OnTheMap: An Innovative Mapping and Reporting Tool

Jeremy S. Wu, Ph.D.[*]
Matthew R. Graham[†]
U.S. Census Bureau

Abstract

Statistical innovation produces new opportunities for advancement of knowledge and policy improvement. Twenty five years ago, a nineteenth century map and chart was cited as "the best statistical graphic ever drawn" because it inspired the visual display of a relatively large amount of data in limited space. It tells a story. Today, modern information technology has ushered in a dynamic, online mapping and reporting tool called *OnTheMap* that can analyze and present an unprecedented amount of detailed data in a short span of time and space. The resulting maps and reports can tell many stories. *OnTheMap* is a product of a visionary idea at the U.S. Census Bureau and its partner states that a new, cost-effective, 21st-century statistical system can be built by integrating existing administrative records with census and survey data. Properly integrated, this statistical system can be better than its individual parts. In turn, ensuring public availability of detailed statistics derived from this system has stimulated the development of state-of-the-art methods to protect confidentiality. In this regard, *OnTheMap* symbolizes continuing American ingenuity to innovate with new statistical data, methods, and dissemination tools.

This work is unofficial and thus has not undergone the review accorded to official Census Bureau publications. All results have been reviewed to ensure that no confidential information is disclosed. The views expressed in the paper are those of the authors and not necessarily those of the U.S. Census Bureau. We wish to express gratitude to John Abowd, Nancy Gordon, Ron Jarmin, and Kathleen Wallman for their helpful comments and edits. All remaining errors belong to the authors.

---

[*] Jeremy S. Wu is an Assistant Division Chief in the Center for Economic Studies (CES) at the U.S. Census Bureau and is the Program Manager for the Longitudinal Employer-Household Dynamics (LEHD) Program. He can be contacted at <Jeremy.S.Wu@census.gov>.

[†] Matthew R. Graham is a Geographer with the Longitudinal Employer-Huosehold Dynamics (LEHD) Program at the U.S. Census Bureau and can be contacted at <Matthew.Graham@census.gov>.

OnTheMap: An Innovative Mapping and Reporting Tool

I. An Exemplary Statistical Dissemination Tool

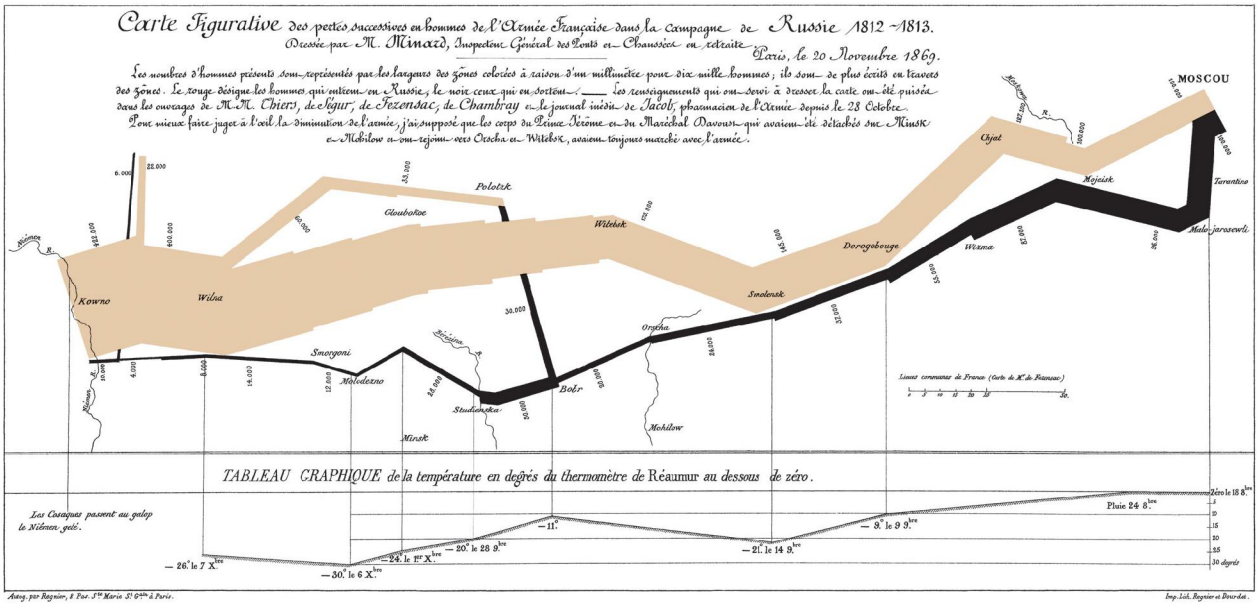
If a picture is worth a thousand words, then a data map should be worth at least ten times more.

