

Digital Cities I : Integrating Data and Information Resources

Roberta Balstad Miller¹ Christopher Small²

¹ CIESIN, Columbia University
61 Rte. 9W, P.O. Box 1000, Palisades, New York, 10964 USA
Tel: (914) 365-8950, Fax: (914) 365-8922
E-mail: roberta@ciesin.columbia.edu

² Lamont-Doherty Earth Observatory, Columbia University
61 Rte. 9W, P.O. Box 1000, Palisades, New York, 10964 USA
Tel: (914) 365-8354, Fax: (914) 365-8179
E-mail: small@ldeo.columbia.edu

ABSTRACT The past century has seen a rapid expansion in the number and size of cities. If the concept of Digital Earth is to present a full and accurate picture of the Earth and its environment, it cannot treat cities the same as other parts of the Earth's surface. Digital Earth must encompass Digital Cities, including the use of both remote sensing images and socioeconomic data to portray them. This paper discusses 1) the nature of the socioeconomic data needed to portray urban areas within Digital Earth, 2) approaches to the integration of these data with remote sensing data, and 3) problems encountered in providing information about the dynamics or processes of change in urban areas. It concludes that the success of Digital Earth depends upon its capacity to encompass Digital Cities as well as the less densely settled portions of the Earth. (For reasons of space, it was not possible to include the images that illustrate the points made in this paper.)

KEY WORDS Digital Earth, cities, urbanization, socioeconomic data

The concept of Digital Earth is global. By definition, it encompasses the entire Earth, including large expanses of oceans, deserts, the cryosphere, forests, rangelands, agricultural land, and other types of land surface and land cover. Most of the surface of the Earth is not densely settled. Roughly 50% of the Earth's population live on 4% of the ice-free land area (excluding Greenland and Antarctica), and 74% of the population (or three out of four people) live on only 10% of the land. Moreover, over 60% of the Earth's land area is effectively uninhabited, with regional densities at less than 10 people per square kilometer (Small and Cohen, 1999; Tobler, et al, 1998). These low densities simplify the data problems in creating Digital Earth. For the large portion of the Earth's surface covered by a thin layer of human settlement, images and digital data can be obtained through Earth observation satellites to characterize specific surface features and ultimately to make it possible to monitor changes. New observation instruments such as MODIS, Landsat 7, ASTER, and several recent or about to be launched commercial instruments will significantly improve our ability to obtain multispectral data that can be used to characterize the sparsely settled surface of the Earth. But the nature of the human presence on the Earth has been changing, particularly over the past fifty years of unprecedented population growth and rapid urbanization. The total global population

has risen rapidly and is now considered to be six billion people. More important for Digital Earth, however, is the fact that the concentration of the population in cities has also been increasing during this period. In 1890, there were only nine cities in the world that had a population of over one million people; by 1980, there were 230 cities in this category (Ponting, 1992). In 1900, the total urban population of the world is estimated to have been 160 million, and by 1985, this had increased more than tenfold to 2,250 million. It is estimated that 61% the world's population will live in large cities in 2025 (O'Meara, 1999).

This post-World War II urban population growth has had significant consequences for the environment, both regionally and globally (Cohen, 1995; Goudie, 1994). Cities have always exerted a disproportionate social, political, and economic influence on the areas that surround them and on the regions and nations of which they are a part (Miller, 1979); but their impact has increased with the development of megacities in the twentieth century. Many cities dominate their hinterlands and influence a wide range of economic, agricultural, and industrial activities within the region through their economic and political power. An example is Karachi, which produces 20% of the GDP of Pakistan and half the revenues of the government (Linden, 1996). Urban and suburban growth absorbs the agricultural land surrounding cities,

resulting in an absolute decline in agricultural acreage in the region. Cities alter their immediate atmosphere by changing the nature of the land's surface and consequently its reflection, radiation, and aerodynamic properties, creating an urban heat island that raises urban temperatures and traps pollutants in the urban atmosphere (Oke, 1982; Berry, 1990). Cities also cause waste disposal problems, alter surface run-off and water quality, and negatively affect regional biodiversity.

Globally, cities exert a disproportional influence on the global environment through their contribution to greenhouse gases. For example, although cities cover only 2% of the surface of the Earth, they account for roughly 78% of the anthropogenic carbon emissions. They are also resource intensive and are responsible for 60% of human use of water and 76% of industrial use of wood (O'Meara, 1999).

