INCORPORATING SECOND-ORDER PROPERTIES IN "DIGITAL POPULATIONS" FOR CLUSTER DETECTION ANALYSIS AND AGENT-BASED MODELING

Charles R. Ehlschlaeger[1]

[1]Department of Geography, Western Illinois University,
Macomb, IL 61455, USA
Tel: +1 309-298-1841
FAX +1 309-298-3003
Email: CR-Ehlschlaeger2@wiu.edu

Abstract

"Digital Populations" research explores the question, "Can the use of publicly available United States Census Bureau data and publicly available land cover data generate models of simulated households in the US that are useful for performing cluster detection analysis in medical geography?" Digital Populations is useful in situations where researchers desire uncertainty estimates of an application requiring US Census Short Form Data, which is commonly available in most GIS at the census block level. Version 1.0 of Digital Populations was developed to represent population data as individual households located in geographic space mimicking census attributes' first-order properties. Digital Populations uses Monte Carlo simulation to represent uncertainties of the data model by creating hundreds or thousands of possible population realities. Digital Populations also identifies statistically significant clusters with results similar to software such as SaTScan or ClusterSeer. This paper discusses Digital Populations version 2.0. Digital Populations 2.0 allows population or household attributes to mimic first- and second-order properties. Digital Populations 2.0 introduces a technique to estimate census attributes' second-order properties from Digital Populations 1.0 results for when survey data is unavailable to represent second-order properties.

1. Introduction

Digital Populations is an ongoing research project attempting to quantify the uncertainty of US Census data, especially the spatially explicit Short Form data commonly used in GIS applications. The term Digital Populations was used for this research as both a reference to "The Digital City," a series of conferences with the goal of forming a more structured presentation of urban space (Craglia, 2004), and an indication that realizations are done at the state level. Traditional measures of US Census data quality have focused on sampling technique errors. However, these measures cannot be easily used to determine whether US Census data can be useful for a particular application. In order to quantify the utility of spatial data for a particular application, researchers often employ Monte Carlo simulation (MCS). MCS requires hundreds or thousands of realizations of each input data layer with the application repeatedly run for each version of the input data layers (Heuvelink, 1998). Ideally, all knowledge of potential data errors is incorporated into the model that generates the data layer realizations.

The simplest implementation of US Census Short Form multiple realizations would be to generate multiple attribute tables with varying enumerations of population counts for each census block. However, multiple attribute tables would not account for the uncertainty of population location from aggregated data. Thus, Digital Populations took a more complex approach by representing every person and household with a separate location in each realization. If the locations in each realization are an accurate representation of possible households, GIS operations could be used that could potentially be more precise than was previously available. Thus, any urban agent-based application can easily represent the data model uncertainty by repeatedly running the application on all Digital Populations realizations. More traditional GIS applications relying on data aggregated to polygons would require separate functions that manipulate point based data layers or functions that rebuild polygon attribute tables from the point data. It is interesting to note that data distributed as a set of point layers ensures it is easy to aggregate the data to any choropleth scheme. For example, building attribute tables for zip code, city and township, or county boundaries would not add uncertainty nor potentially corrupt the data in any way.

The model Digital Populations uses to locate potential household locations is a conflation model (Cobb et al., 1998), and also known as data fusion (Wald, 1999). Conflation is the process by which multiple data layers are combined in order to generate a product containing the best aspects of each layer. Digital Populations uses publicly available United States Census Bureau data and publicly available land cover data to generate models of simulated households in the US. The land cover data is the National Land Cover Dataset (NLCD) and is a good example of conflation at work. NLCD uses Landsat images, census data, and road layers to better classify land use in the United States. Version 1.0 of Digital Populations was developed to represent population data as individual households located in geographic space mimicking census attributes' first-order properties (Ehlschlaeger, 2004).

In order to demonstrate the utility of point based census data realizations, Digital Populations identifies statistically significant clusters with results similar to software such as SaTScan or ClusterSeer. (In the medical community, significant clusters are known as "hotspots.") Identifying hotspots requires a measure of the significance of each hotspot. Significance, when measured as a P-value, is a form of uncertainty estimate. For example, a P-value of 0.001 indicates that only one time in a 1,000 will something measured be inaccurate. Accurate P-values require the uncertainty of the input data to be understood.

