

White Paper:

POPULATION IN HOMELAND SECURITY RISK APPLICATIONS

January 2010

Prepared By:
Digital Sandbox, Inc.
8260 Greensboro Drive, Suite 450
McLean, VA 22102
Phone: +571-297-3800
Web: www.dsbox.com

I. Introduction

Toward the end of his tenure as Secretary of the Department of Homeland Security (DHS), Michael Chertoff gave an address at Georgetown University in which he stated that “the foundation of government is to protect people.” The guiding principle that led to the risk formula used in FEMA’s \$2.5B per year risk-based grants was that these funds were to be used to protect “people and things.” Indeed, population-based metrics account for 40% of the risk in the grants made to States and urban areas, more than any other single factor. Clearly, a thorough understanding of population is required to understand homeland security risks.

In scenario-based formulations of all-hazards risk, the characteristics of a local population have an important impact on risk. Areas known for concentrating large numbers of people at predictable times are attractive terrorist targets. Mass transit hubs have been targeted by Al Qaeda affiliates in London and Madrid and by Aum Shinrikyo in Tokyo. Outdoor markets, cafes, nightclubs and other gathering places have been the frequent target of terrorist attacks in the Middle East and Asia. Understanding the number, density, and movements of a population are central to characterizing and quantifying the risks due to terrorism.

Understanding demographic characteristics of a population can be more important than density in mitigating risks from natural hazards. While nature does not specifically target population centers, the consequences of a natural disaster can be most severe when large numbers of people are involved. The experience of Hurricane Katrina in New Orleans and of the 2008 hurricane season in the Gulf Coast demonstrated that there are segments of the population that are particularly vulnerable to the devastating effects of such disasters: those close to the coast, the poor, those without access to transportation, etc. While the seasonal influenza virus typically affects the elderly most severely, the H1N1 pandemic of 2009 was particularly dangerous to youth and to pregnant women. Preparing for such disasters can be especially difficult when a significant number of the targeted population do not speak English. Understanding these and other characteristics of one’s jurisdiction is essential to addressing the unique risks they represent.

As these examples illustrate, an in-depth understanding of population dynamics and demographics is vital to homeland security and emergency management. To support our customers, Digital Sandbox, the leader in software products and services for analytic risk management, is dedicated to maintaining the best and most current possible population estimates. This paper outlines Digital Sandbox’s approach to building unique and defensible population-related measures, including key demographic variables, which are used in our population risk models and other products.

II. Sources of Population Data

There are a variety of publicly-available sources of data on population in the U.S., and each source has its strengths and weaknesses. In evaluating population-related data for their suitability in homeland security

applications, careful attention should be made to the purpose for which each source was originally constructed, as this may suggest inherent capabilities and limitations of the data.

THE U.S. CENSUS BUREAU

A commonly-cited source of population and demographic information is the U.S. Census Bureau. Every ten years, the Census Bureau conducts a thorough, detailed, and authoritative count of the U.S. resident population. Mandated by the Constitution of the United States to apportion seats in the U.S. House of Representatives, the results of the decennial census are also used in myriad other government and non-government products, including homeland security risk analyses. While this decennial census produces very geographically-granular counts of population and very detailed demographic characteristics, the data are only gathered every 10 years. The most recent census was conducted in 2000, and the results were released beginning in 2002. In the intervening 10 years since the 2000 census was conducted, many population shifts have occurred due to birth and death rates, immigration and other population movements. While the 2010 census is now under way, it may be several years again before the results become available, at which point the data will already be somewhat obsolete. In non-decennial census years, the Census Bureau publishes estimates of several population measures, but these are neither comprehensive nor as granular as the decennial census data. The demographics of urban areas are sampled by the American Community Survey (ACS), another Census Bureau product, but this does not cover all areas of the country uniformly. A consolidated picture of the decennial census data and the annual updates that maintains the granularity and specificity of the decennial census data while being kept current to account for the population shifts between censuses is conspicuously lacking.