From the perspective of the residents, urban environments differ in significant ways from non-urban environments. Modern densely settled cities and suburbs have more contaminated air, less radiation and sunshine, more clouds and fog, more precipitation, and less wind than non-urban areas. The presence of particulates and condensation nuclei in the air, waste disposal, and water quality problems can have serious public health impacts, especially in economically vulnerable portions of the population (Berry, 1990).

Because of their influence on the Earth and its environmental systems, cities cannot be ignored in developing Digital Earth. However, Digital Cities pose a unique set of problems that must be addressed in planning for Digital Earth. Remote sensing data sources that are appropriate for much of Digital Earth are often not applied in urban areas (Warnecke, 1997). Moreover, less attention is being given to the development of urban applications of the newer remote sensing instruments than to applications in land use, agriculture, oceans, and atmosphere (Foresman, 1999; Gubbels, et al., n.d.).

But even if their use in cities were more widespread, remote sensing images alone could not provide the range of data needed for Digital Cities. Many of the ways that cities influence the environment are the result of cumulative human actions based on social, institutional, and economic motivations; that is, they are non-physical in origin. These environmentally significant activities in cities are difficult or impossible to detect using Earth observation satellites. Not only must the data needed for Digital Cities be obtained from multiple sources and sensors, but in addition, socioeconomic, in situ, and remote sensing data

must be integrated into digital urban data bases. The use of socioeconomic data in conjunction with remote sensing data is difficult because the two types of data are collected for different spatial units, at different spatial and temporal resolution, and are calibrated according to different criteria.

The focus of this paper is on the problems that will be encountered in integrating the multiple types of urban data into data bases that can be a foundation for accommodating Digital Cities within the larger Digital Earth program. Our second paper (Small and Miller, 1999) will discuss ways to monitor the urban and suburban environment from space. In both papers, we use digital data on New York City to illustrate the issues discussed. The two papers are premised on the belief that Digital Earth can neither ignore the world's cities nor treat them as if they were no different than other parts of the Earth's surface. Unless there is significant attention paid to the particular issues raised by Digital Cities, Digital Earth will present an inaccurate and biased picture of the Earth and will be incapable of presenting changes over time. The basic differences between cities and the less densely settled parts of the globe are related to the socioeconomic and human complexity of cities and the ways that these interact with the physical environment. Multiple types of disparate data must be integrated in common data frameworks in order to visualize and ultimately understand this complexity. In this paper, we discuss three sets of issues related to data integration:

- A. The nature and diversity of the socioeconomic data that are needed to portray urban areas within Digital Earth;
- B. Approaches to the integration of these data with in situ measurements and remote sensing data;
- C. Problems encountered in the digital representation of the dynamics or processes of change in urban areas.

We have initially focused our attention on New York City for several reasons. Substantively, it is an important city to study. With a population of nearly 7.3 million people, New York City is the largest city in North America. The New York metropolitan region, which consists of 31 counties and encompasses 20 million people, constitutes 8% of the United States population (Yaro and Hiss, 1996). Like many large cities, New York is located on the coast and has 600 miles of coastline. Because of its demographic and physical size and the fact that it is at the center of the Boston to Washington conurbation (the so-called BosWash megalopolis), the city has a major regional environmental impact. A second reason for

examining New York City is that it is functionally a global city with economic and financial, social and cultural, communications, and even political ties with other cities and countries across the globe. Activities and decisions taken in New York can have ramifications in many parts of the world. The interconnections between this city others illustrates another important rationale for a focus on Digital Cities, that is, although they are often spatially separated, they are intersecting rather than isolated centers of anthropogenic activity. A third reason for selecting New York City for study is that the data that allow us to use it as a prototype for examining basic issues in data integration are easily available.

1. Data Needed for Digital Cities

The first issue is the type of data needed for Digital Cities and how these data differ from those used in other parts of Digital Earth. Remotely sensed images are a basic source of information for Digital Earth and for information about human activities in rural areas, however their use as a source of information on cities is less well established. Earth observations have been used to gauge changes in the physical extent of urban settlement (Newman, et al., 1999). This is most effective when it is used in rapidly developing urban areas where land cover changes are pronounced and traditional sources of demographic data are inadequate, outdated, or unavailable. Remote sensing data have also been used to measure changes in urban land use over time, though this requires fairly high spatial resolution (Cowan and Jensen, 1998). In our second paper, we discuss a third use of remote sensing images in urban areas, that is, for measuring the abundance of vegetation.