This paper discusses Digital Populations version 2.0. Digital Populations 2.0 allows population or household attributes to mimic first- and second-order properties. Although field surveys would probably provide superior second-order property estimates, Digital Populations 2.0 introduces a technique to estimate census attributes' second-order properties from Digital Populations 1.0 results. Section two reviews issues of US Census data relevant to cluster mapping and this research. Section three will provide a background on hotspot mapping. Section four covers the methodology behind Digital Populations version 2.0. Section five discusses a Digital Populations case study using Rhode Island data and a sub population with a simulated disease. It will also compare the results against Kulldorff's

Spatial Scan statistic. Finally, section six will discuss this research and issues that must be overcome before Digital Populations becomes

2. Brief review of US Census data

This section provides a brief summary of US Census data as it relates to this research. After a discussion on the types of products is available from the US Census Bureau, this section clarifies which data products have the greatest potential for developing Digital Populations.

There are multiple data products from the US Census. They come from three sources: 1) the Decennial Short Form questionnaire, 2) the Decennial Long Form questionnaire, and 3) the American Community Survey phone survey. All three sources collect information at the household level. They contains brief questions for each person in the household about gender, age, nine categories of race plus “other,” and relationship to 1st person in household, as well as the type of dwelling.

The first two sources, the Decennial Short Form questionnaire and the Decennial Long Form questionnaire, are descriptively named after the forms and are delivered every 10 years. On April 1, 2000, very household in the US was to receive Form d61a. Form d61a only contains the questions described above and is the foundation of most demographic data used in GIS in the US. GIS users see this data as census blocks described as polygons with attribute tables containing the enumeration of various age groups, gender, and race. Summary File One (SF1) (US Census, 2002) and Summary File Two (US Census, 2002b) data products use the information from the Decennial Short Form questions.

On April 1, 2000, one in six households in the US was to receive the Decennial Long Form questionnaire. Summary File Three (SF3) (US Census, 2002c), Summary File Four (US Census, 2002d), and the Public Use Microdata Sample (PUMS) (US Census, 2003) use data from the Long Form questionnaire. All three Long Form data products provide information in the form of two tables: The household table contains detailed information such as number of rooms as well as bedrooms, year moved into, occupants per room, year structure built, expenses, plumbing and kitchen facilities, vehicles available, value of home, monthly rent and more. In addition to the variables available in SF1 and other Short Form data products, the population table contains detailed information such as marital status, whether grandparents are caregivers, language and ability to speak English, ancestry, place of birth, citizenship and veteran status, place of work and commuting distances, educational attainment, disability, income, employment status and poverty status and others. The data in SF3, SF4, and PUMS are presented in the form of “typical” households and “typical” population members. The tabular nature of the data is contusive to SQL queries. The following query, for example, would determine the proportion of people living in a state that were veterans, disabled, and over the age of 65 on April 1, 2000:

```
Select all from SF3p where AGE > 65 AND VETERAN_STATUS = true AND DISABILITY_STATUS = true
```

The third source of public available demographic data is from the American Community Survey (ACS) phone questionnaire (US Congress, 2001b). ACS phone surveys are done every year with a data product similar to PUMS. While several researchers have scoffed at ACS's sampling methodology, there are inherent advantages to annual surveys, especially in rapidly changing neighborhoods. As Goldstein et al. (2004) and many others have discussed, uncertainty increases with the difference in time between a sample and for when data is needed. ACS's sampling methodology does reduce its utility for SQL commands to determine the proper proportion of population or household variables. Using 2000 SF1 and ACS data for Rhode Island, SF1 indicates there are 155,423 people with ages from 50-64. A query of ACS data would indicate 176,300 people in that age bracket. The same is true for occupied households, SF1 has 31,413 while ACS has 23,100.

