

**SEER Special Project #08**  
**Development of High Resolution Population Distribution Data to Enhance Cancer**  
**Prevention and Control Research**  
**RFP No. NCI-PC-25014-20**

**Report**

**Budhendra Bhaduri, Edward Bright, Phillip Coleman**  
Geographic Information Science & Technology  
Oak Ridge National Laboratory  
Oak Ridge, TN 37831-6017

## I. BACKGROUND

Population data are one of the critical elements in cancer prevention and control research. For the US, the source for population data is the US Census Bureau, which reports population counts by census blocks (smallest polygonal unit), block groups (aggregated blocks), and tracts (aggregated block groups). At the highest resolution (block level), a uniform population distribution is assumed and the population values are typically an attribute of the block (polygon) centroids. Similarly, population values for block groups and tracts are reported at the centroids of the block group and tract polygons.

In geospatial analyses, these points are used to represent the population of a census polygon. For example, calculation of travel time to health care providers considers these centroids as the starting points for travel. For exposure and risk analyses, these centroids often serve as "receptor" points for calculating exposure or dosage from any dispersed agent.

In common practice, census data are intersected with buffers of influence (such as those from emission sources) using two primary approaches to quantify population at risk:

- a. tally the entire population (if the centroid is inside the buffer) or zero population (if the centroid is outside the buffer)
- b. an area weighted population accounting approach (based on the ratio of the areas of the polygon included in and excluded from the buffer).

However, it is well understood that uniform population distribution is the weakest assumption and all analytical approaches that use census polygon centroids to represent population are likely to produce results containing substantial but unknown errors.

These limitations, to a large degree, can be overcome by developing population data with a finer geographic resolution than the level of the census block. Geodemographic data at such scales will represent a more realistic non-uniform distribution of population. Such data can then be used either to directly estimate travel times or the exposures of population groups or to make more accurate population centroids for the population data of census blocks.

Oak Ridge National Laboratory (ORNL), as part of their LandScan global population project, has developed a high-resolution (30 arc seconds or approximately 1 km x 1 km cell size) population distribution model (LandScan) for the entire world (Dobson et al., 2000). Collaborating with the National Center for Environmental Assessment (NCEA) of the United States Environmental Protection Agency (USEPA), ORNL has recently demonstrated the feasibility of developing a very high-resolution (3 arc seconds or approximately 90m x 90m cell size) population distribution data (LandScan USA) for the US. Detailed population distribution data including nighttime (residential) as well as daytime distributions at this resolution have been developed for 29 counties covering southeast Texas and western Louisiana. Such data can then be used either to directly estimate travel times or the exposures of population groups or to make more accurate population centroids for the population data of census blocks.

## II. OBJECTIVE

The primary objectives of this project were:

- To develop nighttime (residential) population distribution for 99 SEER counties in Iowa at 90 meter spatial resolution based on the 2000 Census using geographic information system methods.
- Based on the 1990 Census data, develop equivalent population distribution at 90 meter resolution for five Iowa counties (which have experienced the highest population growth between 1990 and 2000).
- To develop estimates of age groups and gender for the 90 meter cells.
- To derive population centroids for various Census polygonal units (blocks, block groups, and tracts) for general population as well as different demographic categories including sex (male and female) and age categories based on the LandScan USA data.

## III. METHODOLOGY

The global LandScan population distribution model involves collection of the best available census counts (usually at sub-province level) for each country and four primary geospatial input datasets, namely land cover, roads, slope, and nighttime lights, that are key indicators of population distribution. Relationships between any of these datasets and population distribution are not globally uniform. Roads play a critical role in human settlements independent of other forms of transport. However, residential population density in proximity to major roads varies significantly across the world. For example, residential population is much higher to primary roads in Southeast Asian countries, which is quite contrary to the same in the United States. Similarly, preference for people to reside on steeper versus gentler (or flat) slopes is generally a function of the abundance of available areas with gentler slope.

