakota State University)

This research was supported by research grants from National Science Foundation (SES-1355693) and National Institutes of Health (1R01HD075785-01A1) and by the staff of the research initiative on Spatial Structures in the Social Sciences at Brown University. The Population Studies and Training Center at Brown University (R24HD041020) provided general support. The authors have full responsibility for the findings and interpretations reported here.
John Logan is the corresponding author, Department of Sociology, Box 1916, Brown University, Providence RI 02912; phone 401-863-2267; email john_logan@brown.edu.

John R. Logan is Professor of Sociology at Brown University, where he has also been director of the initiative on Spatial Structures in the Social Sciences (S4) since 2004. He is co-author (with Harvey Molotch) of *Urban Fortunes: The Political Economy of Place*. He has been working with mapped historical census data on U.S. cities for several years. His Urban Transition Historical GIS Project (www.s4.brown.edu/utp) geocoded 100% microdata for 39 cities in 1880. His goal is to complete mapping for most large cities for every decade 1900-1940.

Weiwei Zhang is an assistant professor of sociology, director of State Data Center at South Dakota State University. Dr. Zhang completed her Ph.D. in Sociology at Brown University in 2014. Her research interests include residential segregation, ethnic neighborhood, demographic and spatial methods. Dr. Zhang has worked on developing new approaches for population estimation, geocoding, and spatial modeling. She is working on projects on assimilation and integration of Asian and Hispanic groups and historical settlements of immigrant groups in the US.

Developing GIS Maps for U.S. Cities in 1930 and 1940

Urban historians and historical geographers have a long tradition of mapping demographic data to study residential patterns, the assimilation or segregation of immigrants and minorities, and processes of neighborhood change, despite the difficulty of working from printed or microfilm copies of city directories and census manuscripts and drawing maps by hand. Dubois' study of Philadelphia was one of the earliest research of this type, including a detailed survey of the predominantly black Seventh Ward to depict the patchwork of poorer and more well to do blocks.[1] The early Chicago School sociologists used census data and data from many other sources to map the social characteristics of Chicago neighborhoods in the 1920s and 1930s. Radford (1976) plotted locations of black and white residents in 1880 in Charleston, distinguishing between those residing on streets, in backyards, and on alleys.[2] Rabinowitz (1978) mapped the streets block by block in four Southern cities to show the degree of racial segregation.[3] Groves and Muller (1975) similarly studied black residential concentrations in post-bellum Washington, DC.[4] Others have focused on white ethnic residential patterns in cities such as New York [5] and Detroit [6].

Historical GIS methods have combined with the digitization of census data from the late 19th and early 20th Centuries to unleash new possibilities for such research. This chapter focuses on methods that exploit digital databases and computerized mapping software to tackle similar issues. Such efforts have become widespread in recent years [7-14]. In the United States, census records for 100% samples of individuals are being made available in harmonized data files for several decades leading up to and including 1940 by the Minnesota Population Center (https://www.nappdata.org/napp/). This means that data can be aggregated easily into enumeration districts (areas smaller than contemporary census tracts) for any variables that were included in each census year. GIS maps are not readily available, but the materials required to create them (paper maps held by the National Archives, street maps for cities in various years, and written descriptions of enumeration district boundaries) are attainable.

In this chapter we begin by reviewing some recent analyses from the Urban Transition HGIS Project (www.s4.brown.edu/UTP) for the period 1880-1940 to illustrate the kinds of analysis that are now possible with mapped 100% samples of the census. We then deal with the concrete questions of how this kind of historical urban research is done – how to move from paper maps to GIS files that reflect a historically accurate street grid, how to determine the boundaries of census administrative areas, and how to transfer census data from computer files to the locations of specific addresses in a city. How is it possible to geocode the residences of virtually all the households in a city many decades ago? Some guidance is already available based on studies of 39 U.S. cities in 1880 [15] and 13 cities in the period 1830-1930 [16]. Here we describe in detail how we plan to develop a GIS database for 69 cities in 1930 and 1940.

Approaches to mapped data in the Urban Transition Project

We begin with a description of the Urban Transition Project. The initial step was to use the 100% samples from the 1880 census from the North Atlantic Population Project (NAPP) to map population characteristics in 39 U.S. cities. Relying primarily on city directories to provide address ranges on city streets, all addresses were geocoded, making available spatial information at a very fine level of resolution. One analysis relied primarily on aggregating population data to enumeration districts in order to study variations in the degree of residential segregation of white

ethnic groups in cities, and therefore it included all cities identified by the Census Bureau in 1880 [9]. The geocoded data were used to probe the relationship between an ethnic group's occupational pattern and residential location. A case that we gave special attention to is Buffalo, NY. Here as in many cities the most segregated ethnic group in was German. But Germans were also highly over-represented in several occupations (sawmills, wood products, and furniture making), while being under-represented in others (paper, printing, and publishing. The question was this: to what extent did occupational segregation contribute to residential segregation? The conclusion was that this effect was modest. Regardless of occupational sector, most Germans were located in a dense enclave east of the city center, while native whites were more widely spread closer to the waterfront. One strong concentration of German sawmill workers in an area north of the city included almost no native whites in the same industry.

Another study exploited data from 1880 in conjunction with similar data from 1900 through 1940 for two cities, New York and Chicago.[17] Here we began with the question of when the black population first became highly residentially segregated. We also asked why blacks lived in residential clusters – was it mainly due to sorting by race, or did other factors such as occupational standing or migrant status (Southern vs. local birthplace) contribute to their separation? In this study more extensive use was made of the flexibility in spatial scale that was provided by having data geocoded to specific building locations in 1880. We compared segregation at the level of city wards (the scale at which census data have previously been easily available), census tracts, enumeration districts, and smaller areas such as specific street segments or even individual buildings. One conclusion was that already at this time, when less than 5% of city residents were black, they were highly segregated by building and street segment. Further, at no spatial scale was their residential concentration attributable to the fact that they were predominantly working class, and there were only small differences between Southern migrants and local blacks. These findings suggest that the origins of black ghettoization were already in place before the turn of the century, decades before the Great Migration that many scholars have considered to be the source of ghettoization in Northern cities. Maps of the location of the black population were used to chart their movement and the expansion of existing black clusters over time. These provided a useful supplement to summary measures of segregation that documented the trend of increasing separation.

A third study expanded this analysis to ten major Northeastern and Midwestern cities for the period 1880-1930.[18] The microdata were drawn from the on-line index of all residents created by Ancestry.com for the decades 1900-1930 (these data will soon be in the public domain through the Minnesota Population Center). Maps were drawn for enumeration districts based on paper maps for each of these decades held by the National Archives. Segregation indices calculated from the aggregated microdata confirmed that in all but one of them the Index of Dissimilarity had reached the "very high" threshold of .60 by 1900 and was above .80 in four of them by that time. Maps for every city are included in the on-line appendix to this chapter and they show that in most cases the location of the eventual large black ghetto was already evident in 1880 or 1900. In this instance the mapped data serve as a supplement to conclusions reached from a non-spatial analysis of small area statistics.