Of the various US Census data products, SF1 and ACS provide the most useful information in a Digital Populations environment. While ACS data doesn't weight the sampling of households and population against the SF1 like SF3 does (US Census, 2002c), the annual surveys reduce the uncertainty that neighborhoods will be misrepresented. Digital Populations conflates SF1 and ACS data together and provides a mechanism to weight SF1 proportions to ACS proportions based on how much confidence the user has in each product.

3. Recent research on hotspot mapping

The section discusses hot spot mapping as it relates to the medical community and the conditions under which US Census and Digital Populations data is used for this purpose.

In order to work properly, typical cluster detection algorithms require two components: 1) events or cases of interest (some illness), and 2) the population, usually people that could potentially become an event. It might normally be the case that the entire population of people is susceptible to the illness. However, many illnesses will only affect a subset of the population. Or it is the case that researchers are interested in looking at clustering patterns of high risk subpopulations. For these situations, it is often easier to create the event data than the population data. For example, medical researchers can usually get access to the events for a disease serious enough to require medical care. It is far more difficult to get an accurate and precise enumeration of a target population in countries without universal health care.

In this research's case study, the goal was to determine whether some areas of Rhode Island were not getting proper medical care for the early detection of breast cancer in the population of older African-American women. If the enumeration of older African-American women were available at the census block level of aggregation, standard cluster detection software such as SaTScan (Jemal et al., 2002) would be used locate potential clusters. Cancer researchers with access to late-stage breast cancer medical records could find all 40-64 year old African-American women with the disease, geocode their addresses, and aggregate those points to census tracks. SaTScan would then take the late-stage breast cancer data, the population of older African-American women, and the centroids of census blocks and perform its spatial scan statistic. SaTScan's spatial scan statistic uses a procedure similar to the geographical analysis machine (GAM) (Openshaw et al., 1987).

GAM determines the P-Value of a potential cluster by creating 100's or 1000's of possible patterns of events by giving each member of the population a chance of being an event. GAM then generates circular regions of varying sizes throughout the study region. Each circular region is considered to be a potential cluster. GAM compares the actual number of events against each of the possible patterns of events to determine the P-Value of that potential cluster. For example, if 999 possible patterns of events were generated for a potential cluster and there were three of patterns with as many or most events than the actual events, the P-Value would be (1+3)/(1+999) or 0.001.

SaTScan, in addition to generating a P-Value for potential clusters, uses the binomial or Poisson model to determine which potential cluster have "the minimum likelihood for random occurrence" (Kulldorff, 2004). The Poisson's minimum likelihood function is:

$$k = (c / n)^{c}([C\text{-}c]/[C\text{-}n])^{(C\text{-}c)} \qquad (1)$$

where *k* is minimum likelihood function, *n* is covariate adjusted expected cases in the potential cluster, *C* is global number of events, and *c* is the number of cases in the potential cluster. SaTScan keeps a record of the most extreme potential clusters allowing while discarding potential clusters that overlap more extreme clusters. O'Sullivan & Unwin (2003) provides an excellent overview of GAM and the mathematics behind SaTScan.

However US Census data is not always easily incorporated into an analysis. There are two potential issues. First, significant proportions of Americans declare themselves to be multi-race, making the race, age, and gender in SF1 less accurate. The ecological fallacy problem makes it difficult to get an accurate count of any subpopulation not explicitly enumerated. Detailed enumeration of exclusive race, age, and gender is aggregated to the census block group in Census Short Form tables. Because of ecological fallacy, it would be impossible to determine how many 40-64 year old African-American women live in a census block if we only have the enumeration of blacks and women of certain ages. (One cannot assume that 25 older African-American women live in a census block with a population of 100 that contains 50 African-Americans, and 50 older women.) Second, Short Form data is tabulated only every ten years. The population of many parts of the United States, neighborhoods catering to recent immigrants in New York City for example, experience massive turnover in only a few years.

Ecological fallacy and time between surveys represents two major issues covered by Digital Populations. Digital Populations also creates a model that represents what the actual distribution of the population might be. The next section discusses Digital Populations methodology.