Based on this variability in cultural and settlement geography, the world is divided into several different regions and each region is considered to have unique settlement characteristics. For each region, the population distribution model calculates a “likelihood” coefficient for each LandScan cell, and applies the coefficients to the census counts, which are employed as control totals for appropriate areas. For LandScan USA, census blocks serve as the polygonal unit control population. Census blocks are divided into finer grid cells (90m) and the total population for that block is then allocated to each cell weighted by the calculated likelihood (population coefficient) of being populated. For sex-age groups, census blocks are the control populations to which the sum of populations of the 90 meter by 90 meter square areas are constrained to meet. GIS is essential for integrating diverse input variables, computation of probability coefficients, allocation of population to cells, and reconciliation of cell totals with aggregate control totals. Remote sensing is an essential source of two input variables-land cover and nighttime lights-and one ancillary database-high-resolution panchromatic and multispectral imagery-used in verification and validation of the population model and resulting LandScan USA database. Large volumes of satellite derived spatial data

including land cover and nighttime lights are used in developing LandScan databases and verification and validation (V&V) of the population model.

## 1. Description of Input Data Sets for the LandScan USA Model:

### 1.1. Population Counts for Administrative Units

Population counts at the census block level were obtained from the US Census 1990 and 2000 Data Products, specifically the Summary File 1 (SF 1). The SF 1 reports demographic characteristics by race, age groups, and sex at the block level along with data on housing units at the same geographic level. An “advanced national” version of the SF 1 file was released on November 16, 2001 that, according to the US Census Bureau, followed by the “final national” release in June of 2002.

### 1.2. Roads

The Topologically Integrated Geographic Encoding and Referencing (TIGER) files, developed by the U. S. Census Bureau in conjunction with the U. S. Geological Survey (USGS) have become the standard geometric reference not only for census data, but for digital street atlases and other high-precision applications as well. The TIGER/Line files are a digital database of geographic features, such as roads, railroads, rivers, lakes, political boundaries, census statistical boundaries, etc. covering the entire United States. The data base contains information about these features such as their location in latitude and longitude, the name, the type of feature, address ranges for most streets, the geographic relationship to other features, and other related information. They are the public product created from the Census Bureau's TIGER (Topologically Integrated Geographic Encoding and Referencing) data base of geographic information. TIGER was developed at the Census Bureau to support the mapping and related geographic activities required by the decennial census and sample survey programs.

### 1.3. National Land Cover Data

This land cover data set was produced by the U.S Geological Survey (USGS) as part of a cooperative project between the USGS and the U.S. Environmental Protection Agency (USEPA) to produce a consistent, land cover data layer for the conterminous U.S. based on 30-meter Landsat Thematic Mapper (TM) data. National Land Cover Data (NLCD) was developed from TM data acquired by the Multi-resolution Land Characterization (MRLC) Consortium. The MRLC Consortium is a partnership of federal agencies that produce or use land cover data. Partners include the USGS (National Mapping, Biological Resources, and Water Resources Divisions), USEPA, the U.S. Forest Service, and the National Oceanic and Atmospheric Administration. A 21-class land cover classification scheme applied consistently over the United States. The spatial

resolution of the data is 30 meters and mapped in the Albers Conic Equal Area projection, NAD 83. The NLCD are provided on a state-by-state basis. The state data sets were cut out from larger "regional" data sets that are mosaics of Landsat TM scenes. At this time, all of the NLCD state files are available for free download as 8-bit binary files and some states are also available on CD-ROM as a Geo-TIFF.

The TM multi-band mosaics were processed using an unsupervised clustering algorithm. Both leaves-off and leaves-on data sets were analyzed. The resulting clusters were then labeled using aerial photography and ground observations. Clusters that represented more than one land cover category were also identified and, using various ancillary data sets, models developed to split the confused clusters into the correct land cover categories. The land cover classification statistics for the state of Iowa is as follows:

<b>Land Cover Classes - Iowa</b>	<b>Square Miles</b>
11 Water	508
12 Perennial Ice Snow	0
21 Low Intensity Residential	423
22 Hi Intensity Residential	160
23 Commercial/Industrial/Transportation	809
31 Bare Rock	7
32 Quarries/ Mines	24
33 Transitional	0
41 Deciduous Forest	4173
42 Evergreen Forest	8
43 Mixed Forest	72
51 Shrubland	0
61 Orchards/ Vineyard	0
71 Grasslands/Herbaceous	2882
81 Pasture/Hay	9217
82 Row Crops	36415
83 Small Grains	341
84 Fallow	0
85 Urban/Recreational Grasses	161
91 Woody Wetlands	695
92 Emergent/Herbaceous Wetlands	377
State/Region Total	56271

