P1.12

ESTIMATING DAYTIME AND NIGHTTIME POPULATION DISTRIBUTIONS IN U.S. CITIES FOR EMERGENCY RESPONSE ACTIVITIES

Timothy N. McPherson and Michael J. Brown
Los Alamos National Laboratory, Los Alamos, NM

1. Introduction

Effective emergency planning and response in the event of a hazardous chemical or biological material release in an urban area requires accurate definition of population exposure. Quantifying the complex plume dispersal in an urban environment is an important aspect in the accuracy of exposure assessments, but an equally important issue is the quantification of the population distribution underlying plumes. The calculation of the exposed population is problematic because population is not static. The spatial distribution of population shifts dramatically between night and day as people migrate from residences to places of work and commerce. Currently, the vast majority of available population datasets represent residential (i.e., nighttime) populations and therefore have no temporal component. For example, in the United States, the population data readily available from the U.S. Census Bureau is based on households (U.S. Census Bureau, 2000). Many dispersion modeling systems utilize the U.S. Census Bureau database when performing exposure assessments for U.S. cities. For releases occurring at night, use of the U.S. Census Bureau population database will provide relatively accurate exposure assessments, but for releases occurring during the day significant errors in exposure assessments will occur due to the daytime migration of people to workplaces, shopping centers, and educational facilities. In this research, we demonstrate a method for estimating urban daytime and nighttime population using US Census, infrastructure, and business demographic data in a GIS.

The underlying value of creating such a database will be improved exposure estimates in daytime release scenarios. The importance of accounting for the daytime distribution of urban populations is illustrated in a case study of a hypothetical material release in Albuquerque, NM. Figure 1 shows the estimated number of affected individuals following a daytime release using nighttime and daytime population distributions. For these plume dispersion simulations, all meteorological conditions and plume dispersal parameters were held constant to facilitate direct comparison of daytime and nighttime population databases. In this scenario, the material release occurred in downtown Albuquerque with moderate winds from the Northwest.

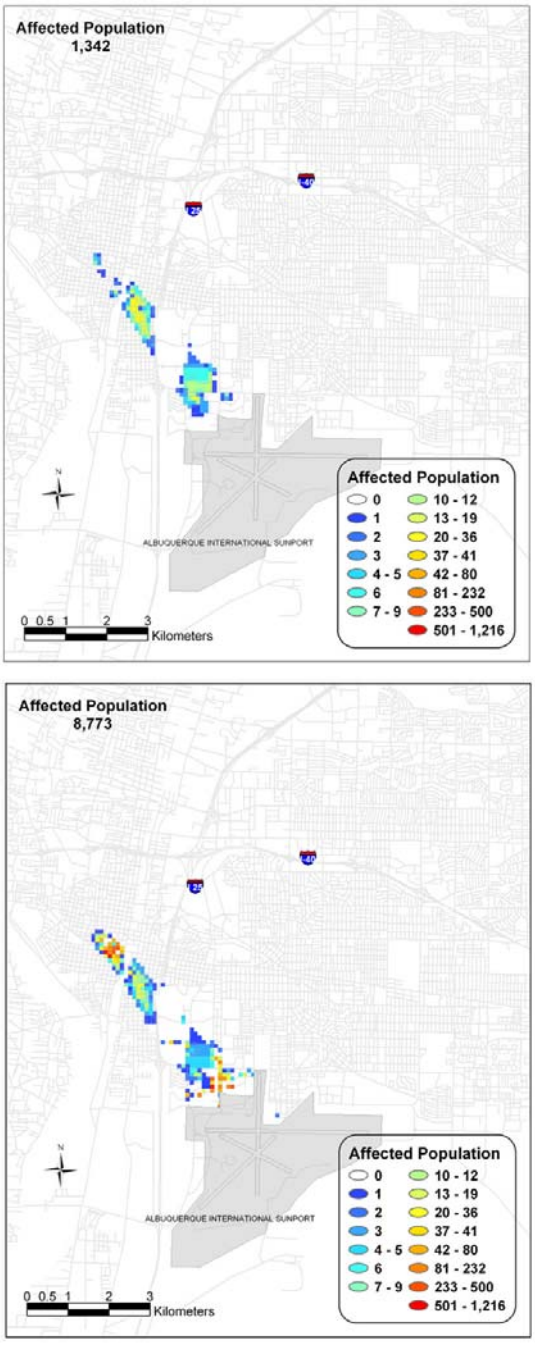

Figure 1. Simulated populations affected by a hypothetical daytime release of a hazardous material in Albuquerque, NM using both nighttime (top) and daytime (bottom) population databases.