4. Digital Populations methodology

GAM and SaTScan's use of multiple realizations of random events in order to estimate P-Value meshes nicely into the Monte Carlo simulation (MCS) approach for uncertainty analysis. Digital Populations has two conceptual algorithms. 1) Generating multiple realizations of potential household location, discussed in Section 4.1, and 2) Identifying potential clusters that have the minimum likelihood of being a random event (Section 4.2).

4.1 Conflating land cover and census data

Digital Populations uses a technique for spatially locating realizations of households in the U.S. by conflating Short Form (SF1) tables, American Community Survey (ACS) tables, and National Land Cover Data (NLCD). Since ACS tables cover entire states, Digital Populations must be created state by state. There is a three step process for generating Digital Populations: 1) Identify the heterogeneous probability function for the study area. 2) Determine how many realizations of each ACS household needs to exist to mimic SF1 data for each Digital Populations realization. 3) Spatially locate realized ACS households in the study area.

4.1.1 Digital Populations heterogeneous probability function

There are three levels of complexity when generating point patterns that mimic real world processes: 1) homogeneous Poisson processes, 2) heterogeneous Poisson processes, and 3) Cox processes (Bailey & Gatrell, 1995).

Homogeneous Poisson processes assume that an object is equally likely to be located at any point in the study as any other. If we were to ignore population information and only study event locations, we could find clusters of events denser than in other areas. However, most clusters will only exist because the events occurred in densely populated areas and events are likely distributed by chance.

Heterogeneous Poisson processes recognize that different parts of the study area have different likelihoods of containing events. SaTScan conceptually assumes a heterogeneous Poisson process since different census blocks will have different population densities. However, since census blocks are identified solely as centroids, it is impossible to discover small diameter cluster. Digital Populations version 1.0 fully implements a heterogeneous Poisson process. Within each census block, Digital Populations assumes each NLCD land cover class has a different density function.

Digital Population version 2.0 has a Cox process approach. The location of Digital Populations version 2.0 ACS households are determined both by the density function defined by census block density, the relative density of different NLCD land cover classes, and the location of ACS households with similar attributes. For example, some neighborhoods have an older population than others because many older people prefer to live near other older people. The same is generally true for race. Digital Populations version 2.0 realization process is described in Section 4.1.3.

Iterative regression analysis of SF1 and NLCD at the census block level determines the relative household density of different NLCD land cover classes. The process is iterative because the regression analysis will determine that some NLCD land cover class will have negative density when the entire set of land cover is used. Land cover classes with negative density are removed from consideration and the regression analysis is performed until all classes have positive density. The regression analysis is shown in equation 2:

$$h_i = SUM_k(d_k\, c_{kj}) + e_i \qquad (2)$$

Where h_i is number households in SF area i, d_k is relative household density in NLCD class k, c_{kj} is area of NLCD class k in SF1 area i, and e_i is error of SF area i.

4.1.2 Digital Populations conflation of disparate SF1 and ACS variables

As mentioned in Section 2, SF1 and ACS variables do not have the same values. Using 2000 SF1 and ACS data for Rhode Island, SF1 indicates there are 46,908 blacks. A query of ACS data would indicate 25,500 blacks. Since PUMS and ACS are sampled households, there is no compelling reason to create 100 households simply because the table is a 1% representation of the population. For example, since the number of ACE households for Rhode Island under represents blacks, there should be more than 100 copies of an ACS household containing blacks than ACS households without blacks.

For each new household in a realization, Digital Populations first picks 10 potential households that will "improve the fit" of the important SF1 variables. In the case study, the important variables are gender, black race, and people with ages from 40 to 64. Digital Populations then determines which of these potential households will more closely maintain the proportion of important SF1 variables should that household be selected. This algorithm is sensitive to the number of households already realized earlier in the process. Digital Populations uses a least squared error approach for both attempting to realization 100 1% ACS households and the exact number of SF1 variables. Users can determine a greater weight on SF1 variables while virtually ignoring ACS household counts or visa versa. Obviously, if the application is for a year when the US Census was done, users should place greater weight on the SF1 variables. However, if the application year is far removed from an actual census, it might be better to add greater weight to the ACS household counts. This step greatly diminishes the sampling methodology drawback to the ACS. This process is iteratively performed until the SF1 is fully enumerated.