Important as they are, these uses of remote sensing images can at best provide observational data; that is, they record evidence of the physical and environmental phenomena that can be observed in a specific urban area. This includes, *inter alia*, the built environment, the transportation and utility infrastructure, vegetation, urban waterways and coastal areas, and land use. It is even possible, if the resolution is fine enough, to observe groups of people. But Earth observations images do not provide a means of determining how the configuration of physical phenomena in cities developed as it did or how the physical patterns and social practices within the city impact its land, water, and atmosphere. Nor can Earth observation images reveal what takes place within the urban built environment that is observed from space. To understand, visualize, and ultimately measure both change and impacts in Digital Cities, remote sensing data must be used in conjunction with in

situ and socio-economic data that are obtained from measurement and observation instruments that differ significantly in their operation and the data they produce.

Most socioeconomic data are obtained from one of three sources: 1) administrative, commercial, and governmental operations, 2) population censuses, or 3) sample surveys of the population. The three differ substantially as to their periodicity and the ways they can be integrated with other sources of data.

1.1. Administrative Data.

Administrative and governmental data are produced by the normal operations of government and its agencies and the private sector. They include, for example, vital statistics (birth, death, and public health records); utilities data, including water and energy usage (these are available by household or for the municipality); transportation data; residence and building permits; coastal, harbor, and water management data; weather and air quality data; and data on other activities that are either regulated by government or monitored for commercial purposes. Increasingly, in many cities, these data have been placed in Geographic Information Systems, although the GIS are frequently not made available to the public. Many of the urban administrative data sets are used by and may be obtained from local planning authorities and non-governmental organizations that track local policy issues. In New York, for example, the Regional Plan Association maintains a number of data series taken from administrative data in the metropolitan area (Yaro and Hiss, 1996).

1.2. Census Data.

Population censuses are enumerations of the total national population and are conducted periodically (usually at ten-year intervals) by most governments. They vary in quality, but because of their official status, scope, and periodicity, constitute a basic source for population data over time. In the United States, the Census Bureau is the only federal statistical agency to provide sub-county data (Cortright and Reamer, 1998).

1.3. Survey Data.

Surveys consist of questions asked of individuals about their behavior, background, perceptions, expectations, and/or attitudes toward a particular issue or subject. The answers to these questions provide subjective data on the individual or the household, but unless the sample is very large, not on sub-sections of the city. The data are called subjective because they consist of individuals'

assessments and cannot be externally calibrated. Unlike censuses, which attempt to obtain data on the entire population, sample surveys are based on data collected from a small subset of the population, often a probability sample of the total population, from which assumptions for the total population can be based. Surveys are usually conducted to meet a specific information need rather than for general purpose monitoring. Surveys can be focussed on an entire country or on specific cities.

These three types of socio-economic data differ from earth observation data in a number of ways, but an important difference concerns the spatial unit of analysis. Remote sensing images are raster-based, which is compatible with their being expressed as a grid; socio-economic data are vector-based and describe the area within the boundaries that describe a specific political jurisdiction. In practice, this means that the spatial coverage of remote sensing data is regular and contiguous, whereas the spatial coverage of socio-economic data sets is characterized by irregularities and is determined by a mixture of historical, demographic, bureaucratic, military, commercial, and ecological factors. The last includes rivers, coastlines, and other natural features which have historically served as political boundaries.

Socioeconomic data that will be useful for Digital Cities can be broadly grouped as economic, demographic, political/institutional, and infrastructural data. In the category of economic data, for example, are economic transactions; household wages, income, and wealth; individual, household, and city consumption; economic transfers and interactions within the region, the nation, and internationally. Examples of demographic data are population, births, and deaths; migration; public health; and age, gender, and education distributions. Infrastructural data include the features and characteristics of the built environment in cities, such as housing, industrial complexes, roads, and bridges (including information on their structural stability and age).

It is also useful to have information on political institutions because of their influence on environmentally significant human behavior through the enforcement of laws, regulations, and common cultural patterns of behavior. Political/institutional data of importance include information on local and national government; legal and regulatory institutions; and the impacts of government policies, such as taxation. For analytical purposes, however, it is important to recognize that knowing the areal coverage of systems of law and regulation is not

the same as knowing the levels of enforcement or compliance within those units.