* *Corresponding author address:* Timothy N. McPherson, LANL, D-3, MS F607, Los Alamos, NM 87545, tmac@lanl.gov.

The resulting plume traveled southeast towards the Albuquerque Airport. In this daytime release case, the number of affected individuals is a factor of 6.5 greater when using the more appropriate daytime population database as opposed to the nighttime population database. Furthermore, the spatial distribution of affected individuals is considerably different. When using a daytime population database for a daytime material release, the affected population is clustered in the downtown business district and in the commercial/industrial zone near the airport. When using a nighttime population database the affected population is spread more evenly across the plume. The differences in the magnitude and spatial distribution of the affected population could impact planning for post-release emergency response and hospital triage.

2. Prior Work

There are two primary methods for constructing population datasets: 1) Demographic counts and 2) Geographic Information Science (GIS)/remote sensing. Demographic counts, e.g., the U.S. Census, are labor and time intensive efforts to enumerate people living in predefined geographical zones. These surveys are conducted for multiple purposes, but the primary focus of data collection is housing units and the people living therein. GIS/Remote sensing techniques use statistical methodology and empirical relationships between population and topography, coastlines or human infrastructure to estimate population distributions from geographic data and remote sensing imagery (Henderson and Xia, 1997; Lo and Welch, 1977; Iisaka and Hegedus,1982; Lo, 2001;Yuan *et al*., 1997; Dobson *et al*., 2000; Harvey, 2002; Langford and Harvey, 2001). The majority of the datasets derived in these research efforts covered small areas in the United States or portions of other countries. Dobson *et al*. (2000) did create a dataset that covered the entire globe, but the 1-kilometer resolution used in that dataset is insufficient for urban exposure analyses.

Previous research has shown that diurnal shifts in population can be accounted for in a spatial database. Dobson *et al.* (2000) constructed a global population database, known as LANDSCAN, at a 30 by 30 arc-second resolution (approximately 1 km grid cells in the lower latitudes) that accounted for diurnal movements of population in order to improve emergency response activities. They distributed country or province population to grid cells using a probability coefficient based on slope, proximity to roads, land cover, nighttime lights, and an urban density factor. Population in urban areas was adjusted to account for urban density using NGDC nighttime lights and Census P-95 Circles data, but they were not able to account for business/employee demographics in their population estimation.
The research group at Oak Ridge National Laboratory that developed the LANDSCAN global population database has recently begun developing high-resolution population datasets in the United States that account for population movement. The LANDSCAN USA population distribution model is a 90m resolution population dataset containing age, race and sex attributes. This model will have a wide range of potential uses when completed. To date, a pilot study in the Houston area has been completed (LANDSCAN USA, 2003).

3. Methodology

We have created a preliminary estimate of the diurnal temporal shift in population due to employment. To accomplish this, separate population grids for nighttime residential, daytime residential, and daytime workplace population were created using existing demographic data and GIS. We constructed day/night population databases using data available for the entire U.S. from the U.S. Census Bureau, Geographic Data Technology (GDT), Navigational Technologies (NAVTECH), and the American Business Directory, Inc.

3.1 Databases

As noted above, four sources of data were used to construct the daytime and nighttime population datasets. Table 1 lists the sources and data types of the databases used in this study.

Table 1: Datasets used to construct daytime and nighttime datasets.

Database	Source	Date	Data type
StreetMap USA	ESRI/GDT	1998	Vector Polyline
NAVTECH Premium Streets Data	NAVTECH	2003	Vector Polyline
State Business Directory	American Business Directory	2000	Vector Point
Census Blockgroups	US Census Bureau	2000	Vector polygon
Census County to County Journey to Work	US Census Bureau	2000	ASCII Text Table

Two general types of data were used: 1) census enumeration data and 2) physiographic data. The census enumeration data provided the population count to be disaggregated and the physiographic data provided the spatial units to which the census enumeration data were disaggregated. In this research, the 2000 U.S. Census data in vector format were used as the starting point from which populations were shifted from residence to workplace. In the decennial census, the U.S. Census defines the

residential population of the United States at varying levels of geographic precision. The census blockgroup data, the second highest resolution dataset from the census, were used in this research. Census blockgroups are clusters of census blocks containing from 600 – 3,000 people. The blockgroup data define the urban nighttime population per polygon. GeoData Technologies (GDT) road network data were used to create a grid of location coefficients that were used to spatially disaggregate the census blockgroup.