The Urban Transition Project: 1930-1940

The public release of census records from 1930 and 1940 has created new opportunities for spatial analysis of population data from this time. The United States had recently become a

predominantly urban nation. The massive waves of international migrants had been interrupted by legislation in the early 1920s, and both the first and second generations of immigrants from Southern and Eastern Europe were establishing their place in cities. At the same time new migrant flows included African Americans' great migration from the South to Northern cities as well as Puerto Ricans heading in large numbers to cities like New York and Chicago. Data from the last two pre-World War II censuses provide rich new opportunities to study these groups' incorporation in urban America. The Urban Transition Historical GIS Project at Brown University seeks to add spatial information to the 100% sample of individual records that have been made available by the Minnesota Population Center's (MPC) ongoing Integrated Public Use Microdata (IPUMS) program. It will then be possible to aggregate data to neighborhoods at varying spatial scales in order to study processes of segregation, and neighborhood data can be combined with individual records to support multilevel analyses. In the longer term it appears that the methods used to create the 1880 and 1930-1940 street maps and geocoding can be applied to additional intermediate census years. It may be possible to have a complete mapped data set for many major cities that includes 1880 and every decade from 1900 through 1940.

Achieving this purpose requires an extensive mapping effort. Thanks to MPC's National Historical GIS Project (NHGIS) there already exists a 1940 tract map for those large cities where the Census Bureau had already defined census tracts. However these maps do not include the historical street grid, and they are of limited use for adding features at a finer spatial scale (enumeration districts, census blocks and street segments). The Urban Transition HGIS aims to create an accurate 1940 street grid for the largest 69 cities in the country, create new layers to represent enumeration districts (EDs) and blocks in both 1930 and 1940, and geocode the addresses of all households in these cities in both years. Figure 1 maps the cities. Even without the city names it is clear that they are most concentrated in the Northeast. But the Midwest and South are well represented, and major cities in the more sparsely populated West (such as San Francisco, Los Angeles, Denver, Dallas, and Houston) are also included.

Figure 1. Location of Cities in the 1930-1940 Mapping Project (n = 69)

These are ambitious goals, but they are feasible through a series of steps that take advantage of several different sources of information. We treat the project as a complex puzzle. There is no single source that provides all the necessary information, but there are ways to piece together bits of data from different sources to complete the puzzle. This chapter describes these steps in some detail. The purpose is partly to document the procedures for future users of the data, pointing out potential sources of error. We also hope that they will prove useful to other HGIS projects with similar goals. Other projects will have different information sources and different challenges in combining them, but they are likely to proceed through many similar steps.

Figure 2 summarizes the approach as a "recipe" for mapping and geocoding these cities. There were many useful sources of information from the Minnesota Population Center (MPC), its National Historical GIS Project (NHGIS), the Census Bureau, and search tools provided by a genealogy website (www.SteveMorse.org).

Figure 2. RECIPE: 1930-1940 GIS Maps and Geocoding of 69 Cities

Ingredients:
2012 GIS maps with current address ranges (Census Tiger files)
1930 100% microdata with address, ED and block ID (Ancestry/MPC)
1940 100% microdata with address and ED (Ancestry/MPC)
1940 street map of cities showing census block IDs (Census)
1940 tract GIS map of tracted cities (NHGIS)
1940 ED and block definition documentation (Census)
Crosswalk linking 1930 and 1940 EDs (StevenMorse.org)
Digitized list of streets by ED (StevenMorse.org)

Prepare the ingredients
Digitize and edit block definition documents
Standardize street names across all sources
Edit 2012 shape files to match the 1940 street maps
Fill in missing street names and house numbers

Combine and stir
Combine sources to create 1940 ED and block polygons
Compare address ranges within blocks for 1930 and 2012 to label many 1930 blocks
Confirm/edit 1930 blocks and derive a full address range every block
Apply the 1930 address ranges to the 1940 street and block map

Construct an ESRI address locator for 1930 and 1940
Run 1930 and 1940 addresses of all city residents through the ESRI geocoder
Document and serve to public through a web-based mapping system

We initially planned only to map cities in 1940. We had broken the problem down into quasi-independent components: 1) to create a 1940 street grid, 2) to develop a standardized list of street names, 3) to create polygons for physical blocks, census blocks, and EDs, 4) to organize and clean the geographic information in the census microdata, and 5) to array addresses along each street (geocoding). The last step could only be approximate: we could place residents on the right street and within the right ED, but we could only guess at the address range for each segment along that street within the ED. We didn't know which block they lived on.

We solved this problem by extending the project to 1930. The 1930 microdata include an extra piece of information that was not transcribed for 1940: the residents' block number. But there was less documentation for mapping in 1930, not even a paper map showing the location of blocks in any standardized form. So we knew people's block number but we didn't know the block's location. We describe below how we combined sources to overcome this obstacle. When we had mapped the 1930 blocks, we could geocode addresses with great accuracy. And having done this for 1930, we could then apply it to 1940.

Though we were led to 1930 for methodological reasons, having another decade of spatially referenced population data has important substantive consequences. First, it will be possible to ask how the composition of any given area, at any spatial scale, changed from 1930 to 1940 and what 1930 characteristics of the area might be considered to be predictors of change. Second, given the elapsed time of only one decade, it should be possible to link data for individuals from 1930, to ask who moved and where they moved, and to distinguish between residents of the area in 1940 who already lived there in 1930 from those who moved there post-1930.

The following sections describe each of the components of the mapping effort, including details on the sources that are used in each one. We draw examples from the city of Chicago, the city that we used to develop these procedures. At the time of publication of this book, the mapping process will still be underway. Additionally most likely we will have uncovered new problems or developed more effective solutions. Hence this chapter is more a report of a project in progress than its final documentation.

The 1940 street grid

The Census Bureau published street maps of major cities in 1940 as part of a series of publications that reported block-level data for each city [19]. The map of a portion of Chicago is shown in the left panel of Figure 3. Note that it identifies boundaries of census tracts and block numbers of blocks within tracts, but it does not identify enumeration districts (EDs) – combinations of blocks that are typically smaller than a tract. In principle an accurate 1940 street grid with a tract and block layer could be created through manual editing of a contemporary GIS street map of a city (from TIGER line files as shown in the right hand panel), using the 1940 map images as a reference.