2. Approaches to the Integration of Socioeconomic with Remote Sensing Data

In order to combine or integrate data from these multiple, diverse sources, there must be a common framework for spatial representation and visualization of the data. Although this is a requirement shared by all aspects of Digital Earth, the task becomes more complicated when both physical and socioeconomic phenomena must be presented in the same data base. The integrating framework must first be capable of displaying attribute data, point measurement, vector data, and raster data. Second, it should be able to accommodate data at multiple time periods and for spatial areas of varying size and shape. We will briefly discuss two approaches to this type of integration: 1) gridding socioeconomic data and 2) using a GIS for data visualization.

To facilitate comparisons among gridded remote sensing images, such as the AVHRR land cover data base, and socioeconomic point data such as population, it is possible to project the socioeconomic data onto the same grid as the remote sensing data. An example is the one-kilometer gridded data base of the US population, produced by CIESIN in 1994. This data base used the physical coordinates of both the AVHRR and the US Census Bureau's 1992 TIGER files as reference points. Population data were taken from the 1990 US Census and registered to the grid. Where polygons were bisected, the data were proportionally allocated. Among the advantages of the gridding approach are that it provides a visual comparison of dissimilar types of data, and once the gridding algorithm has been created, it can be applied to other socioeconomic data from the same data collection instrument. In this case, data on the number of households in the United States were also spatially mapped to a one-kilometer grid. One of the disadvantages of this approach is the requirement that there be two independent spatial reference systems. If either has errors, those errors will impact the entire data base.

The most practical and widely available tool for data integration is GIS, which can be used for both remote sensing and socioeconomic data (Liverman, et al., 1998). Instead of fitting one type of data (socioeconomic) to another (gridded satellite data), a GIS is capable of managing, manipulating, and displaying a variety of spatially referenced information. It provides a means of comparing and visualizing disparate data types, such as population and housing density, in a spatial framework. Even

political and institutional data, which differ in significant ways from demographic and economic data, can be entered into a GIS. Because legal systems are equally applicable across entire political units, which can be referenced as spatial areas, legal attributes of the defined physical space can be represented in a geographic data system.

3. Representing the Dynamics of Change in Digital Cities

It is useful for many purposes to have digital data on cities in a GIS, but by themselves, these data are static and hence inadequate to trace the functions or the impact of Digital Cities. Understanding the dynamics of these patterns and processes requires observations at multiple time periods, which can be used to create time series. Acquiring time series data and using them in conjunction with Earth observation images is complicated by the fact that socioeconomic and physical measurements generally have different temporal resolution. Remote sensing data are relatively continuous—examples are SPOT and Landsat 7 which are available at biweekly intervals. New sources of continuous data will be available when NASA's EOS sensors are launched. Even Ikonos, launched in by a private sector firm in September, 1999, will provide data on a frequently recurring basis. Although it is relatively easy to obtain socioeconomic data over spans of decades, remote sensing data have generally been available only since the early 1970s. In some cases, early observations obtained from space have not been preserved. In other cases, they are considered classified or have deteriorated and are unavailable. There a need to ensure that a carefully selected sample of remote sensing images obtained several decades ago are preserved and properly archived so that they can be made available for multi-temporal comparisons in the future.

In general, it is easier to obtain socioeconomic time series data. Government data are generally archived and are often publicly available for periods of hundreds of years. The periodicity of the time series differs from that of the remotely sensed data, and the data may be available at irregular intervals. Population censuses are generally conducted every ten years, however, and the data can be used to create time series with ten-year intervals. Local and regional estimates of population can be used to create a finer-grained times series. In New York City, these data show the rapid expansion of the city through a process of suburbanization.

Administrative data are available at mixed periodicity. These data may be collected at hourly, daily, monthly, or annual intervals, although the

difficulties in using such large data bases and concerns over privacy and commercial competitiveness is such that only subsets of these data are usually released to the public.

Because sample surveys are expensive to conduct, they often are limited to a single data point that is not repeated. Analysts seeking to create a time series with unrelated surveys can compare responses to both the basic background items and repeated questions from survey to survey. However, the sample population will differ in each survey, and as a result, it is possible to characterize change in a population between surveys, but not to describe the experiences of specific individuals across time. Moreover, the timing of sample surveys is often irregular or not analytically meaningful.

A problem encountered in creating dynamic data sets on Digital Cities is that we need fine-scale socioeconomic data to link with coarse Earth observation data for specific points in space. For example, if we seek data that provide information on the way household practices affect air and water quality in a major city, we would need to link georeferenced household data from multiple sources. This would include aggregated data from enough households to make assumptions about the larger urban population, including the number and ages of residents in households, their food consumption and preparation patterns, their heating and cooling demands and energy costs, their income, and their expectations for future consumption. Data from the 1990 US Census to address some of these issues can be obtained for Census Tracts and Census Block Groups in New York City (See <http://www.plue.sedac.ciesin.org/plue.ddviewer>).

