Secondary data: Engaging numbers critically

Kevin St. Martin, Rutgers University

Marianna Pavlovskaya, Hunter College

Chapter 11 in

Research Methods in Geography: A First Course

Edited by

John Paul Jones III, University of Arizona

&

Basil Gomez, Indiana State University

C:\articles\Res methods\Ch11 SecData Merge 3mp.doc

Last saved 6/25/2007 12:08 PM

Introduction

Secondary data is the data that researchers do not create themselves but use in their research. Compared to primary data that is generated over the course of fieldwork (e.g. measuring water quality or interviewing respondents), “secondary” data is already created by someone else. Secondary data providers include government agencies and private companies or such sources as published scientific studies, archives, or collections. Most commonly, the term “secondary data” refers to relatively large databases that individual researchers would not be able to gather themselves (e.g. census data, newspaper archives, inventories of resources, or satellite imagery). Although called “secondary,” this data informs a great deal of academic work and is central to entire subdisciplines in the social and environmental sciences. Moreover, the importance of secondary data in research and policy development is likely to increase with time. This is because information technologies have facilitated an explosion of a wide range of both environmental and socio-economic digital information as well as methods for its analysis. Widely available and accepted as legitimate, secondary data has come to influence in important ways what kind of knowledge we produce and how. The ubiquity of secondary data, especially within the global north, demands that we carefully evaluate its potentials and limitations before integrating it into any research project or using it to answer specific research questions.

This chapter addresses some of the issues related to the use of secondary data by geographers. We point to the wide variety of secondary data and its many sources and we discuss the important advantages and limitations of secondary data. We then address issues of particular importance to geographers: ecological fallacy and the Modifiable Areal Unit Problem (MAUP) as they relate to secondary data. Finally, we illustrate the need to engage creatively and critically with secondary data by focusing on non-standard approaches to analysis that use “mixed” research methods. Throughout the chapter, we will use examples from our and our students’ work in urban geography (cases from Moscow and New York) as well as resource management (the case of fisheries in the Northeast US).

Many kinds and sources of secondary data

Secondary data includes many different kinds of information about natural and human processes that is collected by various government agencies, non-government organizations, or corporations. Examples of such data include population census data, health statistics, school attainment scores, weather monitoring data, remotely sensed images, ocean surface temperature measurements, fish stock abundance calculations, quantities of hazardous materials released into the environment, results from public opinion polls and other population or business surveys, as well as data often presented in map form such as voting patterns, landuse, or elevation.

In the US, much secondary data is collected and distributed by government organizations such as the Census Bureau, Environmental Protection Agency (EPA), National Institute of Health (NIH), National Oceanic and Atmospheric Agency (NOAA), and Geological Survey (USGS) to name a few. In addition, numerous private agencies collect and sell large amounts of data. They include real estate and environmental consulting firms, insurance and financial companies, marketing companies, and so on. Finally, a number of private agencies re-process government collected data, often doing much of the work that is required before such data can be effectively analyzed, or they operate as distributors for data products the government may not be interested to produce in great quantities.

"Secondary data" clearly encompasses disparate information that originates in a wide variety of sites. As such, one must assume that such data will vary greatly in terms of its form and type, its spatial or temporal coverage, and the categories or classifications through which it is organized. In many cases, these qualifications will determine the utility of a given dataset for a particular research project. In addition, each collection method, technique of recording and aggregating, and resultant dataset is embedded within the historical and social context of the agency or corporation that developed it.

For example, National Marine Fisheries Service (NMFS) data is collected and recorded as a means to quantitatively assess fish stock abundance. This focus clearly emerges from the service's historic mandate to manage fisheries resources such that maximum yield can be obtained rather than, for example, maintaining fishing communities. The service's core datasets, then, concern quantities of fish in the sea. Sea sampling of fish populations is done using a spatial grid with a resolution appropriate for such statistical sampling but too coarse for community level studies. The information gathered is both quantitative and qualitative – quantitative information usually consists of numerical measurements while qualitative information reflects differences in kind – but because the information is stored in a database, the latter information is limited to short string descriptors rather than, for example, the detailed text one might generate from fieldwork (e.g. field notes). Also, while the temporal coverage of the core NMFS data sets is impressive (several decades), much of the data that would be useful to social scientists (e.g. crew size on fishing vessels) has only been collected since 1994. Finally, any NMFS data that might aid socio-economic analysis is organized by the category of fishing vessel rather than by individual industry participants. This makes socio-economic analysis at the level of the fisherman (e.g. issues related to employment, job description, wages and benefits, work tenure) virtually impossible. While NMFS is tasked with collecting data relevant to fisheries in the US, it is clear that the data collected is of limited use to social scientists interested in the scale of community, questions of employment, or socio-economic change over time.

While secondary data varies greatly, is produced by a wide range of organizations, and reflects the idiosyncratic history of those organizations, there are many issues that are common across datasets. This is especially true insofar as information is increasingly stored within digital databases that share principles of organization, methods of query, and forms of reporting.

From paper to digital databases

Just a few decades ago secondary data existed only on paper; all transformations and calculations were made by hand or using a calculator. Paper was the medium to store the data and the results of any query, analytical operation, or interpretation. Today much secondary data, especially in post-industrial societies, is created, stored, analyzed, and distributed digitally. Digital spreadsheets and relational databases have come to replace printed tables. The implications are profound. For example, the volumes of data that are created and stored have increased dramatically, datasets can be accessed much faster, datasets residing in various locations can be remotely linked to act as a single database via the internet, and very large databases can be easily imported, visualized, and analyzed with various software packages that include statistical analytical programs (e.g. SPSS) and geographic information systems (e.g. ArcGIS).

Digital secondary data is most often structured in databases organized as one or multiple tables which can be logically related according to shared attributes (i.e. a relational database). In these tables, rows represent individual cases (e.g., weather stations, land parcels, or census tracts) and columns (i.e. fields) represent their quantitative or qualitative characteristics or variables. While there is much secondary data that is not organized in relational databases, there is clearly a movement in that direction even for those data not normally associated with a tabular form or even digital storage. For example, newspaper articles are now mostly organized in digital form and indexed as cases within a database. The same sort of search and query operations that could return, for example, all sea sample sites where a particular number of juvenile cod were observed (or not) by a NMFS scientist could, given a very different database, return all newspaper articles published in the last five years that mention the crisis in cod fisheries and the loss of local livelihoods. Even archives of visual information such as photographs and maps are being organized via relational tables and ruled by the same principles and logics.

Overall, the amount of digital information has grown dramatically in the last two decades and will do so even more in the near future. Despite being called "secondary" this type of data is becoming the "primary" source for many research projects. As such, it is important to understand its advantages and fundamental limitations as well as the politics surrounding secondary data production, distribution, and use.

Advantages

Among the obvious advantages of secondary data, we will briefly consider its scale and size, professional quality and accessibility, and its association with spatial referencing.