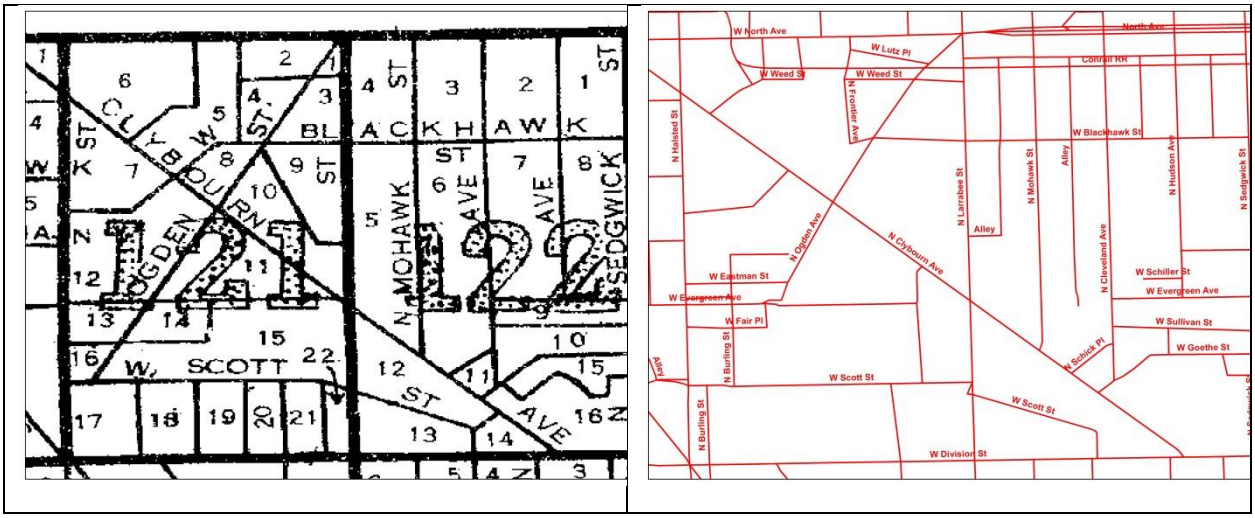


Figure 3. A 1940 block map produced by the Census Bureau (on left) and the 2012 GIS street grid (on right) for a portion of Chicago

The first step in linking these maps is to scan the 1940 map, add it as a layer on the 2012 map, and georeference it. Georeferencing involves identifying some points (typically intersections of major streets) on the scanned map that are known to be the same on the GIS map. After georeferencing the relationship between features in each layer is clear even when looking at them side by side, as in Figure 3. When available in color with one layer superimposed on the other, it

is evident that most streets line up very well even though it is not possible to create an exact correspondence. Differences between the layers are also easy to see: 2012 streets that did not exist in 1940, 1940 streets that are missing in 2012, and the same street with a different name in the two years. In this section of Chicago, for example, West Lutz Place and West Weed Street are found in the upper left quadrant of the 2012 map but not in the 1940 map. West Blackhawk extends on both sides of Clybourne in 1940 but only on its west side in 2012.

Editing the contemporary map backwards to match the 1940 street grid was time consuming. Fortunately it could be completed by undergraduate student research assistants with little training. The editing process preserved information about street segments, such as directionality and address ranges in 2012. The most frequent change was to remove a 2012 street ((including highways and their associated on and off ramps) that did not exist in 1940. In cases of a name change, we adopted the name shown on the 1940 map. Where a name was missing on the 1940 map (e.g., several short north-south streets in the southwest quadrant of Figure 3), we initially applied the 2012 name, which had to be confirmed in a later step. But note that in some of these cases the street was also missing from the 2012 map and the name had to be found in another way. The 1930 and 1940 microdata (searching within EDs) provided candidate names, for example.

Though not shown in Figure 3, the edited 1940 street grid includes other features that are often used to define administrative boundaries, such as the city limits, railroads, and rivers. These were assigned the names that were found on the original 1940 census map and treated as though they were street segments.

The standardized street list

A key concern in creating the 1940 street grid was to maintain standardized street names. This is essential because we collate information from several data sources, and the street names (their spelling or misspelling and the abbreviations used) vary greatly across sources. Is it East 5 St, E Fifth St, or East 5th Street? Was there an S Boardway St in Chicago in 1940? If we change Boardway to Broadway should the full street name be written as S Broadway, S Broadway St., So Broadway Street, or some other variation? Different sources often follow different formats. Uniformity is essential. Achieving it requires procedures that are sometimes referred to as data mining – in cases where the name is nearly unrecognizable (e.g., Bdrwy), we must make an informed inference about the name based not only on the sequence of characters in the name, but also on its location in the city (by ED or tract), which limits the choice set.

We relied on three main sources in order to create a standardized street list. All three sources are available in digital form.

1. Street names from 1930 and 1940 microdata. The transcription of street names by Ancestry.com includes many potential spellings of the same name. However these are the streets that must be on the GIS map in order to geocode residents. The 1930 and 1940 files both identify the ED within which people living on a given street may be found, and we created lists of street names by ED. Because street names were transcribed with no use of naming protocols expected by a GIS (such as including directions, names, and street types in a standard order), these names required extensive cleaning prior to their use. Initial cleaning of street names, though partially automated (making many changes through what are called “regular expressions”

in STATA), was the most labor-intensive part of the project. Every city presented slightly different problems, and an average city could require forty hours to do this initial cleaning even before comparing to other name lists.

2. StevenMorse.org website. Another valuable resource is a website that provides tools mainly to genealogists (www.StevenMorse.org). Among these tools is a listing of all streets found in every 1930 and 1940 ED in major cities. Our experience using this source is that it has a much higher degree of accuracy and consistency in spelling and completeness of names than do the microdata from Ancestry.com. We have been fortunate to obtain the full database that it draws upon (transcribed from original sources by well-trained volunteers).

We compared the microdata and SteveMorse lists within EDs, which greatly reduced the number of possible matches that needed to be evaluated.

3. 2012 GIS map. The 2012 map includes many streets that did not exist in 1930 or 1940. For streets that remained the same, however, it has the advantage that spelling is very uniform and the format of names has already been standardized, including a direction (such as East or South), a name, and a street type (such as Street or Avenue). Therefore the 2012 street list supported many corrections in names. We created a master list of street names from these sources in a standard format including [direction] [street name] [type]. To this we added – where possible – the ED and tract that the street should be found in.

One purpose in standardizing names was to compare which streets were listed in each source. We discovered that the 1940 map from the Census Bureau was incomplete (some streets where people were listed as residing in 1930 or 1940 were not included in SteveMorse.org or block description files). For example, the 1930 microdata included people on streets that did not exist on the 1940 map, but could be found in 2012. It was important, therefore, to retain those streets when creating the 1940 GIS street grid.

