Geographic data misalignment: Practical solutions for integrated data analysis

B. Bhaduri[1], E. A. Bright, P. R. Coleman, A. N. Rose, A. M. Goss Eng[2], W. W. Hargrove[3]

Oak Ridge National Laboratory, PO Box 2008 MS 6017, Oak Ridge, TN 37831-6017
Email: bhaduribl@ornl.gov; brightea@ornl.gov; colemanpr@ornl.gov; rosean@ornl.gov

1. Introduction

Geographic Information Systems (GIS) provide an effective environment for integrated analysis using multiple spatial data. In a GIS, overlay and visual exploration of relationships among two or more geographic features is the simplest spatial analyses. However, taking advantage of the real power of a GIS, integrated analysis of using two or more data layers are more common where the topological relationships between the layers are exploited while modeling a spatial process. Spatial processes are scale dependant and it is desirable that the level of analysis captures the maximum geographic variability (Tobler and Mollering 1972; Woodstock and Strahler 1987; Buttenfield and McMaster 1991) and best describe the process being modeled (Goodchild 2001). Often geographic data are generated at different geographic scales (the ratio between the sizes of an object on a map and in reality), spatial resolutions (smallest object size that can be represented), and times. Moreover, Geographic data production is subjected to variable spatial accuracy (relationship between a measurement and the reality it purports to represent) and precision (the degree of detail in the reporting of a measurement) (Goodchild, 1991). This leads to the common phenomena of spatial and temporal misalignment of geographic data where topological relationships of two different geographic objects remain ambiguous. In many situations, for the purpose of minimizing such ambiguity, it is ideal that the data used in the same analysis should represent the same geographic scale and resolution as well as temporal currency. However, spatial analyses in the real world involve data generated by multiple agencies with disparate scale, accuracy, integrity, and update frequency. It becomes necessary to develop analytical approaches that provide best possible way to clarify the topological ambiguity and rectify the spatial and temporal misalignment. To illustrate such scenarios of spatial and temporal misalignment of data, here we describe two case studies: one on the delineation of contributing upstream watershed from drinking water intake locations and the second on spatial decomposition of Census data with a smart interpolation technique respectively. It should be noted that these national databases used in these studies continue to evolve in their overall qualities. Some of the data challenges encountered at the time of study have subsequently been addressed in the later years. The objective is not to focus on the shortcomings of the databases, but to generalize the problems (and possible solutions) that continues to exist as geographic data of disparate lineage and hence consistency, integrity, and quality are integrated for analysis.

---

[1] Corresponding author.

[2] Presently with the Office of Biomass Program, U.S. Department of Energy, Washington, DC

[3] Presently with the U.S. Forest Service, U.S. Department of Agriculture, Asheville, NC

2. Spatial misalignment: The case of watershed delineation

2.1 Background

Vulnerability of surface water supplies to pesticide residues in raw water depends in part on pesticide usage in a watershed, which in turn depends in part upon land use/cover characteristics, such as percent cropped acreage. To advance the national scale geographically-targeted monitoring and mitigation capabilities at a watershed level, a collaborative effort among the U.S. Environmental Protection Agency (EPA), the United States Geological Survey (USGS), and Oak Ridge National Laboratory (ORNL) was initiated in the year 2000. The goal was to georeference all known community water supply (CWS) surface water intakes in the continental United States to the newly-available National Hydrography Dataset (NHD), and to delineate upstream contributory watersheds for each intake. Using the delineated watersheds, land use/land cover statistics for each watershed were obtained from the National Land Cover Dataset (NLCD), and pesticide usage was estimated from other available data. For the relevance of spatial misalignment, the following discussion focuses on the first task of georeferencing the intake locations to the hydrography network. The characteristics of the two databases are highlighted in Table 1.

Table 1. Brief descriptions of the two data sets used in this analysis.