The classification system used for NLCD is modified from the Anderson land-use and land-cover classification system. Many of the Anderson classes, especially the Level III classes, are best derived using aerial photography. It is not appropriate to attempt to derive some of these classes using Landsat TM data due to issues of spatial resolution and interpretability of data. Thus, no attempt was made to derive classes that were extremely difficult or “impractical” to obtain using Landsat TM data, such as the Level III urban classes. In addition, some Anderson Level II classes were consolidated into a single NLCD class.

#### 1.4. Slope from Digital Terrain Data (DTED) and National Elevation Data

LandScan currently uses the National Imagery and Mapping Agency’s (NIMA) Digital Terrain Elevation Data (DTED) Level 1 for global coverage of roads and slope respectively. DTED-Level 1 (90 m or 3 arc second resolution) data is used for slope. For LandScan USA, the National Elevation Data (NED) is also used to derive higher resolution slope data. The NED has a resolution of 1 arc-second (approximately 30 meters) for the conterminous United States, Hawaii, and Puerto Rico and a resolution of 2 arc-seconds for Alaska.

NED data sources have a variety of elevation units, horizontal datums, and map projections. In the NED assembly process, the elevation values are converted to decimal meters as a consistent unit of measure, North American Datum 1983 is consistently used as horizontal datum, and all the data are recast in a geographic projection. Older digital elevation models produced by methods that are now obsolete have been filtered during the NED assembly process to minimize artifacts that are commonly found in data produced by these methods. Artifact removal greatly improves the quality of the slope, shaded-relief, and synthetic drainage information that can be derived from the elevation data.

#### 1.5. Nighttime Lights

Nighttime lights are an invaluable global resource available from the National Geophysical Data Center (NGDC). Chris Elvidge of the Desert Research Institute, in residence at NGDC, produced the original frequency data (Elvidge, Baugh, Kihn, Kroehl and Davis 1997) and the new Stable Lights and Radiance Calibrated Lights of the World.

The frequency data available for LandScan1998 were good for moderate settlement densities but tended to miss the diffused lights of small settlements and to saturate in large cities (Sutton 1997; Sutton, Roberts, Elvidge and Meij 1997). Radiance-calibrated images are derived from the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) Nighttime Imagery from 1996 and 1997 at 30 arc second resolution. Their main advantage over the previously available frequency data is three separate gain settings to capture low-intensity lights in the countryside, saturated lights in cities, and everything in between. The new data available for LandScan2000 vastly improved our ability to detect small towns and villages and to discriminate densities within large cities.

The nighttime lights data resolution is sub-critical for enhancing the population distribution model at 3 arc second resolution for LandScan USA. It is usually used to delineate urban areas from suburban or rural areas and LandScan USA cells are not weighed with respect to the lights database because of the resolution discrepancy.

#### 1.6. Landmark Polygons

Landmark polygons, developed and distributed as part of the Census TIGER data, denote places of specialized land use including parks, cemeteries, golf courses, other recreational areas. These areas are usually utilized to demark exclusion areas for residential land use.

#### 1.7. Schools and Prisons

Two primary data bases are used for schools and prisons, the locations of which are critical for estimating residential population. In the land cover data, schools are usually captured as impervious structures in residential neighborhoods. However, they serve the purpose of exclusion zones (i.e. zero residential population). The Public Elementary/Secondary School Universe Survey for 2000-2001 (from Center for Education Statistics) is used along with another school database available with ESRI GIS data. Prisons typically indicate anomalously high population density and locations of prisons are critical to confine the prison population to the respective blocks. The National Jail Census, 1999 database (developed by Inter-university Consortium for Political and Social Research and available from the US Department of Justice) is used to delineate the prison locations and prison population.

An extensive effort was made to geolocate the schools and prisons using orthophotographs. A large number of the schools and prisons are geocoded to the zip code centroids and aerial images were used to refine their positional accuracy before they were used in the model.