Creating an integrated data set with all these types of data is complicated by a concern in most countries for protecting the privacy and confidentiality of the subjects of data collection. In the United States, the release of microdata or individual-level data from the decennial censuses, which provide specific household demographic statistics, is prohibited by law for 72 years. Survey data, which could provide information on consumer expectations, would at best be obtained through a survey of a random sample of the city's population (and would at worst be estimated from a national sample), but such a sample would not necessarily be random with regard to spatial criteria. Moreover, the geographical identifiers in surveys are often suppressed when survey data are released to prevent identification of individuals, so it is difficult to estimate responses from sub-sections of the city. At the individual level, the fewest problems might

be encountered in obtaining administrative data, which could be used for monitoring energy or water consumption. However, many utility companies are unwilling to release their data and even if they were to do so, linking household consumption of energy or water to specific geographic locations could be time consuming and expensive. Finally, it would be necessary to learn what local municipal practices are in terms of waste disposal or energy restrictions.

Although a GIS can be used to integrate multiple layers of data in space, it has significant limitations in dealing with the analysis of data over time. One way of circumventing this problem is to create multiple GIS layers for successive time periods, comparing data on the same phenomenon over multiple time periods rather than data on multiple phenomena in space.

The problems in using socioeconomic data in conjunction with remote sensing data in Digital Earth are significant. There is a need for research that looks both into the measurement and integration problems raised here and into more substantive issues. This would include characterizing types of urban areas and their environmental impacts in the city and its hinterland and the distribution within specific metropolitan areas (including cities, suburbs, and the urban hinterlands) of anthropogenic activities. It is clear that today's large and growing cities are a critical element in shaping the Earth and its environment. For this reason, the success of Digital Earth depends upon its capacity to encompass Digital Cities as well as the less densely settled portions of the Earth's surface.

References

- Berry, B. 1990, *Urbanization in Turner, B.L., II, et al., The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the Past 300 Years: 103-119*, New York: Cambridge University Press.
- Cohen, J. 1995, *How Many People Can the Earth Support?*, New York: Norton.
- Cortright, J. and A. Reamer, 1998, *Socioeconomic Data for Understanding Your Regional Economy*, Washington: Department of Commerce.
- Cowan, D. And J. Jenson, 1998, *Extraction and Modeling of Urban Attributes Using Remote Sensing Technology*, in Liverman, et al., *People and Pixels: Linking Remote Sensing and Social Science*, Washington: National Academies Press.
- Foresman, T. 1999, *Government and Planning Applications for Remote Sensing*, EOM, 8(3):30-33.
- Goudie, A. 1994, *The Human Impact on the Natural Environment*, Cambridge, MA: MIT Press.
- Gubbels, T. et al., n.d., *Putting NASA's Earth Science To Work: Remote Sensing Applications*, Washington: Raytheon.
- Linden, E. 1996, *The Exploding Cities of the Developing World in Foreign Affairs*, Jan/Feb.
- Liverman, D. et al., 1998, *People and Pixels: Linking Remote Sensing and Social Science*, Washington: National Academies Press.
- Miller, R. B., 1979, *City and Hinterland: A case study of urban growth and regional development*, Westport: Greenwood Press.
- Oke, T. 1982, *The Energetic Basis of the Urban Heat Island*, *Quarterly Journal of the Royal Meteorological Society*, 108 (1).
- O'Meara, M. 1999, *Reinventing Cities for People and the Planet*, Washington: Worldwatch.
- Ponting, C. 1992, *A Green History of the World*, New York: Penguin.
- Small, C. and J. Cohen, 1999, *Continental Physiography, Climate and the Global Distribution of Human Population*, this volume.
- Small, C. and R. Miller, 1999, *Digital Cities II: Monitoring the Urban Environment from Space*, this volume.
- Tobler, W. et al., 1997, *World population in a grid of spherical quadrilaterals*, *International Journal of Population Geography*, v. 3, pp. 203-225.
- Warnecke, L. 1997, *NASA as a Catalyst: Use of Satellite Data in the States*, Washington: NASA.
- Yaro, R. and T. Hiss, 1996, *A Region at Risk: The Third Regional Plan for the NY-NJ-CT Metropolitan Area*, New York: Regional Plan Association.