The Census Bureau provides population data at a number of different geographic levels, including State, County, Place, Tract, Block Group, and Block. At each geographic level, the entities are chosen to align with existing jurisdictional or physical boundaries. This makes reporting population by jurisdiction straightforward (its intended purpose), but it also inevitably results in entities at a given level with vastly different physical sizes. For instance, Arlington County, Virginia is only 26 square miles, while some of the boroughs (county equivalents) in Alaska are tens of thousands of square miles in size. Besides choosing to use existing jurisdictional or physical boundaries, where appropriate, the Census Bureau also tries to normalize each entity to a similar number of people. For instance, most blocks have tens of people, most block groups have a few hundred residents, etc. Normalizing to a set number of people, rather than to a common land area, ensures that the uncertainties in counting population are roughly equivalent across entities at a given level of geographic hierarchy. This is a strength of the Census Bureau's methodology.

The most commonly-referenced population metric produced by the Census Bureau is the total residential population of a given geographic region or jurisdiction. While this metric is simple and easily verified, it may not reflect the risks facing a region from a given hazard. According to Census data, no one lived in the World Trade Center or at the Pentagon in 2000, yet on September 11, 2001 2,729 people died in these buildings. Likewise, the residential population of an urban center may not reflect the human cost of a severe earthquake that occurs during the workday, when commuters, business travelers, students, and tourists may all be present. Illegal immigrants are notoriously

undercounted in the decennial census, yet their presence must be accounted for when planning for a severe epidemic, since they would surely impact the health care system. Finally, the residential population of popular tourist destinations systematically under-represents the typical number of people present.

LANDSCAN (OAK RIDGE NATIONAL LABORATORY)

While the Census Bureau maintains population for the U.S., another Federal government project tries to estimate population any place in the world using a totally different technique. The Oak Ridge National Laboratory's LandScan project produces population estimates using non-intrusive techniques such as the analysis of aerial imagery. By analyzing such observables as land use and light output at night, LandScan can produce internally-consistent estimates of worldwide population. Since the techniques measure the people who are actually present in an area, LandScan results naturally reflect the complexities of population movements. The data does not distinguish between residents, commuters, visitors, and other types of population; instead, it represents a consolidated view of the typical number of people present in an area. LandScan purports to estimate both daytime and nighttime populations, and the results are presented in a uniform orthonormal grid, the cells of which are approximately one kilometer wide on the ground. The use of an orthonormal grid is convenient for mapping applications and makes the results independent of the vagaries of political boundaries, but it also does not provide as much geographical granularity as the Census Bureau provides in its decennial census. Obviously, by using remote measurement, LandScan cannot reveal much about the demographic characteristics of a population, but it is an intriguing source of global population.

TYPES OF COMMERCIAL DATA SOURCES

Besides government-sponsored efforts, there are many commercially available sources of population data. While many vendors simply package existing sources of data for convenient use (e.g., in geographic information system applications), there are several efforts that are unique and complementary to Federal data sources. Several States and municipalities publish roadway traffic data, either periodically or in near-real-time. These data are consolidated by commercial data providers for traffic and route analyses (e.g., to set volume-based rental rates for billboards). While not directly related to population, traffic data can reveal interesting patterns in the movements of a population.

Related to traffic data, consumer spending data (including point-of-sale data, real estate occupancy, etc., etc.) gives an indicator of where people conduct their commercial activity throughout the day. Such data can sometimes provide very granular temporal indications of the movements of a population. While important for a retail operation, consumer spending data is difficult to convert to population data without making assumptions about how people spend.

One challenge facing users of Census population data is to account for new development between decennial censuses. An area that was farmland in 2000 may be converted to high-density housing in the subsequent years. While county or municipal population may be estimated by the Census Bureau, it is more difficult to determine where those new people live within the jurisdiction without additional knowledge. One technique to estimate the impact of

new development is to review local building permit applications, and there are companies which do this as a service. While the availability of such data is not ubiquitous, and not all approved applications are realized as new housing, there is some correlation between housing permits and new development that can be used to estimate changes in settlement patterns.