According to Tufte [1], a two-dimensional map about the United States and its approximately 3,000 counties will contain at least 12,000 numbers, because a county will require at least 4 points to describe even with no other information in it.

Charles Joseph Minard (1781-1870) showed Napoleon's Russian campaign of 1812 in one of his well-known "Carte Figuratives" in Figure 1 below. This combination of a map and a chart tells the full multivariate, time series story about the location and size of the army at the start of the campaign in June 1812, its changing size and direction of movement of the army over time, and temperatures on selected days of the army's retreat from Moscow.

Tufte [1] cited this map and chart as perhaps "the best statistical graphic ever drawn" in his 1983 book on *The Visual Display of Quantitative Information.*

Figure 1: Minard's Map and Chart on Napoleon's Campaign of 1812



II. Today's Needs for Official Statistics

Today as we try to build a more skillful workforce and develop more effective policies at both the local and the global level, we require informative maps illustrating the movement and change of economic forces at work. For example, we offer these questions:

Where are jobs located? Where do workers live and work? What are the characteristics of the workers and employers? How have these numbers and locations changed over time? These are just a few fundamental questions for understanding the dynamics of a nation's economy and society, as well as its global competitiveness.

The United States has about 150 million workers and 190 million jobs. These workers are the engine of the U.S. economy. Traditional approaches to provide public-use official statistics on jobs and workers [e.g., 2] often have been point-in-time estimates provided in multiple tables and occasional static maps for well established political boundaries, such as states and counties.

However, there are three continuing, interacting trends:

1. Increasing demands for geographically flexible data

In the United States, the county is the next level of local government below the state. However, counties vary widely in population and physical size. The largest of the over 3,000 counties in the United States has a population of 10 million in California, while the smallest about 100 in Texas. Therefore, the county is not always the most appropriate geographic unit for reporting and comparing public-use official statistics.

In particular, data needs for today's regional or local economic development tend to go beyond the traditional political boundaries of state and county. For example, the U.S. Department of Labor's Workforce Innovation in Regional Economic Development (WIRED) Initiative [3] integrates talent and skills development into the larger economic strategies, focusing on labor market areas comprised of multiple jurisdictions (partial or whole counties) within states or across states borders.

Natural or man-made disasters also do not recognize political boundaries. Data and visual aids required for emergency response or management of a hurricane or mountain wildfires usually require more flexibility in the underlying geography.

Geographic boundaries are not static. A hurricane may reshape the coastlines, or a city may expand by annexing a smaller town or by extending its boundaries into unincorporated land. In either case, geographies must always satisfy the most current data needs.

2. Easy visualization of massive amount of data

While the amount of public-use data grows exponentially with the increase in geographic flexibility, so does the need for easy, rapid, remote, and secure access to the data according to the needs and wishes of the user.

Currently, the common approach is to release public-use data files through the World Wide Web and provide links for download. Examples include the 2000 Census Transportation Planning Package distributed by the U.S. Department of Transportation [4] and web-based applications such as the American FactFinder [5], which provides tables and some mapping of data from the American Community Survey and other Census Bureau sources.