Workers from the residential population were then routed to businesses within their county of work using the U.S. Census "County to County Journey to Work" data coupled with the American Business Directory, Inc. data. The American Business Directory creates a commercial and industrial business database for the United States called the State Business Directory (SBD). The SBD is a commercial database containing information on approximately 12,000,000 businesses in the United States. The SBD is constructed from yellow pages directories, SEC annual reports, local, state and federal government data and verified by extensive telephone research. Each record in the database contains the company name, address, geolocation (latitude-longitude), type of business and a range of the number of employees. Business types in the database are defined by 6 digit Standard Industrial Classification Codes (SIC). The SBD data were used to create a set of location coefficients that were used to spatially disaggregate the county worker data. The U.S. Census "County to County Journey to Work" data within the U.S. Census Transportation Planning Package defines the number of people that migrate from one county (i.e., their county of residence) to another county for work. Figure 2 shows an example of the county to county worker migration data in Utah.

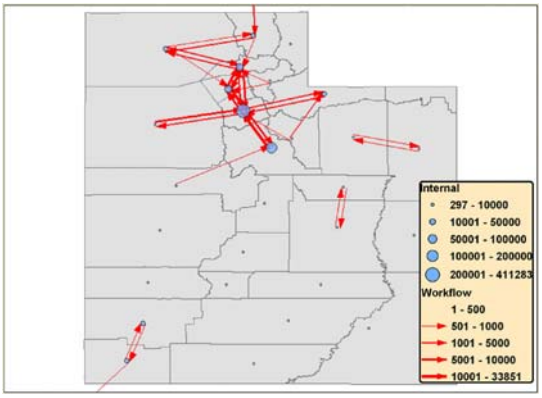

Figure 2. County to county worker flows in Utah. The graduated blue points represent the population working within their county of residence. The graduated red arrows represent the worker flow from county to county. Although county workflow below 500 people is excluded in this figure, all worker flows were used in the daytime population model development.

The county data were used because they are the best available data on worker counts that cover the entire United States. The U.S. Census Transportation Planning Package does have a worker flow dataset with a higher spatial resolution based on census tracts, but those data are not yet available from the 2000 Census.

3.2 Nighttime Residential Population

Nighttime residential population databases for the United States are freely available from the U.S. Census Bureau. Although these vector data represent one of the most complete population databases available, a few modifications were required to facilitate use of the data for exposure assessments. The use of the vector data in exposure assessments typically requires the analyst to assume an areal average of population across the entire polygon. In our methodology, population is assumed to be evenly distributed on road areal coverage (i.e., evenly distributed based on the population per road grid cell within a census blockgroup). In this manner, we are eliminating areas where the probability of housing is low according to road density. Distributing the population evenly across roads will still produce spatial errors, but less so than an areal average across the census polygon. Figure 3 shows the difference between datasets in which the population is distributed evenly across the polygon versus distributed to roads.

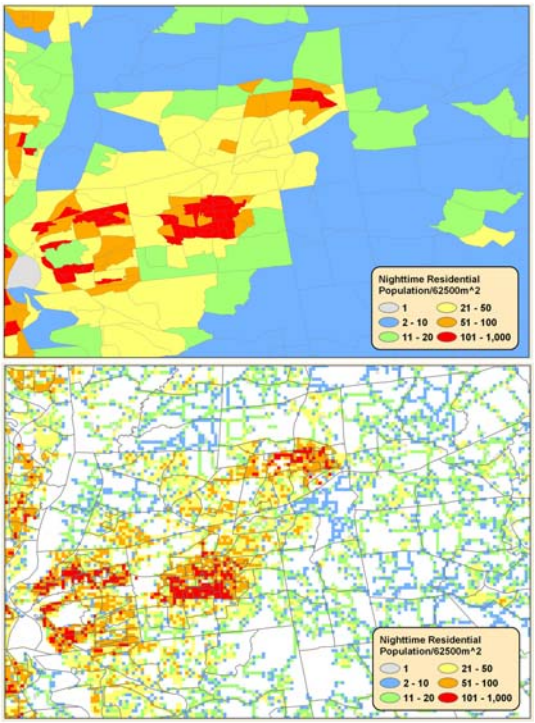

Figure 3. Comparison of areal average population (top) versus population attributed to roads (bottom).