Another indicator of changing settlement patterns can be discovered by examining U.S. Postal Service (USPS) carrier route information. Similarly to the Census Bureau philosophy, postal routes are determined partly by jurisdictional considerations (e.g., zip codes), and partly by the approximate (semi-fixed) number of residents any single postal carrier can service. Postal routes are adjusted periodically as settlement patterns evolve to maintain efficiency in postal delivery, and the most current postal routes can be obtained from the USPS. Analysis of these routes can yield indications of new development as well as the presence of significant unaccounted-for (i.e., illegal immigrant) populations. These results, when combined with data from decennial censuses, can yield detailed demographic indicators on a zip code (or zip+4) basis, and such data is a mainstay of mass mailing campaigns.

III. The Digital Sandbox Approach

The Digital Sandbox approach to population-related data is uniquely tailored for homeland security applications, including quantitative analysis of risk and indicators of preparedness. We have not constructed a one-size-fits-all “universal population index,” and the approach we have taken may not be optimal for non-homeland security applications. We understand that most homeland security analysts require defensible and transparent data sources, but that they will also require the flexibility to explore effective populations by time of day, and to examine any number of demographic characteristics of the population that may make it more or less vulnerable to various hazard scenarios.

Our 11 years of experience with homeland security risk analysis has caused us to “expect the unexpected.” Where population density is important for terrorism scenarios, immigration and trans-border visitors are more important for border risk issues, and various age-related characteristics are important for pandemic flu scenarios. Experience shows that the next major hazard may require a novel metric to understand the nature of its risk to the country, and we have therefore constructed very fast and powerful analytic tools to create new population-based metrics as needed.

Based on our experience in homeland security, we have constructed our population data approach to have the following characteristics:

- **Defensible Data Sources** – Digital Sandbox uses open source, government data wherever possible; we avoid using restricted-use commercial data.
- **Uniform National Coverage** – Digital Sandbox restricts data sources to those that cover all parts of the U.S.; we do not want to have different data for one region than for another, and we want to maintain comparability between all entities at a given level of jurisdictional hierarchy.

- **Geographic Granularity** – We report population data at the smallest available geographic granularity available from authoritative government sources, the census block level, which in urban areas is roughly the size of one square city block.
- **Up-to-Date Demographics** – Unlike many methodologies that report data from the 2000 census, we incorporate annual Census estimates and American Community Survey (ACS)¹ data to “grow” data to the current year.
- **Real-Time Population** – We provide both daytime and nighttime population to account for the movement of commuters (and the presence of visitors).
- **Custom Metrics** – We retain the ability to create custom metrics as needed using our proprietary Geographic Metric Engine (GME)².

Ensuring the first two characteristics (defensible data sources, uniform national coverage) restricts us to using only Federal government data. Commercially-derived data cannot be guaranteed to be acceptable by all homeland security stakeholders, and State and local government data is not universally available. The Census Bureau’s decennial census provides excellent geographic granularity, but we must project the decennial results forward in time to produce up-to-date demographics. We use the annual Census estimates to guide our time evolution of the decennial census data, so that our results remain completely consistent with all Census Bureau products. We incorporate residential population and commuter data, along with an estimate of international visitors, all derived from Federal data sources, to estimate the daytime and nighttime populations of each region separately. Although more detailed time-of-day distributions are possible, daytime and nighttime populations typically represent two extrema in the population distributions and serve as useful guides for most homeland security risk analyses. Finally, Digital Sandbox has created a software application called the Geometric Metric Engine (GME) to enable an advanced user to create almost any mathematical combination of raw population and/or demographic data required to satisfy unique analytical challenges. For instance, Digital Sandbox used its GME to construct the quantity called “Population Index” used in FEMA’s grant programs, which is constructed as a density-weighted population count (population squared divided by land area) calculated for each individual Census block and aggregated over all blocks in a State or urban area.

IV. Methods

The following section details the techniques Digital Sandbox uses to arrive at its population data.

UP-TO-DATE RESIDENTS

¹ The American Community Survey (ACS) is a nationwide survey conducted by the U.S. Census Bureau designed to provide communities with a fresh look at how they are changing. It is a critical element in the Census Bureau’s reengineered decennial census program. The ACS collects and produces population and housing information every year instead of decennially.