These attributes of secondary data provide opportunities for particular forms of analysis that simply would not otherwise exist. Yet, as many historians of science have made clear, the type of data collected, its scale and form, its categories and classification schemes will advance the interests of some but not all. For example, we may point to how NMFS datasets are aligned with the interests of a corporate and large scale fishing industry; indeed, the close relationship between economic power and state sponsored data collection is not uncommon (or undocumented). Yet, a close examination of any dataset can reveal its potential to do unintended or unimagined work. Some variables within those same NMFS datasets can be re-interpreted by social scientists in new ways. For example, the number of crewmembers on a vessel is collected as an indicator of fishing pressure, an important variable in biological assessments of fish stock, but that same variable could be re-interpreted and used as an indicator of employment and its change over time despite its being buried within a table concerned with fish stock rather than socio-economic analysis. To whom secondary data can be an advantage (e.g. corporate vessel owners or crew member/labor organizations) is never fixed; its advantages are open to those willing to spend the time to "get to know" the data and who can then take advantage of its scale, legitimacy, and accessibility.

Scale

Most secondary data, because of its extensive spatial coverage and masses of information collected, simply has no substitute. Individual researchers or even research teams could not possibly produce datasets of comparable size or scale. Government population censuses, for example, cover national territories and entire populations. They generate hundreds of variables for detailed spatial units and do so as often as every ten years. Real estate databases, too, describe housing stock in great detail and, in some countries, local real estate databases are integrated through the internet such that hundreds of thousands of properties can be queried. Inventories of resources, such as fisheries, are not only nationally collected but integrated into international systems of data collection and reporting (e.g. FAO fisheries databases) that makes global environmental analyses possible. In addition, some secondary datasets contain data that is nearly technically impossible to achieve without considerable government investment, for example, satellite imagery, radar, or LIDAR datasets.

Data collection often begins with individual cases and small areas; data is then aggregated to include multiple cases and larger spatial units. In the past, when data was stored on paper, aggregations were fixed. For example, mapped census data would have been aggregated to one type of spatial unit (e.g. census tracts) such that it would literally only exist at that scale (in addition to the raw data). With digital databases, aggregation levels are no longer fixed and, in most cases, the data provider or researcher must, themselves, specify the appropriate and desired level of aggregation given the project at hand (e.g. census blocks and block groups in addition to census tracts). Furthermore, if the spatial scale of a secondary dataset is received at a fixed level or in static map form (e.g. census data at the state level when one’s project focuses on local communities), it is worth inquiring with the agency that created the data as other choices might be available (e.g. census data at a finer spatial resolution such as county, zip code, or census tract). Indeed, it is likely that data exists such that it can be output at a variety of scales that differ from the scale of standard data products.

Legitimacy

Information in secondary datasets is usually organized consistently making the latter well suited for many types of quantitative or statistical analysis, often the very reason such data is being collected. In addition, secondary data is created by specially trained professionals who pre-test questions and verify categories in order to produce standard and comparable information, both across time and space (e.g. for examining trends or comparing information across similar areal units such as counties or provinces). The standard form of secondary data also allows researchers to design data collection projects that add to or can be compared with existing secondary datasets.

Importantly, the professional systems of collection, assembly, storage, and retrieval that constitute secondary data confer a legitimacy that is widely recognized and works to empower secondary data, make it rhetorically convincing, and allow it to convince in ways other datasets cannot. For example, many datasets are derived from dubious information that is self-reported by businesses, individuals, or resource users. Yet, once aggregated in a consistent and organized form, such information, despite its origins, becomes the basis for formal scientific analyses. In fisheries, “log book” data from fishing vessels is a form of self-reporting where vessel captains report fish catch, discards, trip location, and other variables to NMFS. While fishermen’s individual stories are often derided as anecdotal or exaggeration, their “log book” entries are made believable via the technical systems within which they are embedded.

Similarly, we observe that a great deal of the digital spatial data (map layers) currently available in secondary databases were digitized from paper maps that might be decades old, interpolated from sparse control points, or simply geocoded incorrectly (e.g. the location of the Chinese embassy in Belgrade). Yet, such layers, once in digital form, appear to exude accuracy and instill confidence in the analyses being performed.

Accessibility

Importantly, the largest and most comprehensive datasets (e.g. census data) are often produced by public agencies and are publicly available (e.g. at a low or no cost). This makes them accessible to academics but also analysts working for NGO's and grass-root organizations who can analyze these data with respect to their needs or political causes.[1] Overall, such democratization of digital technologies and information serves to empower a variety of social actors beyond the state and corporations.

The increasing accessibility of secondary data also facilitates its use as an exploratory first step in research projects that then focus on primary data collection. Widely available, affordable, and easy to use, secondary data can be used to more efficiently target costly and time-consuming primary data collection. Among other things, it is often used to identify places and/or populations for more in-depth qualitative or quantitative study. In one of our projects, for example, we used census data to identify neighborhoods within New York City that contain large numbers of Spanish and Russian speakers. In addition, municipal level information (available from the New York Department of Education website) was examined to estimate the number of immigrant students attending the public schools within those same areas. Taken together, we could identify neighborhoods where recent immigrants with young children reside. These populations were then the target of a major interview-based research project that focused on the multiple economic practices of immigrant households.

While seemingly ubiquitous from the perspective of the global north, there are limits to the accessibility of secondary data. In particular, a large gap exists in the relative abilities of rich and poor countries to access, produce, utilize, and control digital information. As this gap reflects differences in economic and political power, the advanced post-industrial societies have obvious advantages. Countries of the global south, however, are increasingly conscious about the need to narrow the digital divide and, as digital technologies become more affordable and easy to use, their governments are launching their own data collection projects. International corporations, too, fill their digital data banks with information about new resource, labor, and consumption markets in the global south. It would seem that, for better or worse, the digital coverage of the world is rapidly expanding and providing every more sources of information for research.

Finally, as geographers, we note that the growing accessibility of digital secondary data is closely linked to the growth of geomatic technologies such that access to secondary data increasingly implies access to georeferenced data. Much of the data in secondary datasets is either collected by spatial units (e.g. census tracts or electoral districts) or includes other locational information (street address or geographic coordinates). This data, therefore, can be visualized, explored, and analyzed using Geographic Information Systems.

Working with secondary data has many advantages; its scale(s) and magnitude, its widespread legitimacy, and its ever growing accessibility make it an incomparable source of both social and environmental information to the geographer. And yet we cannot uncritically rely upon secondary data. It is important to remember that the advantages of secondary data should be evaluated relative to its limitations which can be, at times, severe.

Limitations

Despite the advantages of secondary data, its use may, ironically, narrow research opportunities and decrease the quality of findings. In this section, we will discuss the limitations of secondary data that can, without critical interrogation, hamper one's project. We will discuss how secondary data simply is not explicitly created for your particular project, how datasets may become internally inconsistent over time or across space, how what appears as full coverage may be based on sampling, how such data may not represent the population that you think it does, and how its precision must be balanced with issues of privacy, errors, and locational inaccuracy.

Created for which purposes?

Using secondary data means that we use the data created by someone else and for their own purposes to answer our specific research questions. Even large multiuse datasets are structured according to some original purpose (e.g. census data is for voting or taxation purposes and NMFS data is for biological assessment of fish stock). Embedded in the data, the initial design influences and limits our research questions, methods, and findings. For example, it would be very difficult to study some aspect of global climate change that has not been already incorporated into pre-existing global datasets. The latter compile many but certainly not all variables of interest to researchers of global climate change. Similarly, a social scientist researching poverty must rely on a particular definition of income that has been built into particular census categories. That is, a census will typically report a household's official monetary income but is unlikely to include other types of economic activity (e.g. informal and/or unpaid production of goods and services) that may be important for coping with poverty. Domestic work, informal work for cash, in-home childcare, and exchanges between households and within a community are as important for social reproduction as formal wages yet they are absent from census data.