Files and tables, however, do not promote quick understanding of massive amount of data that have a spatial structure. Static data maps provide some visualization of massive amounts of

data. For example, the U.S. Department of Education [6] employs a mapping tool to show demographic and economic statistics for its School District Demographics System, although the information is provided by school district only. The U.S. Department of Agriculture [7] also maintains a gallery of static maps on crop acreage and a time series of maps on vegetation condition for the conterminous United States, but the user cannot select smaller geographies. The U.S. Geological Survey [8] has an excellent collection of maps and even a map maker for agriculture, biology, geology, hydrology, and other physical characteristics, but does not have the constraints of protecting confidentiality of respondents.

3. Strict protection of individual and business respondents' confidentiality

The dilemma of making detailed data available for public access and yet protecting the specific identity and actual attributes of the data provider has a long history in statistical disclosure limitation (avoidance) and privacy-preserving data-mining [e.g., 9, 10, 11, and 12].

Wu and Abowd [12] observed that "(b)usinesses and individuals, who supplied the original data either as direct respondents or as the source of administrative records, are entitled to this confidentiality protection according to both legal and ethical standards. The foundation of the American official statistical system is predicated on the trust that citizens place on the stewards of the data to uphold these standards."

While users want the maximum amount of data to be placed in the smallest available space in a written report or on a computer monitor, there is also an increased likelihood that confidentiality protection will prevent public access to the data in the first place. Concerns about potential disclosures partly explain the relative lack of dynamic, detailed depictions of demographic or business statistics on maps.

Each of these identified needs is non-trivial for development and implementation; together they pose a major challenge to statistical agencies from data creation to result dissemination.

III. The Coming of *OnTheMap*

OnTheMap is an innovative mapping and reporting tool that is designed to meet many of the growing needs for geographic flexibility and easy visualization of massive amount of data on jobs and workers in the United States, while still protecting respondents' confidentiality. The U.S. Census Bureau first released *OnTheMap* in February 2006, and has since expanded it several times. Version 3.2 of *OnTheMap* was released in December 2008. The online mapping and reporting tool is located at <http://lehd.did.census.gov>, and is available 24/7 with no direct cost to the user.

OnTheMap allows the user to select a defined geographical area, such as a county or a city, or a geographic area drawn freehand on the map by the user. *OnTheMap* supplies 22 common geographic boundary layers from which users may choose to select predefined geographies. These layers include cities/towns, counties, states, Congressional and state legislative districts, metropolitan/micropolitan areas, county subdivisions, postal areas (ZIP Codes), Workforce Investment Areas, WIRED Regions, school districts, census tracts, and traffic analysis zones, and Native American tribal areas.

The user may also choose up to three years for data viewing and analysis. When selected, an animation feature displays the series of maps automatically over the selected years.

Once the geography and time period have been selected, the user may optionally view and perform analysis on:

- ***Work Area Characteristics***. The number, spatial distribution, and characteristics of jobs or workers who are employed in the selected area.

- ***Residential Area Characteristics***. The number, spatial distribution, and characteristics of jobs or workers who reside in the selected area.

- ***Origin-Destination Travel Sheds***. The relationship between home and work areas for selected workers. Users may identify (a) where workers employed in a selected area live or (b) where workers living in the selected area are employed.

At the conclusion of the feature selections, two primary types of companion reports may be optionally produced for the selected time period:

- **Work or Home Area Profile Reports**. The number of jobs or workers in a selected work or residential area by year, and the corresponding distribution of jobs or workers by age categories, earnings categories, and industry type according to the 2-digit North American Industry Classification System (NAICS) codes [14].

- **Labor or Commute Shed Reports**. The number of jobs or workers in a selected work or residential area by year, and the corresponding top 10 locations to or from where the workers are expected to commute.

An Example

Suppose a user selects by freehand (inside the green circle) the Lower Manhattan area in New York City in Figure 2.