² The Geographic Metric Engine (GME) is a proprietary software application developed by Digital Sandbox to calculate various metrics based on geo-coded data. The GME utilizes algorithms for a broad set of input data and equations, at user-defined calculation and aggregation levels. For any geographic-based data (e.g., population, demographics) the GME can compute at varying degrees of granularity and summary levels for a wide range of applications.

To calculate resident population Digital Sandbox begins with 2000 census data at the census block level. These are reported for the 8M+ census blocks in the country. Annually after 2000, some of the areas defined by the Census Bureau are updated to account for changing jurisdictional boundaries, fixing boundary errors, etc., and this has an impact on subsequent Census Bureau estimates. We track all changes in block boundaries and affiliation as published by the Census Bureau, and recalculate 2000 block populations based on updated block boundaries. Using subsequent estimates of jurisdictional³ populations, we calculate a jurisdictional organic growth rate from 2000 to the estimate year and we apply that rate to the populations of all blocks in the jurisdiction. The relative growth is estimated at the place level using trends in the annual population estimates.

The U.S. Census Bureau publishes a national POPClock⁴ which estimates the population of the entire United States up to the second. This clock uses the 2000 Census, modified based on trends from monthly estimates for birth rates, death rates, and net immigration equally distributed to the second. For example, the POPClock estimates that every 7 seconds a new child is born to a U.S. resident woman. Digital Sandbox uses the POPClock to update the national population total in real-time then distribute these totals to states, counties, places, block groups, and blocks based on the distribution of population of each of these levels of geography from the 2000 census and the relative growth of each based on monthly and annual Census estimates. This provides the most current and defensible estimation of residents at the census block level based on government sources.

COMMUTERS

Residential population needs to be augmented by daily population movements to account for the actual number of people in a given jurisdiction at a certain date and time. Digital Sandbox uses census 2000 commuter data at the tract level, which is the most geographically granular available. The tract level data is allocated down to the block level based on block land area, assuming that the commercial land usage of every block in a given tract is equal. While this is not always the case, it qualitatively reproduces the distribution of commuters in the highest-density tracts, such as in New York City. Commuter data is evolved from 2000 to the present using place-based growth rates as for residents.

VISITORS

Digital Sandbox uses a variety of sources to account for the average daily number of visitors to an area. Visitors include both business travelers and leisure travelers (tourists), and they can come from other places in the U.S. (domestic visitors) or from other countries (international visitors). Digital Sandbox has identified no Federal government source for domestic visitors, but some statistics may be obtained from commercial sources, including the

³ We use Census Bureau population estimates at the most geographically granular level published, the Census Place. Annual population estimates are also published by the Census Bureau at the county and State levels, but these are less geographically precise than the Census Place level. A Census Place corresponds roughly to incorporated places such as cities, boroughs, and towns, and minor civil divisions such as town and townships.

⁴ The U.S. POPClock is based on the national population estimates. The U.S. Census Bureau produces national population estimates annually using the latest available data on births, deaths, and international migration. Each year, they recalibrate the population clock when they release the new set of population estimates.

tourism and travel analytics company D.K. Shifflet and Associates. Since this is not Federal data, we do not include this data in our commercial products; but, using the GME application, we have used it to derive domestic visitor estimates for FEMA.

The number of international visitors is derived from several different government sources. Overseas visitors must fill out an I-94 immigration form, and these data are available from the Office of Travel and Tourism Industries (OTTI) in the International Trade Administration of the Department of Commerce. Visitors arriving from Canada are queried by the Canadian government's International Travel Survey, and "interior" visitors from Mexico are also captured through the I-94 cards. All three sources are used by OTTI in their report on travel and tourism, so Digital Sandbox adopts the same data. Finally, "border" visitors (day and overnight) can be estimated from border crossing data released by the Bureau of Transportation Statistics. These data usually indicate a destination state and in some cases an estimate of length of stay can be deduced. These are used to allocate average daily visitors to states. Within each state, the international visitors are allocated to blocks proportionally to the total number of (domestic) commuters destined for those blocks. (Border visitors are an exception: they are assumed to stay within urban area where they cross the border.)