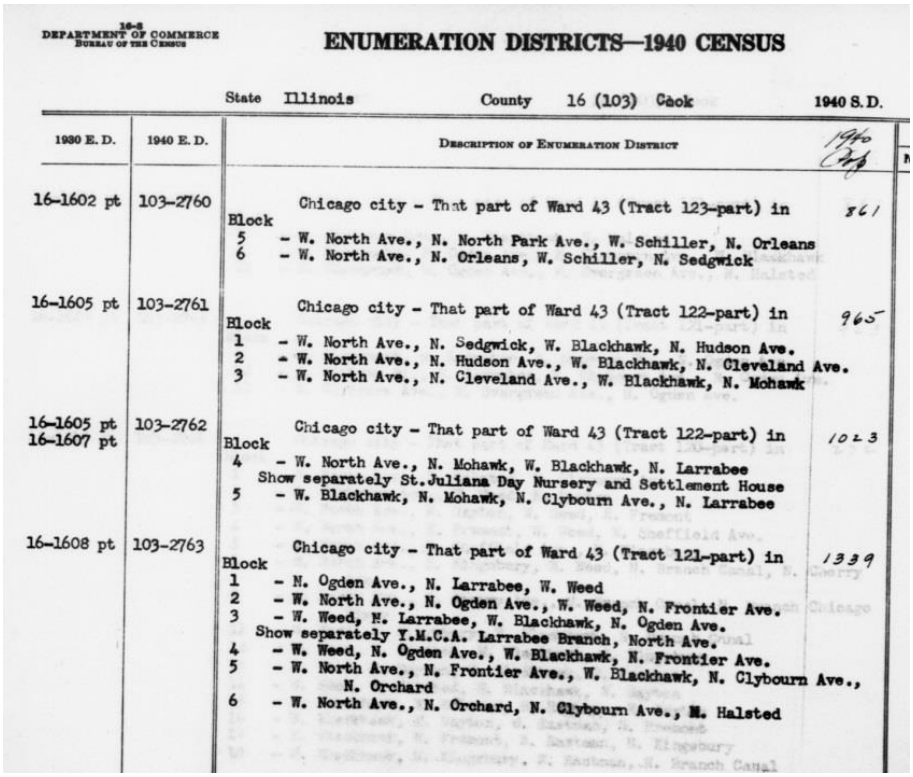
Many street names – especially from the microdata – could not be matched to a street name in another source. This was usually because they were spelled too badly to make a good match (or they included stray characters such as "??"). At this stage we kept these unmatched names in our master list and corrected (imputed) them (often manually) at a later point. We also used the master list to correct street names in another kind of file: 1940 block definitions from the census bureau as discussed below. These files list the boundary streets for 1940 blocks, and also provide the 1940 ED and tract where the block was located.

Mapping blocks and EDs

At this stage we are working with a 2012 GIS street grid that has been edited to match the 1940 census map features, with a partially standardized list of street names (and features such as rivers and railroads), and a layer identifying 1940 census tracts. The next step is to add 1940 block and ED layers to the map.

We automatically drew physical blocks (polygons bounded by streets, rivers, or railroads) using the "features to polygon" tool in ArcGIS. Physical blocks based on the street grid are not necessarily "census blocks" and they do not have census ED or block numbers. We learned these from block definition documentation provided by the Census Bureau (http://www.archives.gov/research/census/1940/finding-aids.html#desc). This documentation

lists all 1940 EDs in major cities and includes the block number and the boundary streets (or other geographic features used as boundaries) for every block in the ED. A portion of a page of block definitions is reproduced as Figure 4.



16-2
DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

ENUMERATION DISTRICTS—1940 CENSUS

State Illinois County 16 (103) Cook 1940 S.D.

1930 E. D.	1940 E. D.	Description of Enumeration District
16-1602 pt	103-2760	Chicago city - That part of Ward 43 (Tract 123-part) in 861 Block 5 - W. North Ave., N. North Park Ave., W. Schiller, N. Orleans 6 - W. North Ave., N. Orleans, W. Schiller, N. Sedgwick
16-1605 pt	103-2761	Chicago city - That part of Ward 43 (Tract 122-part) in 965 Block 1 - W. North Ave., N. Sedgwick, W. Blackhawk, N. Hudson Ave. 2 - W. North Ave., N. Hudson Ave., W. Blackhawk, N. Cleveland Ave. 3 - W. North Ave., N. Cleveland Ave., W. Blackhawk, N. Mohawk
16-1605 pt 16-1607 pt	103-2762	Chicago city - That part of Ward 43 (Tract 122-part) in 1023 Block 4 - W. North Ave., N. Mohawk, W. Blackhawk, N. Larrabee Show separately St. Juliana Day Nursery and Settlement House 5 - W. Blackhawk, N. Mohawk, N. Clybourn Ave., N. Larrabee
16-1608 pt	103-2763	Chicago city - That part of Ward 43 (Tract 121-part) in 1339 Block 1 - N. Ogden Ave., N. Larrabee, W. Weed 2 - W. North Ave., N. Ogden Ave., W. Weed, N. Frontier Ave. 3 - W. Weed, N. Larrabee, W. Blackhawk, N. Ogden Ave. Show separately Y.M.C.A. Larrabee Branch, North Ave. 4 - W. Weed, N. Ogden Ave., W. Blackhawk, N. Frontier Ave. 5 - W. North Ave., N. Frontier Ave., W. Blackhawk, N. Clybourn Ave., N. Orchard 6 - W. North Ave., N. Orchard, N. Clybourn Ave., N. Halsted

Figure 4. Block definitions for a portion of Chicago in 1940

In order to draws census blocks and EDs, we needed to convert these images into accurate digital files. We used an OCR program (FineReader), yielding a transcription that requires further editing (shown in Figure 5). We extracted and manually edited the lines that list the 1930 ED number (beginning with 16-) and 1940 ED number (beginning with 103-) for a set of blocks, along with a ward and tract number that those blocks are found in. These numbers were essential for geographic identification. We also corrected block numbers (such as ^ 4 changed to 4).