4.1.3 Spatially realizing ACS households conditionally with a Cox Process

At this stage of the Digital Populations process, ACS households are placed within the study area. There is a three step process in to generate a realization with Digital Populations version 2.0: 1) ACS households are randomly located throughout the state with a heterogeneous Poisson process. 2) ACS households that contain population members that could be events are then searched to find those closest to known events. These ACS households are then shifted to the event locations. 3) ACS households are stochastically shifted to different census blocks if the shift will cause a better fit to the important SF1 attributes.

ACS households are randomly located throughout the state with a heterogeneous Poisson distribution algorithm using household density that fits any number of SF1 variables' first-order properties. It is probably best to only fits SF1 variables relevant to target population. In experiments so far, SF1 variables were almost always exactly matched when three or fewer SF1 variables were fit. Fitting six or more SF1 variables inevitably caused some SF1 variables to be poorly fit. (This was expected as similar results occurred in other algorithms recreating multiple statistics (Ehlschlaeger, 2002). At this stage of the realization process, each NLCD land cover class in each census block has a uniform household density. For example, all multi-family housing within a census block will have the same density. This might cause minor misrepresentations in the final product should all the housing on the north side of a census block be tall buildings while multi-family housing on the census block's south side are single story. If positive spatial autocorrelation of housing density exists, Digital Populations' first-order heterogeneous Poisson distribution realizer would not capture that phenomenon.

To correct for a uniform household density, the variogram for specific SF1 variables is fitted. If the range of the second-order property's variogram was large enough, Digital Populations version 2.0 would typically cluster more households within a census block towards nearby census blocks with higher density. This will convert the uniform household density function to something with a topography more similar to (and more realistic than) Tobler et al's (1979) pycnophylactic method. Results would be more realistic than the pycnophylactic method at distances shorter than the range, because Digital Populations 2.0's realizer will create varying population densities within census blocks. Digital Populations 2.0's second-order interpolator uses data from ACS households and SF1 tables to estimate a variogram. The variogram's nugget is the variance of the American Community Survey population attribute within households. Variogram lags with distances greater than typical subdivisions can be determined by measuring Digital Populations 1.0 realizations.

While Digital Populations 1.0 only demonstrated a first-order heterogeneous Poisson distribution realizer, this research demonstrates how sampling of second-order properties of race, and/or other SF1 variables would provide better distribution results. ACS households and ACS derived Short Form tables in Digital Populations keeps the population more up-to-date than the once in ten years traditional SF1 data. Finally, distributing American Community Survey households over land use maps reduces the errors caused by aggregating census block information to that block's center.

4.2 Hotspot mapping accounting for uncertainty

The Monte Carlo cluster detection algorithm is conceptually similar to Openshaw's (1987) Geographical Analysis Machine. A regular lattice is laid across the study area. Each lattice point is the center of kernel functions with varying diameters. Householders are weighted by distance to the lattice point. For each kernel function, the proportion of events against population is compared against hundreds or thousands of realized Digital Populations with simulated events to determine a P-value and minimum likelihood of randomness function value similar to Kulldorff's Spatial Scan statistic (Jemal et al. 2002). Since popular cluster detection techniques such as Kulldorff's Spatial Scan statistic already use Monte Carlo simulation to represent the P-value of unlikely event clusters, this methodology provides an efficient potential solution to the ecological fallacy problem as well as a more accurate representation of the maximum likelihood function.

There are two differences between the mathematics behind Digital Population's spatial scan statistic and Kulldorff's spatial scan statistic:

1) Digital Populations' equation must account for the varying population numbers within a potential cluster. Digital Populations' minimum likelihood function is:

$$k = (c / n)^{c}([C-c]/[C-n])^{(C-c)} \qquad (3)$$

where k is minimum likelihood function, n is the average covariate adjusted expected cases in the potential cluster across all realizations, C is global number of events, and c is the number of cases in the potential cluster.