Description	Resolution	Notes
Drinking Water Intakes Database		
Public Supply Database (PSDB) created by USGS (Focazio et al 2000) 6945 point locations	Unknown and variable resolution and quality Points	A 2001 version originally extracted from USEPA's Safe Drinking Water Information System (SDWIS) was used. 6361 locations for the Contiguous U.S. were used in the analysis.
Hydrography Database		
Medium resolution National Hydrography Database (NHD) (Simley and Carswell 2009) Approximately 3 million stream segments and water bodies	1:100,000 Lines and polygons	A May 2001 version was primarily used. However, n earlier version from October 1999 and a later version of NHD from June 2002 for solving individual analytical errors that resulted from identifiable anomalies in the May 2001 version of the NHD data.

Initial examination of the intakes locations revealed a spatial misalignment with the hydrography network. Consequently, contributory upstream watershed delineation for drinking water intake locations required topologically associating the CWS intakes locations to the hydrography network, in particular the stream centerlines. Given the variability in scale and resolution, and uncertainty in the positional accuracy of the intake locations, intakes were often close to two (or more) NHD reaches (Figure 1). Moreover, some of the intakes were associated with hydrographic features that were not captured in the 1:100K scale NHD

data. Thus associating them with the nearest stream reach using a simple GIS 'snap' function was not deemed to be the optimal method. Instead, an approach through attribute matching, between the NHD and the intake database, was developed to increase the level of certainty in those associations.

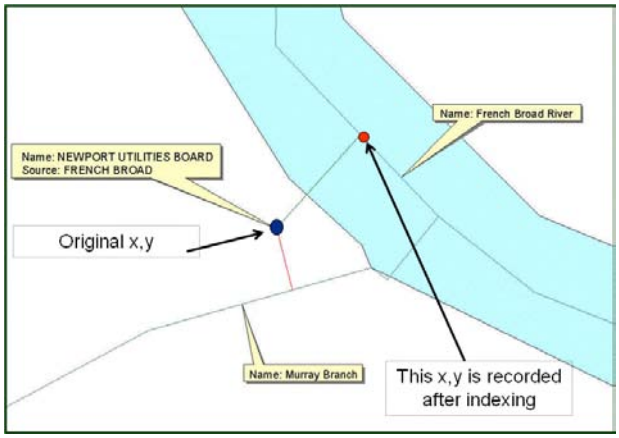


Figure 1. Example of georeferencing of a CWS intake location to the NHD.

2.2 Technical Approach

A vastly automated algorithm (based on attribute matching, feature characterization, and proximity analysis) that evaluates multiple spatial and non-spatial attributes from the CWS database and the NHD was developed that assigned each intake to the most appropriate stream/reservoir network in the NHD.

For each CWS location, two nearest NHD reaches (linear and/or polygon) were selected. Both selected features were tested with the georeferencing algorithm. The georeferencing model (algorithm) performed an extensive conditional test based on five primary parameters:

1. Name of source for intakes (N),
2. Name of river associated with intakes (A),
3. Distance to Reach (D),
4. Linear vs. Polygon Reach condition (P), and
5. Ratio of the two Distances from the intake location to the two nearest reaches (R).

Alphanumeric flag values were assigned for each of the five parameters in the form of [*x*N *x*A *x*D *x*P *x*R], where N, A, D, P, and R represents the five parameters respectively. For both NHD reaches, the first four numeric values (*x* from the alphanumeric part) were added to produce final flag values (Flag 1 and 2 for the nearest and the second nearest reach respectively). Initial testing of the algorithm indicated that inclusion of the fifth parameter (R) in determining the final flag value introduces undesired error in the results. Consequently, values for the fifth parameter (R) is determined and stored but not evaluated as part of the final flag value calculation. It is only consulted in certain situations where anomalies were detected during the verification and validation process. Because of the design (of the algorithm), lower flag value indicated a better match for any intake location

and the corresponding reach was selected as the best or most appropriate reach the intake should be georeferenced to. The following section describes the conditional testing algorithm to determine the most logically appropriate reach to which an intake should be indexed.