### 2. Algorithm for Population Distribution

Based on this variability in cultural and settlement geography, the area under investigation is divided into several different regions and each region is considered to have unique settlement characteristics. For each region, the population distribution model calculates a “likelihood” coefficient for each LandScan cell, and applies the coefficients to the census counts, which are employed as control totals for appropriate areas. For LandScan USA, census blocks serve as the polygonal unit control population.

Census blocks are divided into finer grid cells (90m) and each cell is evaluated based on the primary input data layers influencing the likeliness of that cell for being populated. Relative weights are empirically assigned to each cell for a number of data layers and all weights assigned from different data layers are

combined to develop a cumulative weight for each cell (in a i,j matrix of cells) as follows:

$$W_{Cell\ i,j} = LC_{i,j} \times PR_{i,j} \times S_{i,j} \times LM_{i,j}$$

Where,

LC = Weight for Land Cover

PR = Weight for proximity to roads

S = Weight for slope factor

LM = Weight for land mark polygon feature

Once the individual cumulative cell weights are derived, these are combined and weighed with respect to the total population of the block to develop a block level population (or likelihood) coefficient as follows:

$$PC_{Block} = \frac{Total\ Population_{Block}}{\sum_1^n W_{Cell\ i,j}}$$

Where,

PC = Population Coefficient

N = Number of LandScan cells describing the block

Subsequently, the total population for that block is then allocated to each cell weighted by the calculated likelihood (population coefficient) of being populated as shown below:

$$Population_{Cell\ i,j} = PC_{Block} \times W_{Cell\ i,j}$$

For sex-age groups, census blocks are the control populations to which the sum of populations of the 90 meter by 90 meter square areas are constrained to meet.

GIS is essential for integrating diverse input variables, computation of probability coefficients, allocation of population to cells, and reconciliation of cell totals with aggregate control totals. Remote sensing is an essential source of two input variables-land cover and nighttime lights-and one ancillary database-high-resolution panchromatic and multispectral imagery-used in verification and validation of the population model and resulting LandScan USA database. Large volumes of satellite derived spatial data including land cover and nighttime lights are used in developing LandScan databases and qualitative verification and validation (V&V) of the population model and resulting output data.

### 3. Algorithm for Population Centroids Derivation

The population weighted centroids for Census accounting units were derived by calculating the x and y coordinates of the centroids as follows:

$$X_{Centroid} = \frac{\sum_1^n (X_{i,j} \times Population_{i,j})}{\sum_1^n Population_{i,j}} \quad \text{and} \quad Y_{Centroid} = \frac{\sum_1^n (Y_{i,j} \times Population_{i,j})}{\sum_1^n Population_{i,j}}$$

Where,

X = X coordinate of centroid (latitude)

Y = Y coordinate of centroid (longitude)

n = Total number of cells in the census accounting unit

#### IV. RESULTS

The 99 Iowa counties were considered as an individual region because geographic variability of the input data layers including roads, land cover, and slope were not significant enough. However, the population distribution model was refined to evaluate and characterize each block with respect to the input data layers.

LandScan USA population distribution data were developed for the Iowa counties at 3 arc-second (approximately 90m) spatial resolution. The data is in Geographic projection (latitude-longitude) with NAD 83 datum and represents residential or nighttime population. In addition, LandScan USA based population centroids for various Census accounting units were also derived. The following is a list of all data products that were developed:

Based on 2000 Census:

- Nighttime (residential) population distribution for 99 SEER counties in Iowa at 90 meter spatial resolution.
- Population estimates by age groups and gender for the 90 meter cells.
- General population centroids for census blocks, block groups, tracts, and zip code polygons.
- Sex based (male and female) population centroids for block groups and tracts.
- Age based (various age groups) population centroids for block groups and tracts.

Based on 1990 Census:

For a five county (Dallas, Jasper, Johnson, Linn, and Polk) area in Iowa in which substantial population change has occurred, the equivalent 1990 population distribution was computed by taking the 1990 Census Block population values and assuming the same relative distribution within the block area as for 2000. For the rest of the state of Iowa, the grid area population values for each county will be proportionally allocated so that they sum to the 1990 County population totals. It is important to consider that the Census accounting units, roads, and demographic data from 1990 Census were used to develop the population distribution for 1990 while the land cover and slope data were the same for 1990 and 2000. Because of the lack of historical and consistent land cover data (assuming slope does not vary drastically over a 10 year period) for two

different time periods, the results of the analysis do not represent a true difference in population distribution between 1990 and 2000. This fallacy can be removed by using different but representative land cover data for the two time periods (1990 and 2000) or by refining the existing land cover data using other spatial information (imagery) to adjust for the temporal change.