As in the case of social scientists' use of NMFS databases, categories designed for one purpose may be creatively re-interpreted for another. This reinterpretation is, however, limited by the history and context of the agency or organization which is then reflected in the databases they create. Clearly, secondary datasets are initiated and maintained for particular purposes and, therefore, may only be useable if researchers can creatively reinterpret existing data or, as in all too many cases, modify their original research questions to fit the data. The use of secondary data has the capacity to limit analytical possibilities such that original research questions may, in the end, remain unanswered.

Data collection practices change

Large-scale data collection practices do not stay constant and researchers who use secondary data have no control over these changes. Even in such a uniform and consistent data set as the US census, analytical categories (variables) or the boundaries of spatial units (e.g. census tracts) may change from one decade to another. In addition, new variables are often added and existing spatial units (dis)aggregated. Consequently, making longitudinal (i.e. over time) comparisons becomes difficult (see also MAUP below) and sometimes simply impossible. For example, after the Soviet Union collapsed, the new administration of Moscow reduced the number of major districts from 33 in 1992 to only 9 in 1993. This was done in order to radically modify and break away from Soviet era power structures. As a result, it was no longer possible to directly compare the socio-economic situation in Moscow before and after the transition to capitalism because all socio-economic data was tied to one of two incomparable spatial systems. Unfortunately, very important and interesting research opportunities were closed by the change in data collection and organization.[2]

Full coverage or interpolated sample?

Despite their size and scale, few databases fully cover the populations they claim to represent. Most often, variables are estimated from selected samples and, therefore, may be subject to sampling errors and biases. As a census, the Bureau of Census provides full coverage every 10 years. Only 1 in 6 households, however, fill out "the long form" that solicits some of the most important socio-economic data. Similarly, the so-called micro-data from the census (Public Use Microdata Sample or PUMS) provides detailed information about housing units and people in them (as opposed to geographic areas such as census blocks, block groups, or tracts). While it enables the tabulation of information in the ways that the regular census dataset does not, the findings are valid only if the sample is adequate (i.e. it includes a statistically valid number of cases).

In the case of fisheries, NMFS sea sampling of fish stock is organized at a spatial resolution that is appropriate for a regional inventory and regional-wide regulatory mechanisms (e.g. limits on catch). The resolution at which NMFS samples, however, makes the assessment of local fish stocks and habitats difficult at best. As a result, small-scale fishermen whose fishing practices are local find it difficult to relate to NMFS pronouncements of stock health/demise and find the ensuing region-wide regulations out of step with their local experiences or needs.

Recognizing "silences" in the secondary data

Secondary datasets are fundamentally partial representations. They only contain information about selected phenomena or their aspects and, therefore, always omit information about other phenomena or their aspects. The result is the effective silencing and disempowerment of processes, people, or places that are not represented. For example, only certain types of "formal" phenomena are described by socio-economic data that is regularly collected by state agencies. Such phenomena are, however, only "the tip of the iceberg" and other "informal" economic or social practices go undocumented and remain unseen within state sanctioned datasets.[3] Similarly, while environmental processes and change over time are clearly affected by human activities, only some of these are accounted for in formal databases related to environmental management thus hampering our estimation of the "drivers" of environmental change. For example, while commercial fishing activities are carefully monitored relative to fisheries management, recreational and subsistence fishing are not even though they are thought to have considerable impact on particular fisheries resources).

Formal activities are accounted for and measured and, therefore, can be made part of a secondary database. Formal employment, formal health care services, formal childcare, formal consumption, formal resource harvesting and the like are clearly important to document; yet they cannot capture the totality of human experience. No consistent information exists concerning the informal household, networks, subsistence use of resources, or community economies. What or who is not included, (i.e. the "silences") in secondary datasets will clearly effect and limit our ability to construct explanations using secondary data.

Furthermore, the datasets that do target a particular formal phenomenon may not represent it completely. The US Census bureau, for instance, conducts an economic census of US businesses classified into industries (e.g. real estate or professional services) and aggregated into geographic units (e.g. states, counties, and even zip codes). This vast dataset allows one to research regional and local economies but it incorporates only those establishments with hired workers (e.g. paid employees) and, therefore, excludes many small businesses (e.g. all self-employed workers and many family businesses). Consequently, this dataset provides only limited insight into local economies and services.

Understanding what variables actually measure

While quantitative data, especially if distributed by government and professional agencies, seems objective, precise, and unambiguous, it is important to understand just how the variables are constructed and what exactly they may or may not measure. The most telling example is the concept of “race” as used in the US census. Prior to the 2000 census people could only choose one racial category; this prevented them from identifying with more than one race. The resultant statistics on race concealed the racial diversity of many individuals and oversimplified the racial composition of the US population. In addition, the limited number of racial categories used by the census had a disciplining power insofar as they forced people to identify themselves and others in those terms – a powerful process that for centuries has worked to construct and maintain class and other hierarchies based on particular racial categories.

Analyzing health of individuals against other socio-economic or public health variables may also create problems. The state of health is often self-reported and, therefore, is a highly subjective measure that depends on how respondents understand the meaning of being in good or poor health. And yet, it is used in conjunction with other variables that are less subjective because they are not “self-reported.”

Categories

The ambiguity of data categories (in addition to the ambiguity of variables) is another important consideration. Our research with one of our graduate students on the diverse economies of Arab American communities in the Northeast US exemplifies this issue. To identify these communities she used census data, one of the only sources for such information. Yet, the census does not define Arab Americans directly via a single variable. Racial categories subsume Middle Easterners as “white.” “Arab” populations can, however, be discerned in other ways using census data. For example, “Arabs” could be defined as those who speak the Arabic language, or those who come from a predominantly Arab country (national origin), or those who declare “Arab” as their ancestry. These definitions offer three overlapping but incongruent ways to count Arab Americans. Relatively recent immigrants are more likely to speak Arabic, national origin includes non-Arabic groups, and ancestry is an ambiguous category in itself. Using these definitions, our student produced three maps that show three different although overlapping distributions of the Arab American population. Even such comprehensive datasets as the US Census sometimes offer only partial representations.

Privacy

The contradiction between the need for detailed data and the need to protect the privacy of individuals sometimes demands that researchers make important decisions about how their research may or may not proceed. For example, certain datasets contain sensitive information collected at the level of individuals or households. On the one hand, such detail might be essential to analysis. On the other hand, its utilization in research might actually disclose an individual’s private information. In order to avoid such a violation of privacy, the data are typically aggregated to relatively large spatial units which necessarily leads to information and accuracy loss. This is true for much health-related data as well as PUMS micro data from the US population census.

Where high resolution data is available and its utilization is acceptable, it can still be problematic. For example, mapping detailed information on income, education levels, race, crime rates, etc. should be done carefully as it may lead to the stigmatization of particular people and places with implications for their economic and social well-being. In this case, researchers may decide to map their analytical results at scales which are smaller (i.e. larger spatial units) than the scale of the actual analysis.[4]

Errors and accuracy

Finally, all secondary datasets contain errors. Even professionally done surveys, including public opinion polls, may have unknown sampling problems and misrepresent the population in question. The US Census bureau, for example, consistently undercounts millions of mainly illegal immigrants as well as those at addresses not included in the census database. Some of these errors may systematically distort the population they represent thereby contributing to inaccurate policy decisions. For example, census undercounts of immigrant populations who satisfy the demand for cheap labor also politically disempower such working populations. In addition, where undercounting includes families with children, the demand for schools and other services may be underestimated.