The companion Work Area Profile Report shows about 200,000 workers employed in this selected area in 2006. About 40 percent of these workers are in the Finance and Insurance industry; 15 percent are age 55 or over; and 65 percent earn a relatively high wage of more than $3,400 a month.

Figure 2: Freehand-selected Lower Manhattan Area

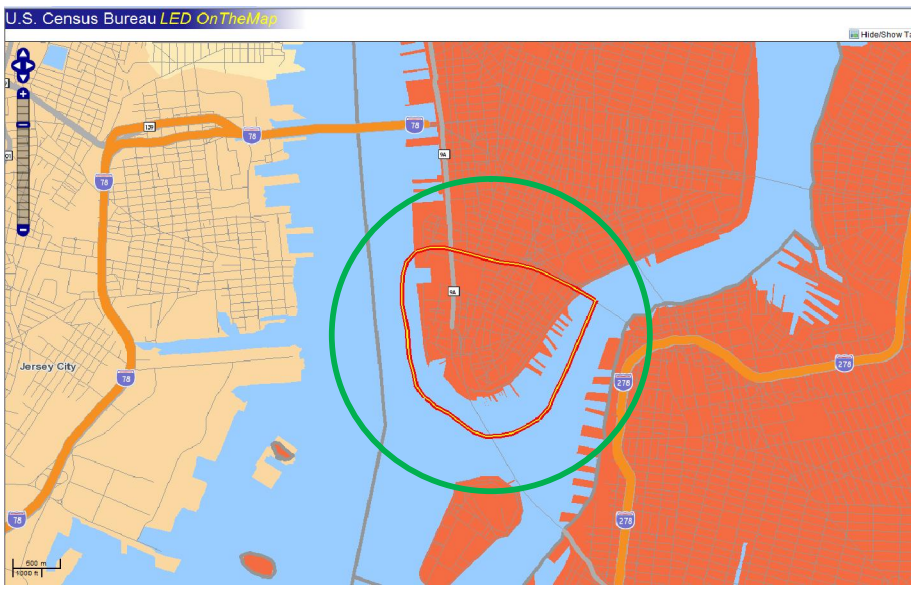
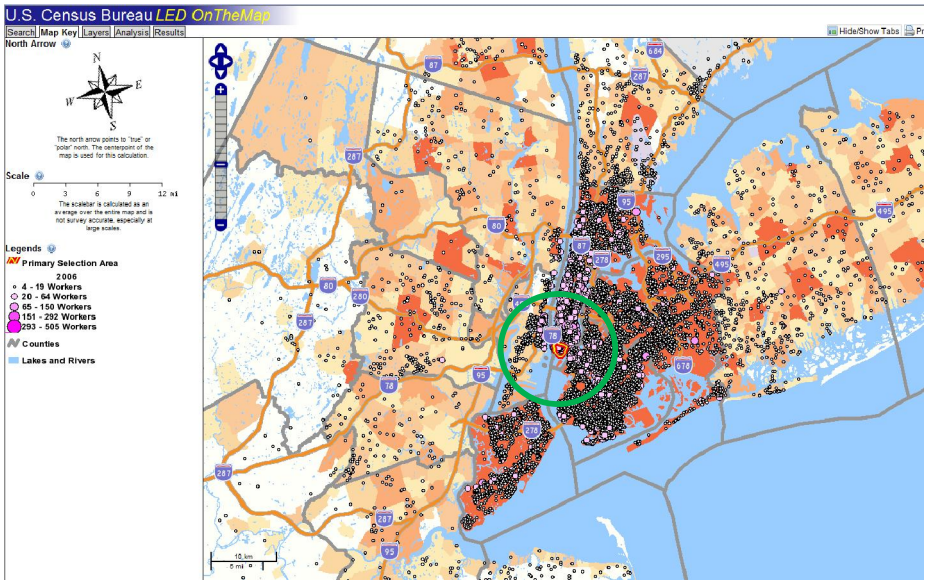


Figure 3 shows the residential locations of the 200,000 workers in dots of varying sizes, representing the range of workers at each location according to the legend on the left panel of Figure 3.