16-1602 pt 103-2760 Chicago city - Thit part of Ward 43 (Trsct 123-part) in xC /
5 - W. North Ave., N. North Park Ave., W. Schiller, N. Orleans
6 - W. North Ave., N. Orleans, W. Schiller, N. Sedgwick
16-1605 pt 103-2761 Chicago city - That part of Ward 43 (Tract 122-part) in
1 - W. North Ave., N. Sedgwick, W. Blackhawk, N. Hudson Ave.
2 - W. North Ave., N. Hudson Ave., W. Blackhawk, N. Cleveland Ave.
3 - W. North Ave., N. Cleveland Ave., W. Blackhawk, N. Mohawk
16-1605 pt 103-2762 Chicago city - That part of Ward 43 (Tract 122-part) in /o 2- 3
^ 4 W. North Ave., N. Mohawk, W. Blackhawk, N. Larrabee
Show separately St.Juliana Day Nursery and Settlement House
5 - W. Blackhawk, N. Mohawk, N. Clybourn Ave., N. Larrabee
16-1608 pt 103-2763 Chicago city - That part of Ward 43 (Tract 121-part) in J3 3^
ij 1 - N. Ogden Ave., N. Larrabee, W. Weed
| 2 - W. North Ave., N. Ogden Ave., W. Weed, N. Frontier Ave.
3 - W. Weed, N. Larrabee, W. Blackhawk, N. Ogden Ave.
Show separately Y.M.C.A. Larrabee Branch, North Ave.
4 - W. Weed, N. Ogden Ave., W. Blackhawk, N. Frontier Ave.
5 - W. North Ave., N. Frontier Avs., W. Blackhawk, N. Clybourn Ave.,
j 6 - W. North Ave., N. Orchard, N. Clybourn Ave., M. Halsted

Figure 5. FineReader product of the sample image in Figure 4

We created a python program to compare the street names in a given ED in the block definition file with corresponding names in the standard street list. The set of possible matches was greatly reduced by being able to limit the search to streets in a specific ED instead of having to consider all street names in the city. Like other standard data mining procedures, the matching program relied on calculating (for every name in one file in comparison with a potential matching names in the other file) the number of matching letters and the number of letters found in the same sequence. Some street names were unrecognizable, but a large share could be matched and corrected.

Another python code automated drawing polygons that are bounded by these listed streets with standardized names. In the majority of cases these polygons were identical to a physical block, and in these cases we assigned the ED and block number of the polygon to this block. In some cases more than one physical block was linked to a census block, and we merged them.

Manual editing was required to confirm block numbers or (for unlabeled blocks) to discover them. Editing was facilitated by having multiple sources of information. Within the area of a 1940 ED we knew what 1930 block numbers should be found. We also knew which block numbers should be near one another because they were part of the same 1930 ED. There is also a pattern in the way block numbers were originally assigned by the Census Bureau, so that usually consecutive block numbers are found adjacent to each other, following a spatial sequence (often clockwise) within an ED. Consequently it was often a simple process of elimination to fill in a missing block number or a short series of block numbers. However it was sometimes necessary to refer back to the original block definition page image to read the boundary streets for a given block or to check the list of populated streets in the 1930 microdata. Finally the correct ED and block number were entered into the attribute table. Note that once blocks were correctly labeled, they could easily be aggregated into EDs and tracts because ED and tract ID numbers were assigned from the block description file or manual editing.

Adding addresses from the microdata

At this point we have constructed a historically accurate GIS street shapefile with layers for the 1940 labeled blocks, EDs, and tracts. The next step is to add information from the 100% microdata, and to place addresses on the map. The U.S. 1940 full count census data include all individuals enumerated in the census with the person's name, age, gender, race, marital status, highest grade completed, place of birth, occupation, and income (https://usa.ipums.org/usa/voliii/items1940.shtml). Housing characteristics include whether the home is owned or rented and home value or monthly rental cost. Information on each person's relationship to the head of household is the basis for describing various aspects of household composition. Household identifiers also make it possible to determine the composition of the whole building in cases where there is more than one household at a given address. The address is provided as a street name and a house number. Other geographic identifiers include the state, county, city, ward, tract, and ED. The original census also includes a block number, but the block number has not been transcribed – a serious omission given our intention to geocode addresses.

There are many kinds of problems in the transcribed street addresses in the file provided by MPC. The street name may be completely missing (the field may be blank or coded as "???"), often because enumerators or transcribers omitted it or expected the user to assume that the street name previously listed would continue for subsequent households on the same page or next page. The house number is often missing. It may also have a value that is out of range for that part of the city. For example a Chicago address is transcribed as 3417 W Scott St in ED 2763; no other address on W Scott St in ED 2763 is larger than 400. Sometimes the information coded in the house number or street name field refers to some other geographic feature (e.g., the name of an apartment building, hotel or boarding house) that may have an address embedded in it (e.g., *1250 South Broadway Apartments*).

Many such errors could be corrected by checking the original census manuscript, which is readily available on-line. However in a project dealing with millions of records, this is impractical. Instead, based on spot checking a non-random set of apparent problems, we have developed standard cleaning procedures.

1. Extracting the street name. The "street name" field sometimes contains extra information. For example, it may include a word like "Cont" (presumably short for "continued"). In this case, we consider the record to have the same street name as the previous record. The street name field sometimes contains house numbers. This situation happens often for apartment complexes, where the numbers in the "house number" field are actually apartment numbers and the real house number is found in the street name field. We used regular expressions in STATA to parse these variables, looking for specific words (e.g., "apartment," "hotel") in the street name and then re-assembling the information.

2. Carrying forward a street name. Some addresses have valid house numbers but no street names. Often the street name for the household on the previous line should be carried forward. We did this under two conditions. First, we borrowed street names only from the same enumeration page. Second, the adjacent cases should not have a large skip in the house number (after experimentation we set this skip at not greater than 6. We also took into account the distinction between odd and even house numbers, assuming that the enumerator generally stayed on the same side of the street when moving from building to building. Each time that a street

name is carried forward this way, we update the file and move ahead to the next missing name. Sometimes the same name is carried forward several times on the same page.

3. Cleaning house numbers. There is considerable variation in the contents of the house number field, and these need to be standardized before we turn our attention to numbers that are entirely missing. The following invalid fields were all recoded to missing values that needed to be filled in by other means.

a. A continuation of the previous house number indicated by "continued", "con", "con't" etc. in the text.
b. A location nearby the previous house indicated by "1/2", "basement", "front", "back", "rear", "top", "bottom", etc. in the text. These are recoded to missing numbers except when there is a new house number within the text. For example, 175rear is recoded as 175. We extract and store the extraneous text in a new variable and keep only house numbers.
c. A different level in the same building indicated by "floor", "[0-9] 1st", "1F [0-9]" etc.
d. An apartment indicated by "Apt" in the text.
e. A miscellaneous group indicated by "[0-9][ ][a-zA-Z]", "[a-zA-Z][-][0-9] " etc. in the house number variable. The uniqueness of this category is that there is no other text or number except a single letter and a single number, sometimes with a space or a dash sign. This category is most likely the room in a hotel, like 9c, a5, 7-B.

4. Dealing with missing house numbers. Missing numbers will be dealt with in a similar way to missing street names, except that in addition to carrying forward we also interpolate numbers. Some house numbers are suspicious and need to be removed from consideration in this process. For example some house numbers are far outside of the logical possible range for a particular street segment and we wish to consider them as outliers (i.e., transcription errors). To identify these outliers, we compare all house numbers of the addresses on the same street in the same ED. The distribution of these numbers tells us the reasonable range for the segment of that street. This reasonable range can be predefined by us depending on prior knowledge about the size of an ED in a particular city. These "suspicious" cases would otherwise mislead us in future steps.