2) Digital Populations allows for the use of kernel function to define a potential cluster. Ergo the variable c is no longer an integer, and both variables c and n are adjusted based on the distance to the center of the potential cluster. Ergo, events and non events at the edge of a potential cluster have much less weight than events and non events at the center of a potential cluster.

5. Case study results

This section compares the results from Digital Populations against an analysis done with the appropriate data using SaTScan. The case study used simulated data of older African-American women with late stage breast cancer in Rhode Island for the year 2000. The analysis was done at the census tract level instead of census block to more easily see result patterns. Digital Populations was able to take advantage of the exact locations of the simulated events while the simulated events were aggregated to census tracts for use in SaTScan.

Digital Populations took over a month of computer time to generate 250 realizations of alternative breast cancer and population realizations on a 3.2 MHz Pentium IV computer. SaTScan performed its analysis in seconds. However, SaTScan was unable to find the two most "unlikely to be random" Digital Populations clusters in Rhode Island. Both were smaller clusters that were diffused by the census tract aggregation. The most extreme Digital Populations cluster was part of SaTScan's 110th most extreme cluster with an insignificant P-Value. These results were mainly caused by the heterogeneous Poisson ACS household realizer.

6. Conclusion and discussion

While Digital Populations shows the potential for providing much superior results than traditional spatial scan statistics, there are many techniques that can improve the quality of Digital Populations. The most important technique that needs to be implemented is the preconstruction of Digital Populations realizations. While it took a month to build the realizations, Digital Populations' q-tree implementation of its spatial scan statistic generated results within minutes. Digital Populations will only become a useful product if the household realizations are pre-built with algorithms to conditionally fit event data and better fit relevant SF1 variables when needed. Accuracy and uncertainty issues are also a critical area for improvement.

Here is a list of model improvements those that would improve the uncertainty analysis capabilities of Digital Populations.

- Improve population density 1st order properties: Instead of locating individual households, locate multi-unit housing as single locations in proportion to ACS Units in Structure (BLD).
- Improve population density 2nd order properties: 2nd order properties are modeled by semi-variogram in Digital Populations' software. Variogram lags are normally determined by distance. Instead, lags could be based on household order. This way, 2nd order properties in high density areas will not overwhelm the modeling of 2nd order properties in low density areas.
- The regression of raw NLCD classes is not best measure of relative household density. For example, open water has positive household density. Digital Populations versions 1.0 and 2.0 automatically excluded open water from the regression analysis. Also, orchards have VERY high household density. Did the NLCD classify migrant worker housing as part of the orchard class? Or, does land near orchards have higher household density? One possible solution would be to model each NLCD grid cell as a vector of distances to closest cell of each NLCD class. The regression analysis would then solve for each NLCD land cover class's distance decay formula.
- So far, Digital Populations has treated SF1 as accurate. Uncertainty in SF1 data is well known, but not well quantified. Theoretical research is necessary to best determine ways of realizing SF1 uncertainty. Currently, the US Census samples SF1 errors by region. These error samples are reported (US Congress, 2001), however, it would difficult and somewhat arbitrary to use these reports to build an uncertainty model. In one tries, they would probably assume a normal distribution of errors and apply uncertainty modeling to create realizations of SF1 variable enumerations.
- Realizing land use and land cover uncertainty (Ehlschlaeger & Goodchild, 1994; Ehlschlaeger, 2000). Each Digital Populations realization would get its own land use map accounting for uncertainty of NLCD. The regression analysis determining relative land cover density would need to be rerun for each realization. This stage of the Digital Populations process only takes seconds while generating realizations is many times more computationally expensive.

Browsing through the list of possible improvements above, it is easy to realize that Digital Populations could easily become an extremely complicated model. As anyone who has researched uncertainty analysis should know, a data uncertainty model may be as simple or complex as the designer wishes. Scientists usually try to determine the simplest model to explain a phenomenon. This approach used in uncertainty modeling will often generate realizations that have little to no chance of being an actual representation of reality.

8. References

Bailey, T., Gatrell, A., 1995, *Interactive Spatial Data Analysis*, London: Longmann.