Figure 3: Labor Shed for Selected Lower Manhattan Area



The companion Labor Shed Report shows that 81 percent of these 200,000 workers live in New York state (63 percent in New York City), 16 percent in New Jersey state, and the remaining 3 percent in the states of Connecticut, Pennsylvania, Massachusetts, and other states.

Similar maps and Home Area Profile and Commute Shed Reports may be generated for the 14,000 workers residing in the selected Lower Manhattan area within a matter of seconds.

The Lower Manhattan examples tell several stories about local jobs and workers by employing only the most basic features of *OnTheMap*. According to Tufte's standard [1], the amount of data values easily exceeds millions in Figure 3 and its companion reports. The number of maps that can be dynamically produced by selected area and year of available data in *OnTheMap* is practically unlimited.

IV. Major Innovative Features of OnTheMap

1. OnTheMap offers dynamic mapping and reporting at the census block level

A census block is the smallest geographic unit used by the U.S. Census Bureau for tabulation of data. There are over 8,000,000 mutually exclusive and exhaustive census blocks in the United States. For reference, the hand-selected Lower Manhattan example above contains 181 census blocks and the city of New York has almost 36,900 census blocks.

The underlying data in *OnTheMap* are loaded and displayed by default at the census block level. A user-selected area will be defined by *OnTheMap* as the aggregate of census blocks in the selection. For example, each dot in Figure 3 represents a range of worker counts in a given census block. Statistics for the selected area are calculated in real time as aggregates of the corresponding census block statistics.

The data display and analysis may be optionally segmented for workers in categories defined by age, earnings, or industry. The display of dots may also be replaced or supplemented by chloropleth (thermal) contours to indicate different levels of densities (jobs or workers per square mile).

The *OnTheMap* approach offers practically unlimited possibilities that have never been available before in statistical dissemination. The modular design allows the inclusion of existing census-block-based geographies or the creation of new census-block-based geographies for custom reporting.

2. *OnTheMap data are derived from integrated data and a unique partnership*

OnTheMap is a product of a visionary idea at the U.S. Census Bureau and its partner states that a new, cost-effective 21st-century statistical system can be built by integrating administrative records with census and survey data, with the anticipation that the sum will be better than its individual parts.

Under a voluntary federal-state partnership known as Local Employment Dynamics (LED), the partner states supply historical and ongoing employment administrative records to the Census Bureau, which in turn adds value by integrating them with demographic and economic data from its available sources in censuses, surveys, and additional administrative records.

The emerging statistical system is a longitudinal national frame of jobs, designed to cover the employment history and characteristics of up to 150 million current workers, the payroll history and characteristics of more than 20 million current employers, and their relationship through jobs. There are well over 6 billion records in the current statistical system, which is rich in content, continuously growing, and cost-effective. The cost of processing a record under the LED partnership is a minute fraction of the cost of collecting a new survey or census response.

The Census Bureau and the LED partner states share responsibilities in data creation, quality, and use. Unlike traditional sampling frames that are used for selection of samples, new data and products such as *OnTheMap* are derived from this emerging 21st-century statistical system on workers, employers, and jobs.

The emerging LED statistical system shows that innovative development and implementation of integrated data can supplement time-tested census and century-old random surveys in the study of statistics.

3. *OnTheMap is protected by state-of-the-art disclosure avoidance methods*

Despite its inherent richness, data from this new statistical system cannot be loaded directly into *OnTheMap* and similar applications without adequate confidentiality protection. The possibility exists that an individual may be identified with loading of raw data.

The customary approach is to suppress the publication or use of the data in question to ensure disclosure avoidance, but this action also denies public access. Therefore, the Census Bureau has developed and implemented state-of-the-art disclosure avoidance methods.