The intra-state allocation of visitors is based on the assumption that both business and leisure travelers go to where the commuters go. For business travelers, this is intuitive, since both business travelers and commuters are presumably headed for centers of commerce. For leisure travelers, this is only approximately true. In areas where tourism is a significant component to the economy, such as Las Vegas and the Virgin Islands, the tourism centers and the commercial centers are identical. Thus, the assumption for leisure travelers holds best in the places where it is the largest component of population.

ILLEGAL IMMIGRANTS

There are approximately 11.5 million illegal immigrants in the U.S., which is nearly 4% of the total resident population of the country. In California and Texas the number of illegal immigrants accounts for 7.7% and 6.7% of their total resident population respectively. This figure can significantly increase the number of people within a given census tract or block. Digital Sandbox calculates illegal immigrant population by immigrant arrival data from DHS, visa data from the State Department, border crossing data from the Bureau of Transportation Statistics, and resident nationality data from the 2000 Census, American Community Survey, and annual characteristic estimates.

Within a given state, and for a given nationality, we allocate the estimated number of illegal immigrants to census blocks based on the racial composition of the legal residents of the state, as reported by the Census Bureau. For instance, if all people of Hispanic race are located in one county of a state by the Census Bureau, then all estimated illegal immigrants from any Hispanic country of origin (e.g., Mexico, Honduras, etc., but excluding Spain) would be allocated to that same county by the Digital Sandbox procedure. The assumption we make in performing such an allocation is that illegal immigrants are most likely to congregate in areas where their legal compatriots are.

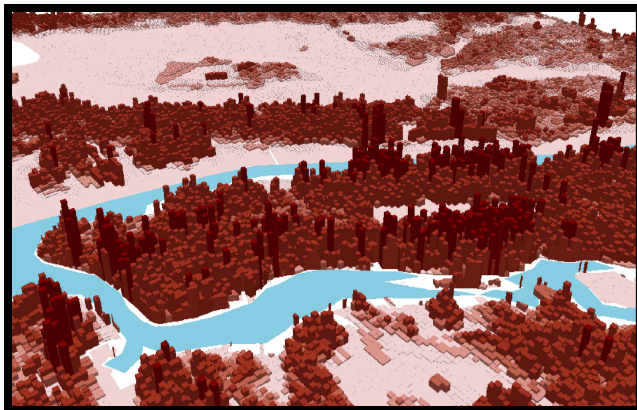
DAYTIME AND NIGHTTIME POPULATION

Once we have estimated the up-to-date populations (resident, commuter, visitor, immigrant) at the census block level, we combine them to create daytime and nighttime populations using our GME. We define daytime population as residents plus visitors plus net commuters (commuters coming into a block minus commuters leaving the same block), and we define nighttime population to include only residents and visitors. (We assume most visitors spend the night close to the locations they are visiting.) In our base population metrics, we do not include illegal immigrants, since not all homeland security applications call for them; but, at the user's request, we can include illegal immigrants into our population metrics using our GME.

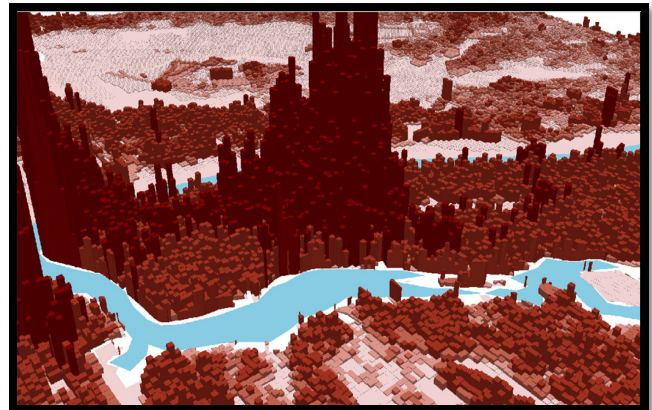
Many high-risk jurisdictions undergo significant fluctuations in population on a daily basis. For example, New York City (NYC) had a resident population of more than 8 million in 2008. But that number spikes during the workday by as much as 7%, with an additional 560,000 commuters coming in to work in the city. That spike from commuters alone is more than the total resident population of Atlanta, GA or Portland, OR. In addition to commuters, tourists can also significantly increase the population of some jurisdictions. On an average day there are more than 730,000 tourists in NYC. Therefore, there can be regular daily fluctuations in the population that can account for an increase of over a million people in the city. These fluctuations can significantly alter the homeland security risk profile of the city and change how city officials must plan for a number of hazard scenarios.