In this research, the nighttime residential population grid was constructed using U.S. Census 2000 shortform population data coupled with the GDT street data. The vector based road data were converted to raster and then used to create a grid of location coefficients, which were used to disaggregate the census polygon data. Figure 3 shows an example of the residential population raster construction. The process begins with a vector map of 2000 Census blockgroups containing population counts. The GDT road data is then queried to define residential roads, i.e., those roads having a CFCC classification of A4*. Those residential road vectors are then converted to raster at a resolution of 25m, and the road cells in each census polygon are counted. The residential population location coefficient is then calculated according to equation 1:

$$p_{ri} = 1/r_{bg} \quad (1)$$

where p_{ri} is the residential location coefficient for each grid cell i, and r_{bg} is the number of road grid cells per census blockgroup. The residential location coefficient represents the proportion of the total residential population in a blockgroup that may live in a given grid cell within that blockgroup. The raster map of location coefficients created using this method was used to disaggregate the total residential population in the blockgroup into each grid cell according to equation 2:

$$NRP_i = W_{bg} \cdot p_{ri} \quad (2)$$

where NRP_i is nighttime residential population in each 25m grid cell i, p_{ri} is the nighttime residential location coefficient for each grid cell i, and W_{bg} is the residential population in each blockgroup. Following that step the 25m raster of nighttime residential population is aggregated to 250m for further calculations.

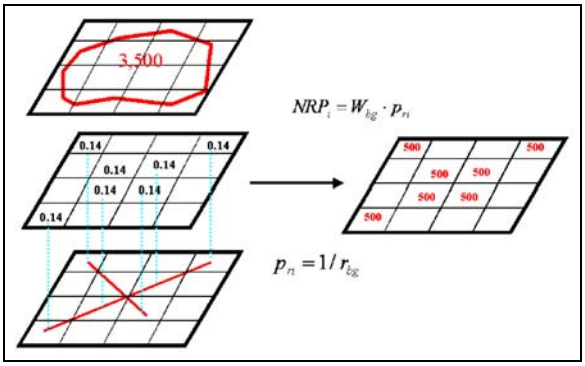

Figure 4. Example of nighttime population calculation. The roads (bottom left) are converted to raster and used to derive a location coefficient grid (middle left), which is subsequently used to spatially disaggregate the census count (top left) thereby producing the nighttime raster model (right).

3.3 Daytime Population

The daytime population model is composed of two components as shown in Figure 4. The first component is an estimate of the daytime worker population density. The second component is a raster of residential population that remains home during the day. In this case, we are assuming that non-working individuals remain in their residences during the day.

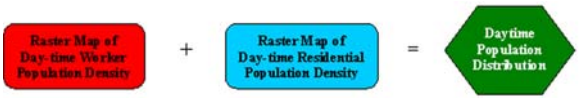

Figure 4: Derivation of the daytime population dataset.

The daytime worker population grid was constructed from two datasets: the 1999 State Business Directory and the 2000 U.S. Census "County to County Journey to Work" database. The daytime residential population was derived from our nighttime residential model and the Census "County to County Journey to Work" database. The SBD data was used to spatially disaggregate the number of workers in each county as defined by the Census Journey to Work data. The SBD includes three data fields that are useful in the placement of workers at facilities: 1) facility street address, 2) latitude-longitude, and 3) range of employees. The former two data types provide information on geolocation, while the latter provides information on the magnitude of the employment. Facilities in the SBD table for each state were geographically placed using either the ArcGIS geocoding tool with the Navtech Premium Street Data providing the street index or the latitude and longitude contained in the SBD for each record. Geocoding using the street address was the primary method for specifying facility location because there were no metadata available on the accuracy of the latitude and longitudes imbedded in the SBD. The accuracy of the placement of facilities was checked to assure each facility was placed in its correct 5-digit zip code area. Facilities that could not be placed due to poor address information and those that were placed in the incorrect zip code area were geographically placed again using the latitude and longitude imbedded in the SBD. These facilities were also checked against their zip code area, and those facilities that were placed in the wrong zip code were removed from the analysis. In this way, the geographic placement of facilities in the final vector file of businesses was accurate to at least the zip code area. A further improvement on the methodology would be to assess the spatial accuracy of the geolocations by using GPS to define the true geolocations of facilities. This task was not undertaken due to the expected expense of such an effort.