There are also random errors that may not distort overall averages but do decrease the quality of the data and the researcher’s ability to work at finer resolutions. In particular, errors at the data-entry stage (e.g. typos in attribute information or mistakes in a spatial layer) are very common. For example, discrepancies in street addresses may result in the elimination of many records (in some cases as much as 40%) that cannot be matched to an address database (this process is called geo-coding). Similarly, the “log book” data from individual fishing trips collected by NMFS are riddled with errors. While many errors are the result of poor data-entry (e.g. NMFS hired companies that use prison labor to enter data from forms where entries were hand-written by fishermen at sea) others derive from fishermen’s deliberate misreporting. In addition, random errors may occur because of technological faults such as instrument calibration problems that reduce the quality of satellite imagery.

Locational errors are especially important in geo-referenced data. They can lead to the wrong conclusions concerning the spatial overlap of phenomena in question. For example, places may be erroneously identified with some negative social phenomenon (see section on privacy) or their exposure to industrial hazards as measured within the Toxics Release Inventory (TRI) database may be underrepresented.[5]

Analytical problems

While the inherent limitations of secondary datasets are cause for concern, so too is the relationship between secondary datasets and a number of analytical issues. In particular, we examine secondary data and its propensity to increase the occurrence of ecological fallacy as well as its relationship to the modifiable areal unit problem (MAUP), two important concerns for geographers.

Ecological fallacy

Ecological fallacy refers to the assumption that all individuals in a group share the average characteristics of that group. In the case of spatial data, we should be careful to not assume that all people residing in a particular geographic area (e.g. a census tract or school district) have properties identical to the average for the area as a whole. The following example from another of our graduate students illustrates this problem. The objective of the student's research was to find out whether differences in the recycling behavior of New Yorkers are determined by differences in their attitudes toward and knowledge about recycling. Individuals from areas with low and high levels of recycling answered questions about their attitudes toward recycling. Their recycling behavior, however, was only assessed using the so-called "diversion rate" (percent diverted from disposal) estimated for each district in the city. Survey respondents were assumed to recycle less or more based upon the average statistic for their district rather than their actual behavior, which is a case of ecological fallacy. Avoiding it involves asking the respondents directly about their recycling behavior.

Secondary databases make the occurrence of ecological fallacy more likely insofar as a wealth of data resembling the data the researcher needs (e.g. recycling behavior) already exists and is readily accessible across multiple spatial units.

Modifiable areal unit problem

Another common analytical problem for geographers is the modifiable areal unit problem (MAUP) which refers to the effect of political boundaries on spatial data and its analysis. In particular, these boundaries are social constructs that may have little to do with the phenomenon under study. In the US, for example, the effect of state and especially county boundaries on the diffusion of diseases, residential segregation, or migration may be very limited and yet data are frequently collected, analyzed, and mapped using such boundaries. In other words, we often identify patterns in data based upon boundaries that are unrelated to the phenomena in question.

Two other aspects of the MAUP are important to consider.[6] First, the boundaries of units for which the data are collected change with time making it difficult or impossible to compare datasets that describe the same territory but in different time periods. The dramatic changes in administrative boundaries in Moscow (see above), which are the basis for organizing socio-economic data, represent an extreme case of MAUP. Second, the scale at which data is presented and analyzed can affect one’s results. For example, analyzing the same census data at the level of census blocks, census block groups, or census tracts (three different scales) may yield different statistics and different spatial patterns (also see Practical Exercise 2). An awareness of the effect of choosing one or another spatial scale is vital. In some cases, choosing a single scale for analysis will precisely address the problem at hand, while in other cases analysis at multiple scales will be necessary to capture those processes that manifest themselves differently at different scales.

In our research on Moscow, the geographic evidence for connections between Soviet-era structures of political and economic control (e.g. economic ministries and Komsomol headquarters) and subsequent capitalist development (e.g. new private banking and financial firms) was only visible at the finest spatial resolution of a single street addresses. Only at this scale, could we see the concentration of new enterprises within the very locations (indeed, offices) of Soviet-era structures of power. At more coarse resolutions this locational coincidence was not visible.

A study of access to open space in New York City conducted with another graduate student illustrates the necessity of a multi-scale approach. Open space ratios that measure access to open space (see the next section for details) and their correlations with socio-economic variables were calculated at three levels: that of community board districts (CBD, the largest units), census tracts, and the neighborhood (measured as open space within walking distance). While a number of socio-economic variables were significantly correlated with open space at the scale of the CBD (e.g. positive with median household income and negative with percent people of color), the same variables could not be used to predict access to open space at finer spatial scales. At those scales, associations were more complex. For example, at the neighborhood level both wealthy and poorer neighborhoods had access to open space but in wealthy neighborhoods open spaces were large (e.g. large urban parks) whereas poorer neighborhoods had access to only very small open spaces.

Working with secondary data

While the advantages and limitations of secondary data must carefully be considered, they clearly contribute to and even expand the scope and power of standard forms of analyses such as querying (i.e. asking questions of the database and retrieving data that answer these questions) or statistical analysis. Such standard forms of database analysis are discussed in detail in a variety of introductory texts and we will not review the here.[7] Rather, we will briefly discuss three strategies for creatively using secondary data. Our goal is to suggest that secondary data can be used in ways that complement creative and critical analyses in geography.

The three examples we provide include transforming and adjusting secondary data to better correspond to one's original research questions, designing new measures and indicators, and using secondary data in a "mixed" method approach that combines quantitative spatial analysis with qualitative interview information. In addition, we will examine the opportunities offered by the emerging fields of data mining and geovisualization.

Redesign the data to suit your research needs

As discussed above, categories and variables embedded within secondary datasets can influence and shape research strategies and findings. To avoid this, we need to critically examine the data and, if necessary, update, revise, and/or combine it with primary data collection. Our research on access to open space in New York City is a good example. For analysis with socio-economic census data, our graduate student obtained from the Department of Parks and Recreation a database indicating the location of open spaces in the city. The categories of open space in this database included “publicly accessible facilities of regional importance” but excluded spaces that predominantly serve single communities such as “playgrounds, basketball and handball courts, and community gardens.” And yet, the latter play a very important role in the daily recreation practices of New Yorkers. Without considering them, the analysis of access to open space would be incomplete.

Updating the database was a time- and effort-intensive but necessary part of the research. The student acquired the additional datasets from several public agencies and NGOs and merged them with the original database. The map in Figure 1 illustrates that there are noticeable differences in calculations of access to open space from the original to the updated database. In many districts the difference exceeds 0.5 acres. This is a considerable discrepancy given city standards for defining severely underserved districts (1.5 acres per 1000 residents or less). Updating the database also proved crucial for obtaining one of the key findings of the study: access to open space varies differently in relation to income and minority status depending upon the size of the open spaces available within walking distance.