Cobb, M. Chung, M., Foley III, H., Petry, F., and Shaw, K., 1998, A Rule-based Approach for the Conflation of Attributed Vector Data. *GeoInformatica*, **2**, 7-35.

Craglia, M., 2004, Cogito ergo sum or non-cogito ergo digito? The Digital City revised. *Environment and Planning B*, **31**, 3-4.

Ehlschlaeger, C., Goodchild, M., 1994. Dealing with Uncertainty in Categorical Coverage Maps: Defining, Visualizing, and Managing Data Errors. *Proceedings, Workshop on Geographic Information Systems at the Conference on Information and Knowledge Management*, Gaithersburg, MD.

Ehlschlaeger, C., 2000, Representing Uncertainty of Area Class Maps with a Correlated Inter-Map Cell Swapping Heuristic. *Computers, Environment and Urban Systems*, **24**, 451–469.

Ehlschlaeger, C., 2002, Representing multiple spatial statistics in generalized elevation uncertainty models: moving beyond the variogram. *IJGIS*, **16**, 259-285.

Ehlschlaeger, C., 2004, Digital Populations: Building multiple realizations of population for cluster detection analysis. *GIScience 2004 Extend Abstracts* (Aldephi MD), pp. 77-79.

Goldstein, M., Candau, J., Clarke, K., 2004, Approaches to simulating the "March of Bricks and Mortar". *Computers, Environment and Urban Systems*, **28**, 125–147.

Heuvelink, G., 1998. Error Propagation in Environmental Modelling with GIS, Taylor & Francis, 127 pgs.

Jemal, A., Kulldorff, M., Devesa, S., Hayes, R., and Fraumeni Jr., J., 2002, A Geographic analysis of prostate cancer mortality in the United States, 1970–89. *Int. J. Cancer*, **101**, 168-174.

Kulldorff, M., 2004, *SaTScan^{TM} User Guide for version 5.0*. URL: http://www.satscan.org/.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A., 1987, Developing a mark 1 geographical analysis machine for the automated analysis of point data sets. *Int. J. Geographical Information Systems*, **1**, 335-358.

O'Sullivan, D., Unwin, D., 2003. *Geographic Information Analysis*. John Wiley and Sons.

Tobler, W., Deichmann, U., Gottsegen, J., and Maloy, K., 1997, World population in a grid of spherical quadrilaterals. *Int. J. of Population Geography*, **3**, 203-225.

U.S. Census Bureau, 2002, *Census 2000 Summary File 1 – United States*.

U.S. Census Bureau, 2002b, *Census 2000 Summary File 2 – United States*.

U.S. Census Bureau, 2002c, *Census 2000 Summary File 3 – United States*.

U.S. Census Bureau, 2002d, *Census 2000 Summary File 4 – United States*.

U.S. Census Bureau, 2003, *Census 2000, Public Use Microdata Sample, PUMS, United States*.

U.S. Congress, Committee on Government Reform. Subcommittee on the Census, 2000, *Oversight of the 2000 census: examining the status of key census 2000 operations: hearing before the Subcommittee on the Census of the Committee on Government Reform, House of Representatives, One Hundred Sixth Congress, second session, February 8, 2000*.

U.S. Congress, Committee on Government Reform. Subcommittee on the Census, 2000b, *Oversight of the 2000 census: status of non-response follow-up and closeout: hearing before the Subcommittee on the Census of the Committee on Government Reform,* *House of Representatives, One Hundred Sixth Congress, second session, June 22, 2000*.

U.S. Congress, Committee on Government Reform. Subcommittee on the Census, 2001, *Success of the 2000 census: hearing before the Subcommittee on the Census of the Committee on Government Reform, House of Representatives, One Hundred Seventh Congress, first session, February 14, 2001*.

U.S. Congress, Committee on Government Reform. Subcommittee on the Census, 2001b, *Census Bureau's proposed American Community Survey (ACS): hearing before the Subcommittee on the Census of the Committee on Government Reform, House of Representatives, One Hundred Seventh Congress, first session, June 13, 2001.*

Wald, L., 1999, Some Terms of Reference in Data Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 1190-1193.