As noted earlier, the SBD data only provides a range of employees, while the Census Journey to Work database contains employee numbers but only at the county level. We used the higher spatial resolution SBD data to identify the locations of workers, and without any other information, used the midpoint of the employee range for the worker count. We then used the Census "County to County Journey to Work" database to scale the worker counts such that the total number of workers per county agrees with the Census database. Figure 5 shows an example of the steps used to create the daytime worker population raster.

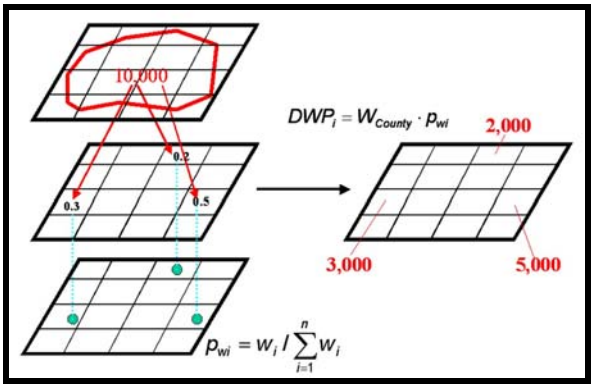

Figure 5. Example of daytime workplace population calculation. The businesses (bottom left) and their estimated workforce size were used to derive a location coefficient grid (middle left). The location coefficient grid was then used to spatially disaggregate the worker count derived for each county from the Census County to County Journey to Work data (top left) thereby producing the daytime workplace raster model (right).

Within the GIS, the SBD employee data were used to compute a grid of urban worker location coefficients that define the probability that an employee in the county works in a given grid cell. The worker location grid was calculated using equation 3:

$$p_{wi} = w_i \Big/ \sum_{i=1}^{n} w_i \qquad (3)$$

where p_{wi} is the worker location coefficient for each grid cell i, and w_i is the estimated number of workers in each grid cell i in each county based on the SBD midpoint of the employee range. The worker location coefficient represents the proportion of the total workers in a county that may work in a given grid cell within that county. The raster map of location coefficients created using this method was used to disaggregate the census defined total number of workers in the county into each grid cell according to equation 4:

$$DWP_i = W_{County} \cdot p_{wi} \qquad (4)$$

where DWP_i is workplace population in each 250m grid cell i, p_{wi} is the worker location coefficient for each grid cell i, and W_{County} is the total number of workers in the county.

The second step in the calculation of the daytime population is to define the daytime residential population distribution following the daily migration to workplaces. Figure 5 displays the calculation of the daytime residential population raster. The daytime residential population was computed using the nighttime population grid and the ratio of non-working population to total population in the county as defined in equation 5:

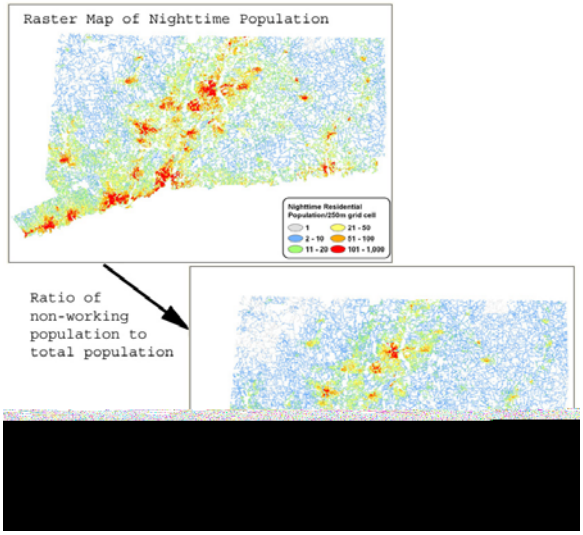

Figure 5. Derivation of the daytime residential population raster.

$$DRP_i = 1 - \left(W_{home} \Big/ P_{2000} \right) \qquad (5)$$

where DRP_i is daytime residential population in each 250m grid cell i, W_{home} is the number of workers residing in the county and P_{2000} is the total population in the county. In this approach, we are assuming the distribution of workers relative to non-workers is constant across the county. This assumption may also produce spatial errors given the large size of counties. A future improvement will be to use a smaller geographic unit in the definition of the non-worker to total population ratio. The resulting raster maps of daytime working population and daytime residential population were added to define the total daytime population for the state.