Design your own analytical measures

When working with secondary data, creative thinking during the analytical stage helps to limit ourselves to standard statistical measures. Indeed, secondary datasets can be manipulated to produce novel variables and measures that lead to illuminating findings.

The research on access to open space in New York again provides a good example. Here, a new measure was designed to overcome the modifiable areal unit problem that was a result of measuring daily human patterns (e.g. use of the urban parks) based upon arbitrary political spatial entities (i.e. Community Board Districts or CBDs). Traditionally, access is measured as a simple ratio at the level of CBDs and is expressed either as a percent open space or acres of open space per 1000 residents in each unit. CBDs, however, are rather large entities and, for daily recreation, New Yorkers will only use open spaces within walking distance. In addition, this standard measure of access to open space is clearly tied to an arbitrary administrative boundary (the CBD) even though people disregard these boundaries and simply go to the nearest park or playground. We wanted to measure access to open space that accounted for how people use open spaces in their daily lives.

The literature suggested that children will utilize parks within a quarter-mile while adults will walk up to a half-mile. To avoid the effect of the CBD boundaries, the student converted the map of open spaces to a raster format with a cell size of 40 feet. This resolution roughly matches the size of a tax lot and, therefore, can account for even the smallest open spaces (e.g. community gardens). In addition, this cell size also approximates the size of a given residence from which access to open space might be measured. We then calculated for each pixel (our “stand in” for residential buildings) a sum in acres of open space within walking distance (for quarter- and half-mile radii). Instead of a single value per large district, we constructed a surface that reflects much finer variations in access to open space (see Figures 2 and 3). These maps not only show access to open space in new terms using new measures, they also show that there are significant (but ignored) differences in the amount of open space accessible to children and adult New Yorkers.

Mixed methods: Mapping the social "landscape" of fishing communities

In the above, data from a secondary database is transformed and re-worked into a new variable across a new space, the space of access to open space. In the case of fisheries we similarly produced a new measure distributed within a new space. Fisheries science and management repeatedly represents the presence of fishermen and fishing communities on the ocean as aggregate fishing effort expressed in terms of quantities of fish caught. While useful for region-wide estimations of remaining fish stock or future yields, the aggregation of effort to a single variable erases local differences and the dependencies of particular communities upon particular resources.

To express the presence of particular fishing communities and their dependence upon particular fishing grounds, we searched NMFS essentially biological datasets looking for some way to map the social "landscape" of fishing communities. We found in the "log book" data locational information by fishing trip that could be tied to vessel "home port;" this would tell us where particular vessels from particular communities fished. In addition, we found data on the number of crew and trip length; this would give us a measure of labor time. For each trip we multiplied crew on board by trip length to create a new variable, "fishermen days," that could then be linked to particular locations or fishing grounds. Using this new spatial variable we created, for a variety of communities, individual and composite maps (Figure 4) showing the areas upon which they depended. As part of a participatory research project, fishermen from each community were then invited to correct or amend the maps. The maps are already proving valuable as communities lobby for more localized assessments of fish stock and for greater community input into stock management. The “mixed method” approach used in this work (i.e. statistical and GIS analysis of secondary data combined with participatory interviews and workshops) is an emerging and robust way to take advantage of secondary datasets, identify their limitations, and employ alternative methods to both address those limitations and, importantly, distribute the power of secondary data to communities and lay people generally.

New research opportunities in a digital world

Secondary databases themselves, their attributes and the ubiquity, are making possible new forms and styles of analysis. Indeed, they are facilitating forms of knowledge production unique to secondary databases. In particular, new methods that deal with specific properties of large datasets have been employed in a number of fields, including geography and GIS. They first appeared in marketing research that demanded new techniques for the integration and analysis of the growing but disconnected and non-systematized information about consumers and their behavior. Statistically “mining” those databases promised to uncover yet unknown patterns in consumer behavior which could then be leveraged for corporate profit. What is important is that these techniques reverse the traditional approach to research; instead of testing hypotheses, the new data mining algorithms aim to detect patterns that are not yet hypothesized or observed, patterns that uniquely emerge from very large digital databases.

Today, exploration of digital data is a cutting-edge research direction in geography and GIS. In addition to statistical approaches that mine spatial data,[8] new geo-visualization approaches similarly allow for the recognition of patterns in secondary data. Spatial exploratory data analysis involves advanced data displays that combine maps with graphs and tables that help the researcher to visually examine the data and discover new spatial patterns.[9]

Conclusion

The quantity and magnitude of public and commercial digital datasets, and especially those with spatial information, has significantly increased and will continue to do so. Secondary data is now and will remain important to geographic research as a primary source of information to a growing number of data-intensive applications. Using this data clearly gives a researcher important advantages in terms of data coverage, quality, and costs, as well as the opportunity to analyze phenomena that otherwise would be impossible to analyze (e.g. population distribution at a national scale). And yet, the important limitations of secondary data such as the danger of “data-driven” research questions, incomplete representation of phenomena, ambiguity of categories, and issues of privacy should be kept in mind. In addition, geographers should clearly understand the potential ease with which secondary data can lead to ecological fallacy or MAUP.

While the advantages and limitations of secondary data are important considerations for any form of analysis, we are enthusiastic about the possibilities for new and creative analytical techniques that secondary data facilitate. In addition, we should not be confined to the original purpose of any dataset; nor should we shy from manipulating and transforming data to build new variables, measures, or maps; nor should we hesitate to combine secondary data analysis with other methods as in "mixed" methods research. While often associated with standard analytical techniques, secondary databases might usefully be thought of as vast territories to be explored, visualized, and understood using new critical and creative approaches.

Supplemental reading and data websites

Cromley, E. K., and S. L. McLafferty. 2002. *GIS and public health*. New York and London: Guilford press.

Introduction to and advanced treatment of spatial databases, mapping, and spatial analysis with a focus on environmental hazards, infectious and vector-borne diseases, and health services.

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2005. *Geographic information systems and science*. 2nd ed., 315-39. John Wiley and Sons.

A major introductory text to GIS that discusses databases and geovisualization.

Heywood, I., S. Cornelius, and S. Carver. 2006. *An introduction to geographical information systems*. 3nd ed. New York: Prentice Hall (Pearson Education).

A well written, accessible, and comprehensive introduction to GIS, database development, management, and analysis.

Shekhar, S., and S. Chawla. 2003. *Spatial databases: A tour*. Upper Saddle River, NJ: Prentice Hall.

GIS and spatial databases, data models, query languages, storage and indexing, query processing and optimization, spatial networks, spatial data mining.

American Fact Finder http://factfinder.census.gov/home/saff/main.html?_lang=en.

The US Bureau of Census on-line service that provides access to population, housing, economic, and geographic data. Also allows to map data interactively.

Social Explorer http://www.socialexplorer.com/pub/home/home.aspx

Provides easy access to interactive demographic maps of the United States including historical data back to 1940.

CIESIN The Center for International Earth Science Information Network http://www.ciesin.columbia.edu/ within the Earth Institute at Columbia University.

On-line datasets for social, natural, and information sciences. Includes PUMS from US census data.

The Economic Census of the US Census Bureau http://www.census.gov/econ/census02/

Detailed portrait of the US economy once every five years from national to local level. All domestic non-farm non-government business establishments with paid employees.

ICPSR-Census 2000, University of Michigan, Institute for Social Research. http://www.icpsr.umich.edu/CENSUS2000/index.html

Access to census 2000 data files. Explains their content.