Parameter 1 and 2: Name of source (xN) and Name of associated river (xA)

The CWS intake database had a source name for every intake location. In addition there is an attribute field for the name of associated river for each intake. Although the source name field was always (100%) populated, it was not true for the associated river name field (49% populated). These two name fields were the only spatial link to the NHD, which had names for linear and polygon reaches. The following procedure was generally applied when matching the "source name (*Sourcename*)" and the associated river name (*Riverresla*)" fields for each intake location to the two nearest NHD reach names.

To facilitate the spatial association, names from these two fields were first standardized by expanding abbreviations (Table 2).

Table 2. Standardized abbreviations for field names.

Abbreviation	Standardized	Abbreviation	Standardized	Abbreviation	Standardized
BK, BROK, BROO	BROOK	NW	NORTHWEST	WBR	WEST BRANCH
BLCKWTR, BLKWTR	BLACKWATER	OFFST	OFFSTREAM	WF, WFK	WEST FORK
BLK	BLACK	QTR	QUARTER	WTRSHD	WATERSHED
BR, BRCH, BRA	BRANCH	R, RI, RV, RVR, RIV, RIVR, RIVE	RIVER	E	EAST
BYWY	BYWAY	RES, RESEVOIR	RESERVOIR	EFK	EAST FORK
CH, CHNL, CHAN	CHANNEL	RD	ROAD	IRR	IRRIGATION
CNL, CAN, CANA	CANAL	RN	RUN	NE	NORTHEAST
CK, CR, CRK, CREE, CREK	CREEK	RT, RTE	ROUTE	U, UP	UPPER
CKS, CRKS	CREEKS	S, SO, SOU	SOUTH	OFFF	OFF
CNTR	CENTER	SBR	SOUTH BRANCH	OS, OFFSTR, OFFCHANNEL	OFFSTREAM
DTCH	DITCH	SFK	SOUTH FORK	DRW	DRAW
F, FK, FRK	FORK	SP, SPR	SPRING	GUL	GULCH
FR, FRM	FROM	SPRRNGS	SPRINGS	UN	UNNAMED
FK S, FKS	FORKS	STR, STRE, STRM	STREAM	UNT, UT	UNNAMED TRIBUTARY
LT, LTL, LTTL	LITTLE	SYS	SYSTEM	OFF STREAM	OFFSTREAM
L, LR, LOW	LOWER	ST	SAINT	OFF CHANNEL	OFFSTREAM
LK	LAKE	SW	SOUTHWEST	OFF	OFFSTREAM
M, MID	MIDDLE	SUPP	SUPPLY	SOUTH	SOUTHWES

				WEST	T
MN	MAIN	TR, TRB, TRI, TRIB	TRIBUTARY	SOUTH EAST	SOUTHEAST
MT	MOUNT	TRIBS, TRS	TRIBUTARIES	NORTH WEST	NORTHWEST
MTN	MOUNTAIN	TWN	TOWN	NORTH EAST	NORTHEAST
N, N0RTH, NO	NORTH	V, VLY, VY	VALLEY		
NF, NFK	NORTH FORK	W	WEST		

Subsequently, the names were split into two components:

1. **Name** Component (For example, Milton, Mississippi, Clinch, Colorado)
2. **Feature** Component (For example, River, Stream, Creek, Reservoir, Fork)

At this stage at least one or both components (Name and Feature) for the CWS source and NHD Reach would be present or absent. That presented 21 different possibilities of matches or mismatches (Table 3). The following rules were applied for matching:

1. Name component matches were ranked higher than feature component matches.
2. The *"present-absent*", *"absent-present"*, and *"absent-absent"* combinations were never considered as matches.
3. A *"present-present"* non-match was worse than a *"present-absent"* non-match, which was worse than an *"absent-absent"* non-match. (Considering no information is better than the presence of wrong information)

The 21 different combinations were further classified into 10 individual categories (Table 3). The first three categories were considered matches and the other seven categories were considered mismatches. By design, the categories indicated an ascending order of mismatch i.e. the categories are scaled between category 1 (Flag 1) that represents a perfect match and category 10 (Flag 10) that represents a complete mismatch.