4. Results

A day and night population database has been derived for the entire continental United States plus Hawaii at 250 m grid resolution. Figure 6 displays the day and night population distribution for the continental U.S.A. at 1 km resolution.

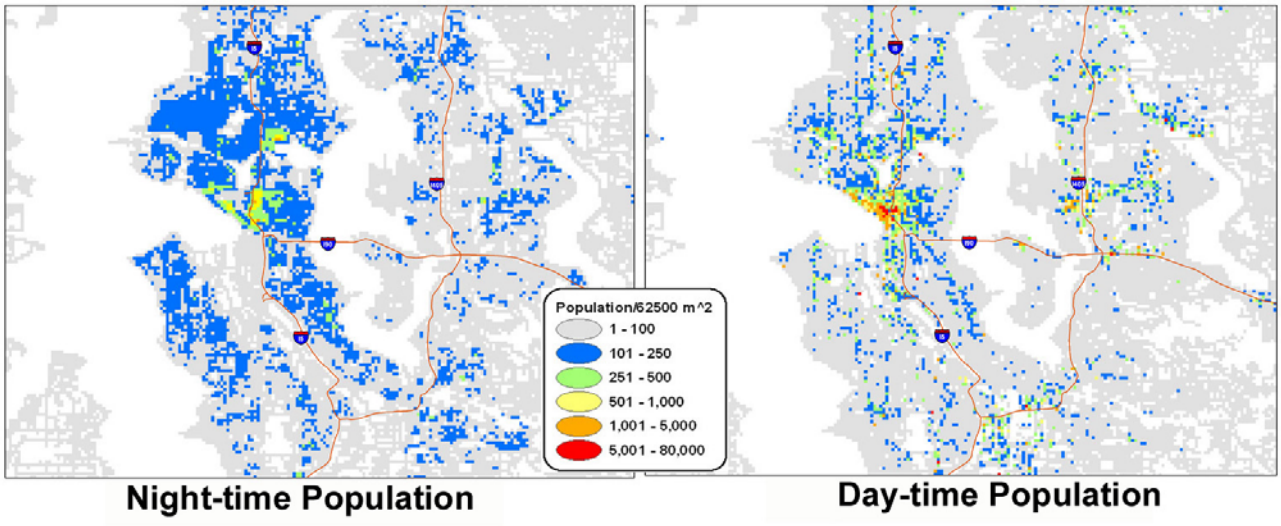

Figure 9. Nighttime population versus daytime population in Seattle, WA.

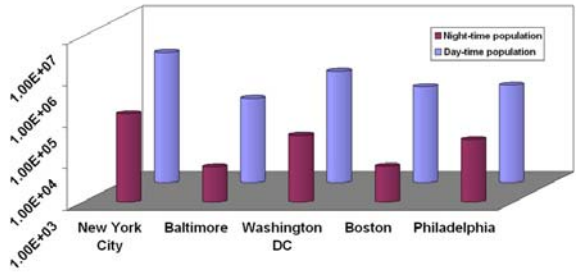

Figure 10. Daytime versus nighttime Population in downtown census tracts.

4.1 Evaluation

Model evaluation is an important aspect in any modeling project. Sufficient "ground-truth" data is needed in order to assess the accuracy of the derived database. Conducting a field campaign to acquire population count per grid cell from a statistically significant sample would be an appropriate validation technique, but such an effort is extremely expensive and beyond the scope of this project. Though not entirely satisfactory, as an alternative we have used county data for the evaluation of our derived nighttime and daytime population raster models. This will not evaluate the accuracy of individual population grid cells, but will determine if errors were introduced into the computations on a somewhat coarser spatial scale.