MPC Minnesota Population Center, University of Minnesota. www.ipums.umn.edu

Integrated Public Use Microdata Series (IPUMS) census microdata for social and economic research. IPUMS-USA database from 1850 to 2005. UMS-International has census data from around the world.

Toxics Release Inventory (TRI) public database by Environmental Protection Agency http://www.epa.gov/tri/.

Information on toxic chemical releases and other waste management activities.

GeoDa - An Introduction to Spatial Data Analysis, https://www.geoda.uiuc.edu/

Developed by the Department of Geography at the University of Illinois, Urbana-Champaign, the increasingly popular free software GeoDa provides tools for exploratory spatial data analysis.

Practical exercises

Exercise 1. Querying on-line TRI database.

In your internet browser, open the Toxics Release Inventory (TRI) site of the Environmental Protection Agency http://www.epa.gov/tri/. This site provides access to a public database with information on the toxic chemical releases and other waste management activities of certain industries in the US. The website provides tools that let you tabulate (summarize) these data by geographic units (e.g. states), industry, and type of released chemicals. It will also let you compile a report on TRI incidents in particular neighborhoods defined by their zip codes. To find out whether any releases occurred in your neighborhood or any other neighborhood of the US, type in the corresponding zip code. For example, typing in 80524 (Fort Collins, Colorado) reveals that in this area a factory that produces malt beverages released ammonia and polycyclic aromatic compounds into the air in 2004.

The TRI website is a user-friendly interface that allows you to query, in a variety of standard ways, an enormous government database spanning many industries, at a national scale, and over many years. To build multi-attribute queries look to the bottom of the results page where there is a link to the TRI Explorer home page. Using TRI Explorer, you can search over several years (1988-2004), by type of the released chemical, by geographic location, and by industry type. To use these data in more creative and geographic ways (e.g. to explore the correlation between releases and poor or minority areas), you can download the type and location (x and y coordinates) of toxics releases and import them into a GIS.

Exercise 2. Mapping with American Fact Finder.

One of the great features of the American Fact Finder, the interactive database of the US Bureau of Census, is its ability to map census and related data. In this exercise, we will map one census variable (median age) at different spatial resolutions (states and counties) and examine the effect of scale on how the data are visualized and can be interpreted (i.e. the MAUP).

Open the American Fact Finder home page http://factfinder.census.gov/ in your browser and click on the link to the "Decennial Census" located in the left-hand banner under "Data Sets." When the page opens, make sure that the radio button for the "2000 Census Summary File 1 (SF 1)" is checked on. This file contains data that cover the entire US population. Click on "Thematic Maps" in the right hand portion of the screen. You now can specify the geographic scale you wish to use for displaying data. To display the whole United States by state, select "Nation" as the geographic type and "United States" as the geographic area. Click "Next." On the next screen select the theme TM-PO17 Median Age: 2000 (a specific variable from summary file 1). This variable, median age, will split the population in half. In other words, half of the population is younger and the other is older than the median age – the higher the median age, then the older the population of that area. The median age in the US in 2000 was 35.3 years. Click the “Show Result” button to load the map of median age by state. The map legend or key is on the left. It indicates what values are included into each of the five categories shown with different colors. According to these values, the median age varies significantly by state, with 10 years separating the younger populations (27.1 ages in Utah) and older populations (38.9 in West Virginia and 38.7 in Florida). Besides Utah, the states of Texas, California, Idaho, Louisiana, Mississippi, Georgia, and Alaska have a relatively low median age (they are shown by the light yellow color). The “*i*” button activates the query function which you can use to find out the median age value for individual states. Determine which state have the oldest population.

Let us now see whether displaying the same data by county changes the distribution of older and younger populations. In the drop-down box “Display map by” above the map, choose “County” instead of “State” as the spatial unit. When the new map loads, look at the legend and note that the minimum and maximum median age values have changed. At this spatial level, the median age varies from 20 to 58.6 years for individual counties, yielding a gap of almost 40 years instead of 10 years as in the previous map. While both statistics were computed from the same data, the county data retain more variation than does averaging to the state level. Examine the map and determine whether the “younger” states (e.g. Utah, Texas, California, etc.) are uniformly young? Do the “older” states have homogeneously old populations?

What are the possible explanations for median age and its variation within different parts of the country? What erroneous conclusion from these data might you draw that would be an obvious ecological fallacy?

Keywords

Database querying – set of techniques that retrieves the data from databases using structured query language.

Data mining – set of statistical techniques for analysis of large databases that seeks to discover the underlying patterns in data. Includes spatial data mining.

Ecological fallacy – an erroneous assumption that all individuals in a group share the average characteristics of that group.

Geovisualization – computer-based multiple and interactive displays of geo-spatial information.

MAUP (Modifiable areal unit problem) – posits that data analysis is scale-dependent.

Mixed methods – methodologies in social science that integrate quantitative and qualitative research techniques within a single project.

Qualitative data – reflect differences in kind or type of phenomenon

Quantitative data – measure differences in quantity or degree of a phenomenon

Relational database – structured by record (row) and allows for connection with other databases based upon a common field.

Sample – a selected subset that represents the population for statistical purposes.

Spatial exploratory data analysis – the search for new spatial patterns in large databases using a variety of statistical and geovisualization means.

Acknowledgements

We used examples from graduate research of our students Jess Bier, Sara Hodges, and Ben Mancell. Thanks to Kimberly Wolff and Valeria Treves for their research assistance on a number of projects. We thank the anonymous reviewer for the useful comments on the draft of this chapter.

Figure captions

Figure 1. Differences in the amount of open space (acres per 1000 residents) calculated by researchers and the Department of City Planning.

Source: Sara Hodges, 2004. MA thesis “Open space in New York City: A GIS-based analysis of equity of distribution and access” Hunter College, New York. Reprinted with permission.

Figure 2. Children’s access to open space in New York.

Source: Sara Hodges, 2004. MA thesis “Open space in New York City: A GIS-based analysis of equity of distribution and access” Hunter College, New York. Reprinted with permission.

Figure 3. Adult’s access to open space in New York.

Source: Sara Hodges, 2004. MA thesis “Open space in New York City: A GIS-based analysis of equity of distribution and access” Hunter College, New York. Reprinted with permission.