The logic of interpolation is to borrow house number information from neighboring households on previous and subsequent lines. We treat renter households with a missing house number as living at the same address as the preceding household, so the house number can simply be carried forward. (For example in institutions like hotels and boarding houses, there may be many households listed, but only the first one carries a house number.)

We believe this is less likely if the household is identified as a home owner, because condominium ownership was rare in this period, and we expect at most one resident owner per building. In these cases we add a house number with the same parity (odd or even) based on interpolation (out of caution, we do this only among addresses that are listed on the same page and are on the same street). There are a few caveats. Sometimes there is no number between the previous and next neighbor addresses that can be used. For example, the previous address has a number 132 followed by an address that has no number and then an address 134. In this situation, we must assume the missing address has the same number as the previous one even if it means two "owner" households are listed at the same address. When there are multiple households that have no numbers, we assign a separate number for each one.

Drawing the 1930 map in order to geocode 1940 addresses

The final step is to assign addresses to locations. One approach would be use a contemporary geocoding engine. That is likely to be effective for many addresses in many cities, but with an unknown reliability. We wish to have more certainty based on period information. If the 1940 block number had been transcribed by Ancestry.com it would have been a simple matter to place addresses on the proper street segment and side of the street to be on that block, and to array them in the correct order along the segment. But the smallest geographic unit that we have available to place addresses in 1940 is the ED.

Working at this scale has become our "fallback" geocoding procedure. Let us define the length of a street that falls within a given ED as an "ED segment." It could be a single block long, or it could extend several blocks but typically not more than three blocks. The information that we have assembled up to this point allows us to place addresses on their ED segment in the correct order. The ambiguity in this procedure is that we don't know which block the address is on, so its position along the street is arbitrary. We divide addresses along a street equally among the street segments in a given ED and space them evenly within the street segment. When a street is a boundary between two EDs, we align addresses independently on either side of the street. This means, for example, that 2147 can fall between 2120 and 2140, because it is in a different ED.

For many purposes this placement is acceptable (and more useful than if data had to be aggregated to the ED level). It is approximately accurate at the scale of the ED segment, and we will apply it when we cannot improve it. However in most cases we can do better by taking advantage of the 1930 full-count microdata file that includes not only addresses, ED and tract numbers, but also block numbers. If we assume that an address that lies in a given block in 1930 can be found at the same location in 1940, this additional information should be able to inform our 1940 geocoding. Our approach is to draw the 1930 block map, geocode addresses in 1930, then apply the same address ranges to 1940. If the result does not contradict other known information (e.g,, such as being placed in the wrong ED) we accept it as correct. We have no additional way to confirm it.

Although we could not locate an original census block map for 1930, we could exploit the progress that we had already made in mapping the historical street grid and ED layer for 1940 to create a 1930 block map. The procedure involves several steps and additional manual editing. It was facilitated by another datum from the 1940 block definitions: next to every 1940 ED number was a list of 1930 EDs that were wholly or partly within it (this crosswalk was collated by SteveMorse.org and made available for our use). This provided a means of locating the approximate location of 1930 EDs.

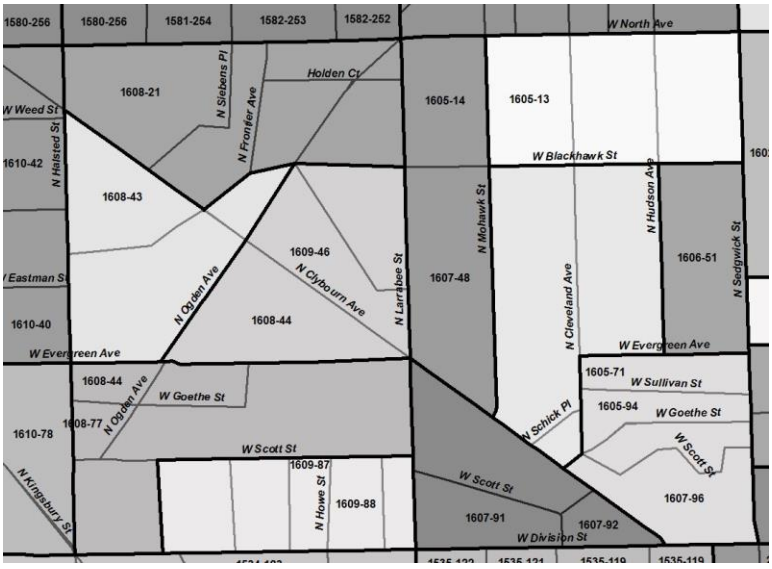
1. 1930 block map: first draft and editing process

In 1930 Chicago contained more than 15,000 populated blocks. However if we could locate a single address on a block in 1930 (if we knew its location and which side of the street it was on) we could assign a block number to that location. The 1930 microdata file includes at least one address in 12,000 blocks, so most blocks in Chicago could be labeled this way. But how could we place these blocks on the map?

For the purpose of making a first draft of the 1930 block map we relied on contemporary 2012 address geocoding in the following way. We treated every street segment in 1930 and 2012 as two cases, an odd numbered and even numbered segment. We also knew in 2012 which side of the street was odd or even. If there were a street segment in 2012 whose address range coincided with the address range on that same street and on a single block in 1930, there was a good chance that these were actually the same block.

A question is how much the 1930 and 2012 address ranges should overlap in order to consider them the same. After some experimentation we decided that if the lowest house number and the highest house number on the street segment in 1930 were within 30 of the lowest and highest numbers in 2012, or if the range of addresses in either year could fit within the range in the other year, it would be a likely match. Of the 15,522 census blocks in Chicago in 1930, more than 13,000 blocks included at least one "matching" street segment by this criterion. If there were a match, then we knew the coordinates of the 1930 street segment. We also knew whether the addresses were on the odd or even side of the street, and based on that we could assign them to a specific 1930 block. That 1930 block number could then be added to the corresponding 1940 census block polygon. A majority of blocks were given a tentative 1930 block number in this way.

The result of this procedure for a portion of Chicago is illustrated in Figure 6 (upper panel). The figure shows several blocks with no label. In some cases two or more blocks are assigned the same block number. And in some cases more than one number has been assigned to a block. Clearly this map needs further attention. However the map also displays a pattern that suggests that many blocks are correctly labeled: there is only one block with a given block number, and there is an apparent logical pattern of block numbering. Upon further inspection we noticed that every Chicago ward had its own series of block numbers (from 1 to as high as 700+). Successive block numbers were usually adjacent to one another. ED numbers showed much less pattern in numbering, but typically each ED contained a set of consecutive block numbers.