In each case, the nighttime residential and daytime workplace raster models were aggregated to the county level and compared to the U.S. Census County residential and worker count data. Error statistics were generated for each county. Since these comparisons do not utilize field-collected data, they are not a complete validation but are more an indicator of general model performance.

The model error analysis indicates the models are performing extremely well for workplace and nighttime populations at the county level. Figures 11 and 12 display log-log plots of the counted census populations versus the estimated populations at workplaces and residences, respectively. In each case, the models accurately capture the population counts for nighttime residential and workplace with correlation coefficients of 0.99. The nighttime population model does appear to under predict the population in counties with populations lower than 10,000 (Fig. 12). However, the total population of the U.S. estimated in the nighttime population model is only 1.5% less than the actual population as counted by the U.S. Census, indicating the aggregate effect of the errors in low population counties is low. In the nighttime population model, the median relative error for the counties is low at 1.6. The median relative error for the workplace population model is also low at 0.24. The aggregate effect of errors in the workplace population model is minimal. The total workplace population for the U.S. estimated in the workplace raster model is only 0.1% less than the U.S. total counted in the census.

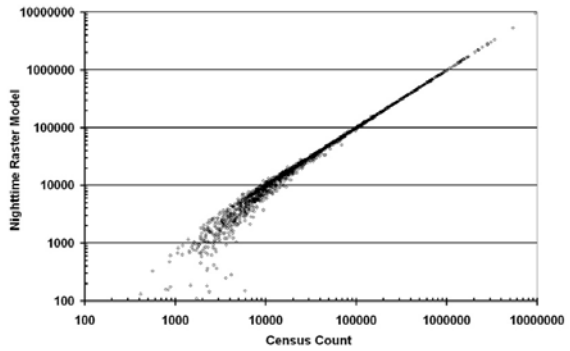

Figure 11. Comparison of nighttime census population per county ("truth") to the LANL-derived nighttime raster model aggregated to the county level.

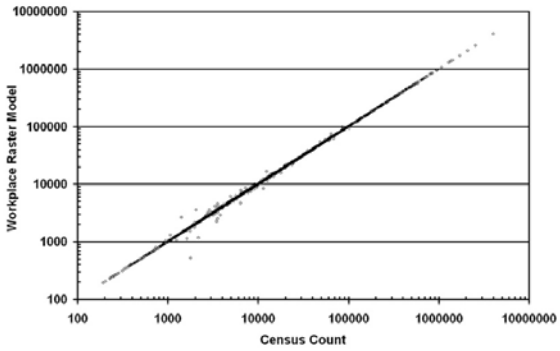

Figure 12. Comparison of U.S. Census county worker counts (truth) to the LANL-derived workplace raster model aggregated to the county level.

5. Progress to Date/Future Efforts

As shown in Figure 6, daytime and nighttime populations have been constructed for all of the continental United States and Hawaii. This area covers 99.8% of the U.S. population. Although daytime and nighttime population datasets have been constructed for all states, these results are preliminary. The results generated using the method documented within this report are representative of maximum daytime workplace population and maximum nighttime residential population. Although these peaks represent an improved estimate of the temporal distribution of urban populations over standard nighttime population databases, there are many potential improvements that could further increase the value of the dataset. For example, greater specificity of both the temporal and spatial distribution of population will allow for better risk assessments in hazardous material release scenarios. Potential improvements to the dataset include, but are not limited to, the following:

1. Determination of the spatial and temporal distribution of populations commuting between work and home, i.e., the traffic component of the population.
2. Determination of the spatial and temporal distribution of populations at educational facilities.
3. Determination of the number of people within buildings, i.e., the indoor and outdoor components of the population distribution.
4. Determination of the spatial and temporal distribution of populations in retail zones.

A further potential improvement would be the use of more advanced numerical techniques in the disaggregation of the population numbers such as maximum likelihood estimation as discussed in Langford and Harvey (2001).

6. Summary

In order to address the potential shortcomings in exposure assessments of hazardous material releases, two population datasets were constructed to more accurately represent the temporal and spatial distributions of populations under hazardous material plumes. Exposed populations are difficult to define because population distributions shift throughout the day according to the work, shopping and mobility habits of urban citizens. Most available population datasets were constructed for purposes other than exposure assessments and often have no temporal component and a limited spatial component. These population datasets are based typically on residential units such as households or families and are often geographically constructed based on sampling design instead of spatial accuracy. Residence-based population datasets may be useful in exposure assessments if analysts can assume the majority of the population is in their homes. This assumption may be valid at night, but during the day it may be significantly inaccurate. Therefore, the use of these population data in exposure assessments may misrepresent the actual population exposed to the material in question.