A detail description of the data base is provided in the appendix.

## V. REFERENCES:

- 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 & Remote Sensing* 66(7), 849-857.
- ELVIDGE, C. D., K. E. BAUGH, E. A. KIHN, H. W. KROEHL, AND E. R. DAVIS, 1997. Mapping city lights with nighttime data from the DMSP Operational Linescan System. *Photogrammetric Engineering & Remote Sensing* 63(6) 727-734.
- SUTTON, P., D. ROBERTS, C. ELVIDGE, AND H. MEIJ, 1997. A comparison of nighttime satellite imagery and population density for the continental United States. *Photogrammetric Engineering & Remote Sensing* 63(11) 1303-1313.

## Appendix I

### LandScan USA Population Distribution Data and Population Centroids for Iowa Counties Based on 2000 and 1990 Census

#### **Description:**

A CDROM was produced that contains the LandScan USA nighttime (residential) population distribution data and population centroids for counties in Iowa. Population distribution for all 99 Iowa counties was calculated using the 2000 census, and 5 Iowa counties (Dallas, Jasper, Johnson, Linn, and Polk) were calculated using the 1990 census. Block level census data with the corresponding TIGER block outlines were used to distribute population in 3 arc second (approximately 90 meter) cells. An ArcView legend file (pop-3sec.avl) has been provided for optimal visualization of the population grid data.

#### **Origination:**

This data was developed by the Geographic Information Science and Technology Group at Oak Ridge National Laboratory (ORNL) using ORNL's LandScan USA population distribution model.

#### **Format:**

The population data is developed in ESRI ArcInfo Grid format.  
[An ArcView legend file (pop-3sec.avl) has been provided for optimal visualization of the population grid data]  
The population centroids (points) are developed as ESRI shape files.

#### **Projection and Resolutions :**

The population data has a 3 arc second (approximately 90 meter) spatial resolution.  
It is in geographic (latitude-longitude) projection with NAD 83 datum.  
It represents a nighttime or residential population distribution.

#### **Data Organization:**

The data for 1990 and 2000 are in the directories IA1990 and IA2000 respectively. These directories also contain the corresponding Census data files used in developing the LandScan USA population distribution data.

For each year, the data for each county is in a separate subdirectory.

Each County subdirectory has the LandScan USA population distribution grid and the population centroids in ArcView shape file format.

[A file, Tbl\_matrix.txt, is included for the Census year 1990 that provides explanation of the naming convention for the centroid shape files. For Census year 2000, the population centroid file names are generally self-explanatory as the file names correspond to the demographic field names in the Census data.]

For the year 2000, each county has 76 centroid point shape files providing the population centroids for blocks, block groups, tracts, and counties. Each population accounting unit, e.g. blocks, has 19 centroid shape files representing each of the 19 different demographic groups.

For the year 1990, each county has 125 centroid point shape files providing the population centroids for blocks, block groups, tracts, and counties. The blocks have 8 demographic shape files, while the others all have 39 demographic shape files. For the 1990 census a limited number of demographic groups were available at the block level compared to the 2000 Census data.

The attribute table for each centroid shape file contains the fields for:

Stfid	Census ID
LatCen	Census Latitude
LonCen	Census Longitude
Flag	A field denoting the source of the population centroid: 0 = Calculated from the LandScan USA population within the shape 1 = Copied from the Census latitude-longitude because of zero population in a specific demographic group in the shape 2 = Copied from the Census latitude-longitude because of zero net population in the shape 3 = Copied from the Census latitude-longitude because a block was too small to grid even at 1 arc second resolution in the LandScan computation.
Lat	Latitude of population centroid (See Flag field above)
Lon	Longitude of population centroid (See Flag field above)

As noted above, only those records with a flag of 0 value are true population centroids. The others are given for completeness.