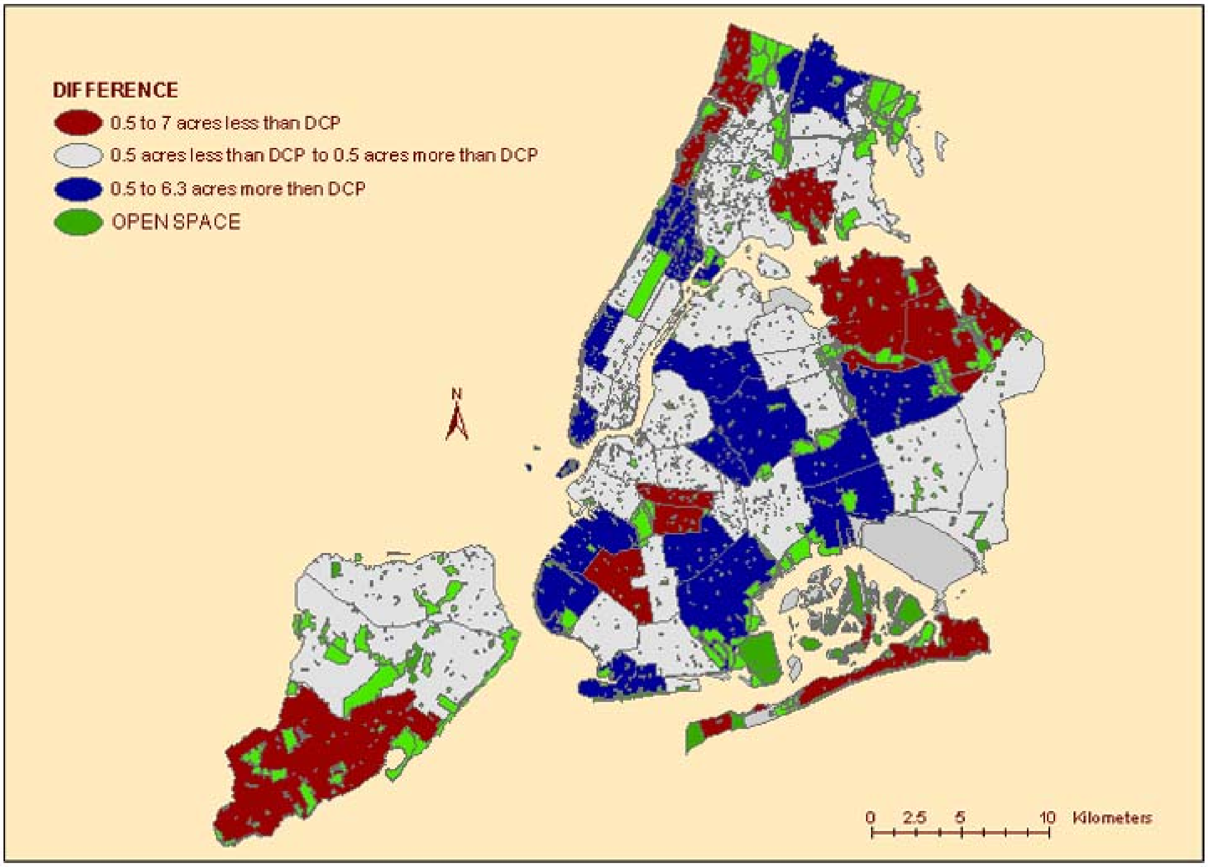
Figure 4. An extract from a map depicting the primary fishing grounds (based on labor time) of small trawl vessels from particular communities/ports in New England (color coded outlines correspond to port markers). The outlines are superimposed upon a NOAA nautical chart. The map also contains a raster density surface (green shading) based on the aggregate of all vessels.

Source: author (St. Martin).

Figures

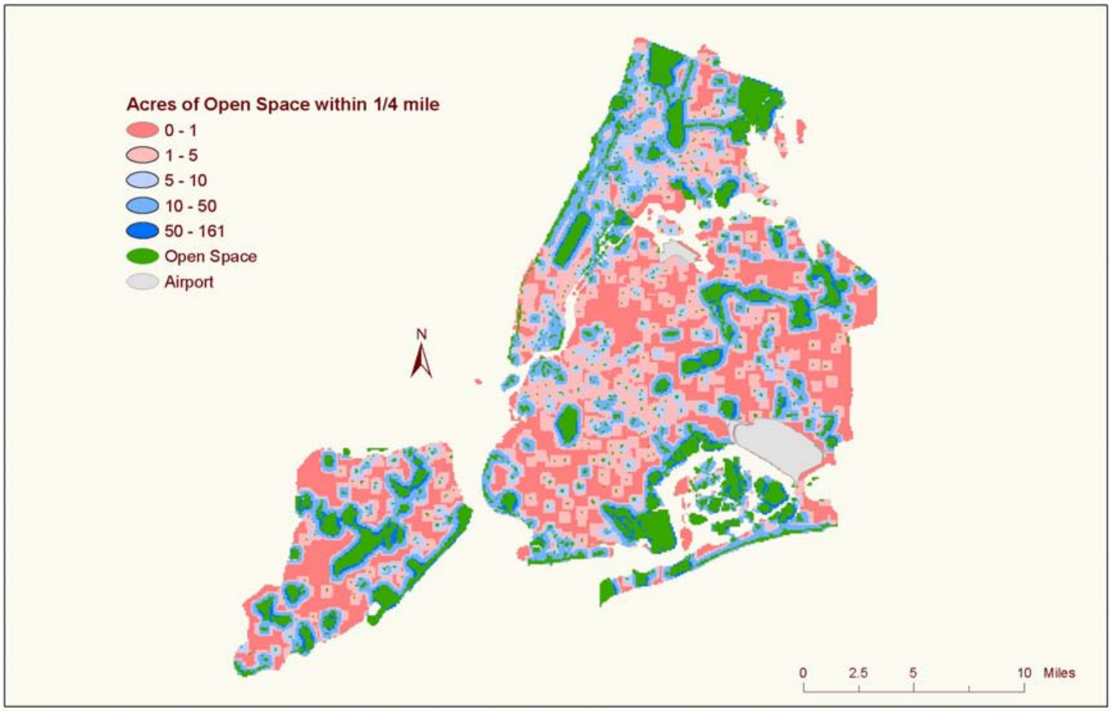
Figure 1.

Difference in the Acres per 1000 Residents Calculated by Us and the Department of City Planning


DIFFERENCE
🟥 0.5 to 7 acres less than DCP
⬜ 0.5 acres less than DCP to 0.5 acres more than DCP
🟦 0.5 to 6.3 acres more then DCP
🟩 OPEN SPACE
N
0 2.5 5 10 Kilometers

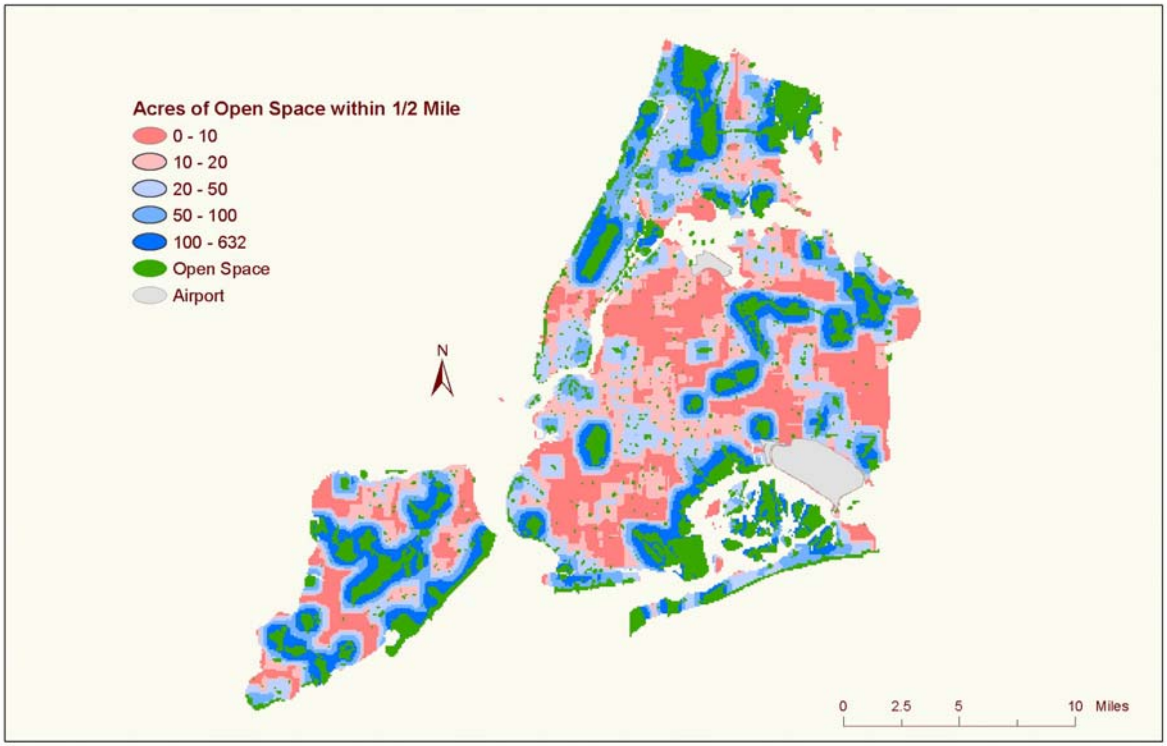
Map created by Sara Hodges March 2003 Data Projected to NAD 1983 New York Long Island State Plane
Funding Provided by the New York City Environmental Justice Alliance