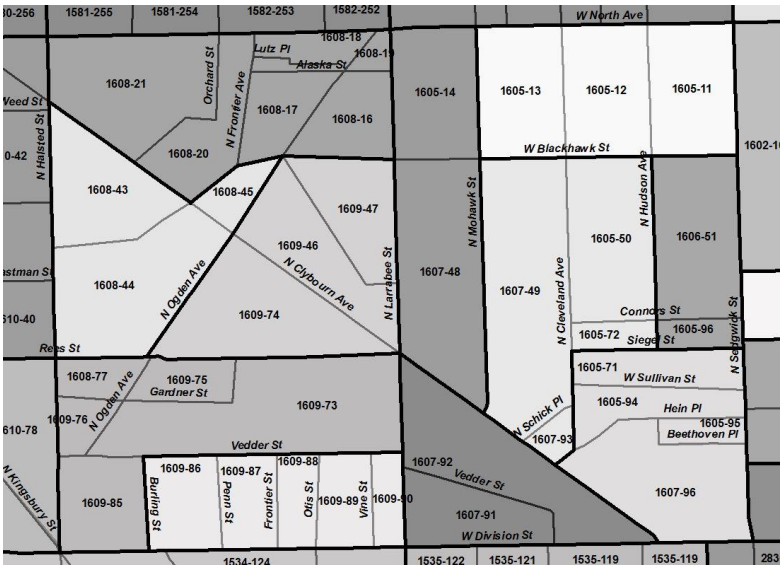


Figure 6. The physical blocks identified as polygons with automated census block labeling (upper panel) and the final corrected map for 1930

The editing process in reflected in the lower panel of Figure 6. This illustration merits close examination. Note first that different shades (and thick boundary lines) have been drawn on each panel to identify the boundaries of 1940 EDs. The 1940 block descriptions list which 1930 EDs (or parts of EDs) are within each 1940 ED. Therefore, for example, we know from the start that all the blocks in 1930 ED 1605 would be found in one of three 1940 EDs in the eastern section of this neighborhood and above N. Clybourne Avenue. One of these blocks (1605-14) was tentatively located north of W. Blackhawk Street and west of N. Mohawk, and we confirmed this location by discovering in the 1930 microdata that people on this block were listed as living on Blackhawk, Mohawk, North, and Larrabee – evidently the boundary streets of this block. In the course of checking block by block, we also found errors on the map. For example note that W. Scott Street in the southwest corner turned out to be Vedder Street. We deduced this because block 1609-73 had no residents on Scott but many on Vedder (and Vedder had to be the southern boundary street for the block be7cause other boundary streets were properly named. Finally, we note missing street names in the initial map in the area below Scott/Vedder. Several north-south one-block street segments were on the Census Bureau's 1940 street map but without names. We followed an interactive process linking 1930 ED numbers and possible boundary streets for those EDs based on the 1930 microdata, correcting one name, adding others, and correcting the two tentative block labels on the initial map.

2. Geocoding the 1930 and 1940 addresses. Based on a nearly complete 1930 ED-block map and knowing from the microdata which addresses were found in which block, it is straightforward to place 1930 addresses in the proper order along a street segment on a single block. In cases where a single block number is unclear the geocoding can often be done by elimination – if there are four blocks along a street in the 1930 ED and three of them have identified block numbers, then any residual addresses are logically on the fourth block. If there is greater uncertainty, the division of residual addresses into blocks in that ED will have to be arbitrary.

Figure 7 illustrates the difference between the original geocoding of 1940 addresses (numbered points with markers on the streets) and the geocoding of 1930 addresses that takes advantage of 1930 block information (numbered points with markers offset from the streets). Note that the original address range for North Larrabee Street between Blackhawk and Holden was 1445-1481. In the revision, the range is 1500-1538. All of the 1400s have been moved to the block south of Blackhawk (which is in the same 1940 ED but a different 1930 ED).

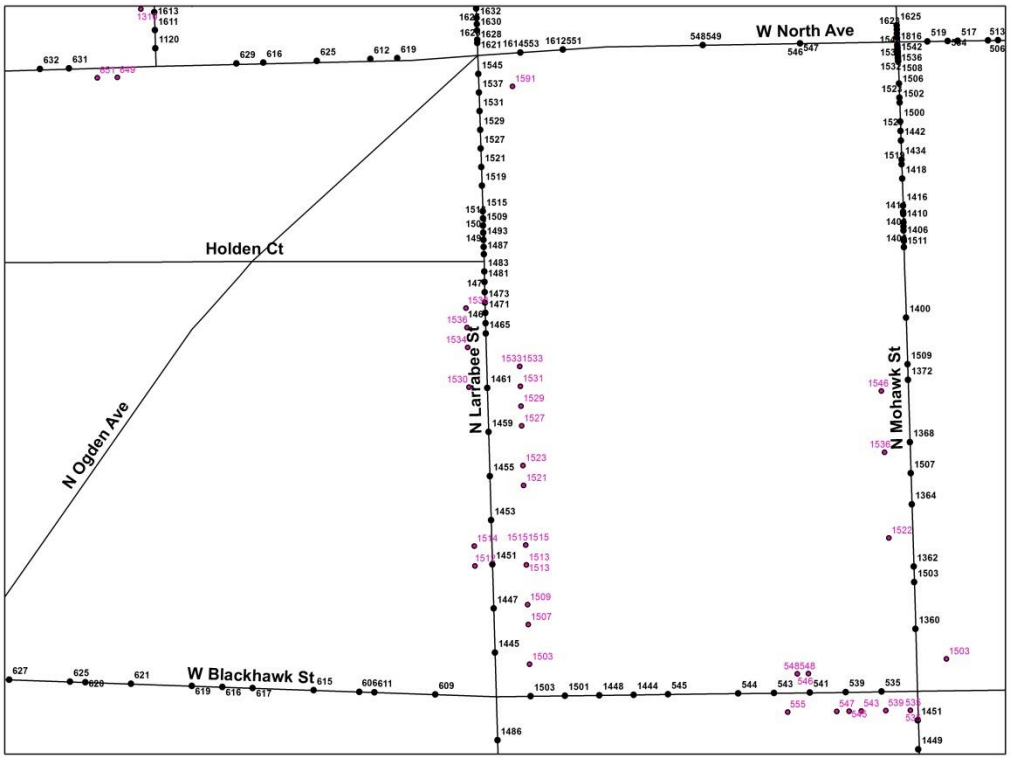


Figure 7. Area of Chicago showing original 1940 geocoded addresses and revised based on 1930 address ranges

We then use the 1930 address range to inform geocoding in 1940. As noted above, we assume that a given address range (again, dealing separately with odd and even numbers) on a given street will lie along the same physical block in 1940 as it did in 1930. The main ambiguity here comes about when there are 1940 addresses on that street that fall outside those address ranges. For example, suppose numbers 320-350 East Fifth Street are on one block in 1930, and 402-486 East Fifth Street are on another. Where would we place 380 East Fifth Street in 1940? We do not know for sure and any decision may introduce error. We hesitate to rely on placement in 2012, especially because in some cases the same street is not found in 2012 but also because we are uncertain whether there has been a change in the numbering scheme. Our approach is first to place 380 on the same block as the closest geocoded address, in this example on the 402-486 block. But since the skip between blocks in this case includes a number evenly divisible by 100, we assume that the actual theoretical range of the 320-350 block is 300-398, and we place 380 on that block.