OnTheMap is the first partially synthetic data product publicly released by the Census Bureau. Wu and Abowd [13] further explained that:

> "The synthetic data for *OnTheMap* are generated for each origin based on a Multinomial model for the origins whose probabilities are drawn from the posterior

Dirichlet (multivariate beta) distribution conditional on unique destination and employee-workplace characteristics. The conjugate prior distribution for the probabilities is also a conditional Dirichlet distribution which must have sufficient empirical support and whose parameters are specified for confidentiality protection. The likelihood function is the multinomial distribution conditional on destination and characteristics. Noise is already infused into the destination count of workers from the Quarterly Workforce Indicator protection system. Thus, only the origin data are synthesized, and *OnTheMap* may be described as a partially synthetic data product, using the current standard nomenclature."

OnTheMap is an innovative mapping and reporting tool that allows for unprecedented dynamic visualization of massive data. However, it would not be possible if it were not supported by a creative integrated data system that is derived from a unique federal-state partnership, as well as leading-edge synthetic data modeling and noise infusion techniques to protect confidentiality and preserve analytical validity.

In this regard, *OnTheMap* symbolizes continuing American ingenuity to innovate with new statistical data, methods, and dissemination tools.

V. References

[1] Tufte, Edward R (1983). The Visual Display of Quantitative Information. Graphics Press, Connecticut.

[2] U.S. Bureau of Labor Statistics. State and Metro Area Employment, Hours, & Earnings. Available at <http://www.bls.gov/sae/tables.htm>, on December 16, 2008.

[3] U.S. Department of Labor (2008). Workforce Innovation in Regional Economic Development Combined Regions. Available at <http://www.doleta.gov/pdf/1%20-%20WIRED%20Combined%20Fact%20Sheet.pdf>, on December 16, 2008.

[4] U.S. Department of Transportation (2008). Census Transportation Planning Package – Data Products. Available at <http://www.fhwa.dot.gov/ctpp/dataprod.htm>, on December 16, 2008.

[5] U.S. Census Bureau (2008). American FactFinder. Available at <http://factfinder.census.gov/home/saff/main.html?_lang=en>, on December 16, 2008.

[6] U.S. Department of Education (2008). School District Demographics System. Available at <http://nces.ed.gov/surveys/sdds/map00.asp>, on December 16, 2008.

[7] U.S. Department of Agriculture (2008). Research Highlights from the Census and Survey Research Branch and Geospatial Information Branch. Available at <http://www.nass.usda.gov/research/>, on December 16, 2008.

[8] U.S. Geological Survey (2008). Maps, Imagery, and Publications. Available at <http://www.usgs.gov/pubprod/maps.html>, on December 16, 2008.

[9] Duncan, G.T., de Wolf, V.A., Jabine, T.B., and Straf, M.L. (1993). "Report of the Panel on Confidentiality and Data Access," *Journal of Official Statistics*, 9, 271-274. Available at <http://www.jos.nu/Contents/issue.asp?vol=9&no=2>, on December 16, 2008.

[10] U.S. Office of Management and Budget (2001). Federal Committee on Statistical Methodology Working brochure. "Confidentiality and Data Access Issues Among Federal Agencies." Available at <http://www.fcsm.gov/committees/cdac/brochur10.pdf>, on December 16, 2008.

[11] U.S. Office of Management and Budget (2005). Federal Committee on Statistical Methodology Working Paper 22. "Report on Statistical Disclosure Limitation Methodology." Available at <http://www.fcsm.gov/working-papers/SPWP22_rev.pdf>, on December 16, 2008.

[12] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, (5):35{64,1977.

[13] Wu, Jeremy, and Abowd, John (2007). "Synthetic Data for Administrative Record Applications at LEHD." PN-2007-05. Joint Statistical Meetings Invited Session 82 presentation. Available at <http://lehd.did.census.gov/led/library/presentations/Wu-Abowd-20070831.pdf>, on December 16, 2008.

[14] U.S. Census Bureau. North American Industry Classification System (NAICS). Available at <http://www.census.gov/eos/www/naics/>, on December 16, 2008.