In this research, we created two databases that account for daily population migrations. Raster based models of nighttime and daytime population were constructed in a GIS with a 250-meter resolution. These datasets cover the continental United States and Hawaii and represent 99.8% of the U.S. population. The raster format was used to facilitate the derivation of the spatial models and to improve the ease of use of the datasets in exposure assessments. The majority of urban dispersion models are grid based and the construction of the population models as grids may hasten their use in those codes.

Error analyses of the model results indicate the models are performing extremely well for workplace and nighttime populations at the county level. In each model, the correlation coefficients between the model and census data was 0.99 indicating the models are accurately capturing the population counts for

nighttime residential and workplace. The nighttime population model does under predict the population in multiple counties with populations lower than 10,000, but the error in the total population estimate for the United States is only 1.5%.

The value of the models in exposure assessments is significant. Considerable differences were found between the derived daytime and nighttime population datasets. Daytime populations in downtown census tracts were found to be 6.9 – 28.6 times greater than the nighttime populations in the same census tracts. These differences could have profound impacts on exposure assessments and emergency management decisions during hazardous material release events.

Although the methodology used in this technique provided national datasets on daytime and nighttime population, uncertainty in the source data warrant significant consideration before applying these data in a wide range of applications. First, the methodology is highly reliant on the Census data, such that errors in that dataset (particularly in the population or worker counts) will be replicated and potentially compounded in these datasets. Secondly, attribution of the population to roads is dependent on the quality of the road data in terms of geolocation and attribution. If an existing road is not listed in the streets data, is incorrectly classified as non-residential, or it is placed in the wrong location, then residential populations will be placed in the wrong locations in the raster model. Thirdly, geolocations of businesses in the SBD were defined using two sources: 1) results from the ArcMap geocoding algorithm and 2) the imbedded geolocation in the SBD data. In both cases, the business geolocations were not validated due to the expense of such an effort. Potential geolocation errors in the workplace population may shift working population to the wrong location. To minimize the impact of the workplace geolocation uncertainty, all geolocations were checked to assure facilities were placed in the correct zip code zone. A further potential error in the workplace population model is the number of employees per business. These data were derived from the employee range information in the SBD data. These data were also not validated and errors in the workplace population may be possible due to the inaccuracies in the employee ranges.

7. References

LANDSCAN. 2003. http://web.ornl.gov/sci/gist//landscan/landscan_usa.htm

Dobson, J.E., E.A. Bright, P.R. Coleman, R.C. Durfee, and B.A. Worley. 2000. LANDSCAN: A global population database for estimating populations at risk. Photogrammetric Engineering and Remote Sensing 66, 7: 849-857.

Harvey, J.T. 2002. Population Estimation Models Based on Individual TM Pixels. Photogrammetric Engineering and Remote Sensing, 68, 11: 1181-1192.

Henderson, F.M. and Z. Xia. 1997. SAR Applications in Human Settlement Detection, Population Estimation and Urban Land Use Pattern Analysis: A Status Report. IEEE Transactions on Geosciences and Remote Sensing, 35, 1: 79-85.

Iisaka, J. and E. Hegedus. 1982. Population Estimation from LANDSAT Imagery. Remote Sensing of the Environment 12: 259-272.

Langford M.L. and J.T. Harvey. 2001. The Use of Remotely Sensed Data for Spatial Disaggregation of Published Census Population Counts. IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, 260-264.

Lo, C.P. and R. Welch. 1977. Chinese Urban Population Estimates. Annals of the Association of American Geographers, 67, 2: 246-253.

Lo, C.P. 2001. Modeling the Population of Chine using DMSP Operational Linescan System Nighttime Data. Photogrammetric Engineering and Remote Sensing, 67, 9: 1037-1047.

Yuan Y., R.M. Smith, and W.F. Limp. 1997. Remodeling Census Population with Spatial Information from LANDSAT TM Imagery. Computers, Environment and Urban Systems, 21, 3/4: 245-258.