Conclusion

Creating a historical GIS infrastructure of U.S. cities will generate many new opportunities for historical analysis. The initial shape files with geocoded census data offer an extremely flexible basis for spatial analysis. It is vastly different from the data on city wards that was for so long the principal source for cross-city and over time research. This also opens up new questions, especially what is the spatial scale at which analyses should be conducted? Our assumption has been that neighborhoods were an essential building block of social life in the period of urban transition that we are studying. But what is a neighborhood, and how do we place boundaries on it? Freed from what many researchers have described as the forced choice of treating census tracts as neighborhoods, what is the alternative? That is a problem that we have begun to focus on [20, 21], a task that relies on the sort of geocoded 100% data that we are developing in this project.

Another opportunity is to add information from other sources to these base maps. For this purpose the accurate historical street grid and address ranges are crucial, because any event or institution or photograph with a known address (or at least an approximate location) can easily be added to the GIS. While many questions can be directly answered with population data, for many other questions the population distribution is only a backdrop. A strength of GIS is its expandability.

This chapter has provided much more detail about how to construct a GIS than on how it can be used. Interested readers can consult the studies referenced here and other studies for that purpose. Our primary goal here is to lay out the methodology of a specific HGIS project, partly to document it but equally to reveal the complexity of the mapping process. Contemporary GIS research counts on shape files of all kinds that are often pre-prepared and validated. Historical studies regularly need to create the spatial data. In this case the innovation lies in how disparate sorts of information could be pieced together. This project nevertheless has much in common with others: the need for an accurately projected base map, the importance of consistent place names and ways to estimate their locations, a tolerance for simplifying assumptions combined with a constant concern for accuracy and replicability. In these respects every HGIS project builds on the experience of previous ones and helps pave the way for the next.

Notes

[1] W. E. B. Du Bois and I. Eaton, *The Philadelphia Negro: A Social Study*: Published for the University, 1899.

[2] J. P. Radford, 'Race, residence and ideology: Charleston, South Carolina in the mid-nineteenth century,' *Journal of Historical Geography 2,* 1976, vol. 4, 329-46.

[3] H. N. Rabinowitz, *Race Relations in the Urban South, 1865-1890.* Athens: University of Georgia Press., 1978.

[4] P. A. Groves and E. K. Muller, 'The evolution of black residential areas in late nineteenth century Cities," *Journal of Historical Geography 1,* 1975, 169-191.

[5] T. Kessner, *The Golden Door: Italian and Jewish Immigrant Mobility in New York City*, New York: Oxford UP, 1977.

[6] O. Zunz, *The Changing Face of Inequality: Urbanization, Industrial Development, and Immigrants in Detroit, 1880-1920.* Chicago, IL: University of Chicago, 1982.

[7] I. N. Gregory and R. G. Healey, 'Historical GIS: structuring, mapping and analysing geographies of the past,' *Progress in Human Geography,* 2007, vol. 31, 638-653.

[8] A. K. Knowles and A. Hillier, *Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship*: ESRI, Inc., 2008.

[9] J. R. Logan and W. Zhang, 'White ethnic residential segregation in historical perspective: US cities in 1880,' *Social Science Research,* 2012, vol. 41, 1292-1306.

[10] O. Zeller, *Historical GIS: The Spatial Turn in Social Science History.—Social Science History, Duke University Press, volume 24, n° 3, automne 2000*, Cahiers d'histoire, 2000.

[11] C. Gaffield, 'Conceptualizing and constructing the Canadian century research infrastructure,' *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 2007, vol. 40, 54-64.

[12] I. N. Gregory, C. Bennett, V. L. Gilham, and H. R. Southall, 'The Great Britain historical GIS project: from maps to changing human geography,' *The Cartographic Journal,* 2002, vol. 39, 37-49.

[13] M. De Moor and T. Wiedemann, 'Reconstructing territorial units and hierarchies: A Belgian example,' *History and Computing, 2001,* vol. 13, 71-97.

[14] P. K. Bol, 'The China historical geographic information system (CHGIS): Choices faced, lessons learned,' in *Conference on Historical Maps and GIS, Nagoya University*, 2007.

[15] J. R. Logan, J. Jindrich, H. Shin, and W. Zhang, 'Mapping America in 1880: The urban transition historical GIS project,' *Historical Methods,* 2011, vol. 44, 49-60.

[16] C. Villarreal, B. Bettenhausen, E. Hanss, and J. Hersh, 'Historical health conditions in major US cities: The HUE data set,' *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 2014, vol. 47, 67-80. See also D. Lafreniere and J. Gilliland, 'All the World's a Stage: A GIS Framework for Recreating Personal Time-Space from Qualitative and Quantitative Sources, '*Transactions in GIS,* 2015, vol. 19, 225-246. J. Gilliland and S. Olson, 'Residential Segregation in the Industrializing City: A Closer Look,' *Urban Geography,* 2010, vol. 31, 29-58.

[17] J. R. Logan, W. Zhang, and M. Chunyu, 'Emergent ghettos: Black neighborhoods in New York and Chicago, 1880–1940,' *American Journal of Sociology,* 2015, vol. 120, 1055-1094.

[18] J. R. Logan, W. Zhang, R. Turner, and A. Shertzer, 'Creating the black ghetto: Black residential patterns before and during the Great Migration," *The ANNALS of the American Academy of Political and Social Science,* 2015, vol. 660, 18-35.

[19] U.S. Bureau of the Census, *Sixteenth Census of the United States: 1940. Housing. Supplement to the First Series [Data for Small Areas] Block Statistics for Cities.* Washington, D.C.: Government Printing Office, 1941-1942.

[20] J. R. Logan, S. Spielman, H. Xu, and P. N. Klein, 'Identifying and bounding ethnic neighborhoods,' *Urban Geography,* 2011, vol. 32, 334-359.
[21] S. E. Spielman and J. R. Logan, 'Using high-resolution population data to identify neighborhoods and establish their boundaries,' *Annals of the Association of American Geographers,* 2013, vol. 103, 67-84.