Figure 2.



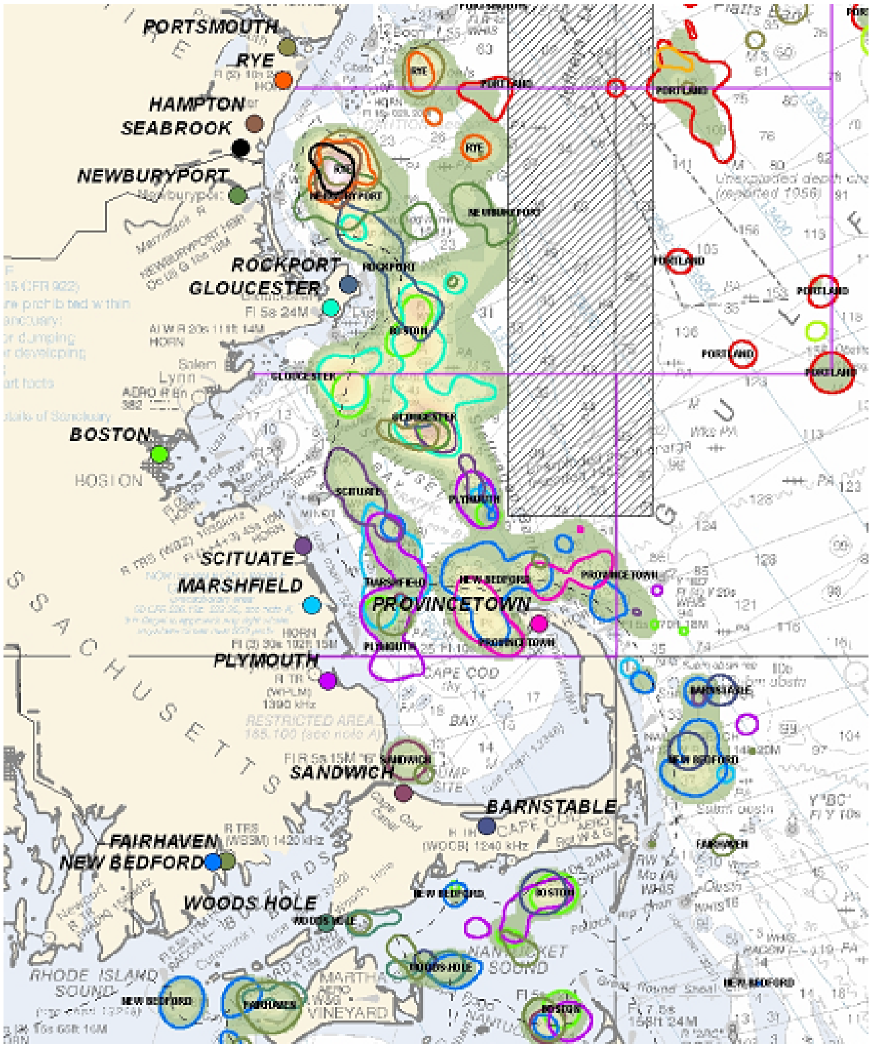
Map created by Sara Hodges
Data Sources: NYCDPR, 2003; CENYC, 2003; NYCDOE, 2003; CARSI, 2002, ComCarto, 2003.

Figure 3.



Map created by Sara Hodges
Data Sources: NYCDPR, 2003; CENYC, 2003; NYCDOE, 2003; CARSI, 2002, ComCarto, 2003.

Figure 4


PORTSMOUTH
RYE
HAMPTON
SEABROOK
NEWBURYPORT
ROCKPORT
GLOUCESTER
BOSTON
SCITUATE
MARSHFIELD
PROVINCETOWN
PLYMOUTH
SANDWICH
BARNSTABLE
FAIRHAVEN
NEW BEDFORD
WOODS HOLE
PORTLAND
CAPE COD BAY
NANTUCKET SOUND
MARTHA'S VINEYARD
RHODE ISLAND SOUND
MASSACHUSETTS

Endnotes

[1] See, for example, Elwood, S. A. 2006. Beyond cooptation or resistance: Urban spatial politics, community organizations, and GIS-based spatial narratives. *Annals of the Association of American Geographers* 92, (6): 323-41.

[2] A discussion of these and other methodological issues relevant to secondary data can be found in Pavlovskaya, M. E. 2002. Mapping urban change and changing GIS: Other views of economic restructuring. *Gender, Place and Culture: A Journal of Feminist Geography* 9, (3): 281-89.

[3] For a conceptualization of economic diversity see Gibson-Graham, J. K. 2006. *A Postcapitalist Politics*. University Of Minnesota Press.

[4] For a discussion of mapping and privacy see Cromley, E. K., and S. L. McLafferty. 2002. *GIS and public health*. New York and London: Guilford Press, especially pp. 207-209.

[5] See Scott, M., S. L. Cutter, C. Menzel, and M. Ji. 1997. Spatial Accuracy of the EPA's Environmental Hazards Databases and Their Use in Environmental Equity Analysis. *Applied Geographic Studies* 1, (1): 45-61, for an analysis of such misrepresentation.

[6] Wong, D. W. S. 2004. The modifiable areal unit problem (MAUP). Ch. 93 in *Worldminds: Geographical perspectives on 100 problems*. eds D. G. Janelle, B. Warf, and K. Hansen, 571-78. Dordrecht, Boston, London: Kluwer Academic Publishers.

[7] See Shekhar, S., and S. Chawla. 2003. *Spatial databases: A tour*. Upper Saddle River, NJ: Prentice Hall.

[8] Shekhar, S., and S. Chawla. 2003. *Spatial databases: A tour*. Upper Saddle River, NJ: Prentice Hall.

[9] See MacEachren, A. M. 1995. *How maps work: Representation, visualization, and design*. New York: The Guilford Press; Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2005. Geovisualization. Ch. 13 in *Geographic information systems and science*. 2nd ed., 289-313. John Wiley and Sons.