Digital Sandbox's population metrics allows users to identify population totals and demographic profiles based on time of day. The population profile of a census block adjusts due to "population movement" via travel and commute. This provides the ability to track commuter totals and demographics based on location and destination at the census block level. Users are able to set the time zone, time, date, year and any location in the country and return a defensible estimation of population for that area.

Manhattan Population – Nighttime



Manhattan Population – Daytime



REAL-TIME DEMOGRAPHICS

In addition to a current and defensible estimation of population totals at various levels of geography, Digital Sandbox estimates demographic variables as well. Certain demographic inputs are particularly valuable to our homeland security clients. Examples of these include age, gender, ethnicity, income, disability, employment, and

education level. All of these variables are published by the Census Bureau. The 2000 Census and the ACS provide detailed demographic profiles including over 2,000 variables. In addition, the Census releases annual characteristic estimates at the county level for age, gender, and ethnicity. Digital Sandbox uses a combination of these three sources as well as the up-to-date population estimate to produce a defensible estimation of demographics.

Digital Sandbox uses the 2000 Census and grows all geographic levels to the current real-time population through the GME. Then Digital Sandbox modifies the population with characteristic estimates released annually for three variables at the county level: age, gender, and ethnicity. Whenever possible, Digital Sandbox then replaces this updated data with the most current ACS data. The ACS does not release data for all levels of geography — just place level and higher and only those that reach certain population thresholds. Using a “waterfall” approach, Digital Sandbox distributes a 1-year estimate for all levels of geography with population >65,000 and the ACS 3-year estimate for the remaining levels of geography with population >20,000. The chart at the right lists some of the demographic characteristics that can be produced by Digital Sandbox. Using the GME we can generate defensible estimates for any of the Census and ACS demographic variables.

INDIVIDUAL VARIABLES*:

- Sex
- Age
- Marital Status
- Race
- Place of Birth
- Citizenship
- Year of Entry
- Education Level
- Language Spoken
- Disability
- Fertility
- Dependents
- Veteran Status
- Occupation
- Income

HOUSEHOLD VARIABLES*:

- Structure Age
- Structure Type
- Own vs. Rent
- Property Value
- Access to Car

**Not a complete list of variables*

V. Conclusion

Understanding the demographics and dynamics of a local population is essential to understanding one’s homeland security risks to terrorism and natural disasters. Homeland security risk analysts need detailed population measures that are based on the best government data available. All analysts need those measures at geographically granular levels that map to recognized jurisdictional boundaries, and national level studies require metrics that are universally available for the entire U.S. For many locations, using 10-year-old census data does not reflect the current population, and for most disaster scenarios, planners need to know about more than simply the resident (nighttime) populations.

Most currently available population data were not produced for a homeland security audience (e.g., Congressional boundaries, mass marketing campaigns, etc.), and do not reflect the unique challenges facing homeland security planners.

Digital Sandbox understands the unique needs of the homeland security, having been in the security and risk analysis business for over 11 years, and we have constructed a set of population data metrics for use in homeland security risk analysis and planning. Our data is based on U.S. Census Bureau and other publicly available Federal government sources; it includes residents, commuters, international visitors and

illegal immigrants separately; and it is available in combined form as daytime population, nighttime population, or as a number of hazard-specific population risk metrics (e.g., the density-weighted “Population Index” used by DHS in their grants risk formulas). In addition, we have built a powerful and flexible tool for combining our population data into custom population-based measures as needed.

For more information on these or any other Digital Sandbox products, please contact:

marketing@dsbox.com