Table 3. The scenarios are arranged in a descending order of "goodness of match".

PM: Present & Match ***PN***: Present & No-match ***A***: Absent (No-match)

Case #	**CWS Source Name**		**NHD Reach Name**		**Flag**
	Name	***Feature***	***Name***	***Feature***	
1	***PM***	***PM***	***PM***	***PM***	1
2a	***PM***	***PN***	***PM***	***PN***	

2b	***PM***	***A***	***PM***	***PN***	2
2c	***PM***	***PN***	***PM***	***A***	
2d	***PM***	***A***	***PM***	***A***	
3a	***PN***	***PM***	***PN***	***PM***	3
3b	***A***	***PM***	***PN***	***PM***	
3c	***PN***	***PM***	***A***	***PM***	
3d	***A***	***PM***	***A***	***PM***	
4	***A***	***A***	***A***	***A***	4
5a	***PN***	***PN***	***A***	***A***	5
5b	***A***	***A***	***PN***	***PN***	
6a	***A***	***PN***	***PN***	***PN***	6
6b	***PN***	***PN***	***A***	***PN***	
6c	***PN***	***A***	***A***	***PN***	7
6d	***A***	***PN***	***PN***	***A***	
6e	***PN***	***A***	***PN***	***A***	8
6f	***A***	***PN***	***A***	***PN***	
6g	***PN***	***A***	***PN***	***PN***	9
6h	***PN***	***PN***	***PN***	***A***	
6i	***PN***	***PN***	***PN***	***PN***	10

Parameter 2: Distance (xD)

The intake locations were previously verified by USGS to an accuracy of 6 arc seconds with respect to the reference data. Distance between an intake location and the two closest reaches were recorded and classified with the following flags:

Distance between intake and NHD reach ≤ 6 arc seconds: Flag = 0

Distance between intake and NHD reach > 6 arc seconds: Flag = 10

Parameter 3: Linear vs. Polygon Reach (xP)

Hydrologic features in the NHD are represented as linear (rivers, streams, canals etc.) and polygon (lakes, reservoirs, large rivers) reaches. Initially available data indicated that significant numbers of the intake points were present close to a water body (polygon) reach that also had a centerline (linear) reach). Because both the polygon and linear reaches may have the same name, it needed to be decided which reach the intake location should be referenced to. Initial investigation showed that in these scenarios, intakes were usually associated with lakes, reservoirs, and wide-body rivers (polygons) rather than linear reaches. Accordingly a polygon reach was rated higher than a linear reach in the algorithm as follows:

If selected reach is a polygon reach: Flag = 0

If selected reach is a linear reach: Flag = 10

Parameter 4: Relative Distance (xR)

Often the two closest NHD reaches for an intake point had similar names (equivalent flag values) or a missing name for the first reach (resulting in a higher match with the second reach). But visual investigations reveal that the distance between the intake point and the second reach is significantly greater than that between the intake and the first reach (Figure 2). For example, an intake location may be close to the "Clinch river", but the name field is empty for the NHD reach that represents the closest reach to that point. The second closest reach may be a long distance away from the point, but if its name field had "Tributary to Clinch River", the algorithm would produce a better match with the second closest reach for that particular intake. To logically identify and eliminate this problem, the algorithm produces a relative distance flag in the following way:

Distance between intake location and the first closest reach = d_1

Distance between intake location and the second closest reach = d_2

If $(d_1/d_2) \leq 3$: Flag = 0

If $(d_1/d_2) > 3$: Flag = Value of (d_1/d_2) rounded to the closest integer.

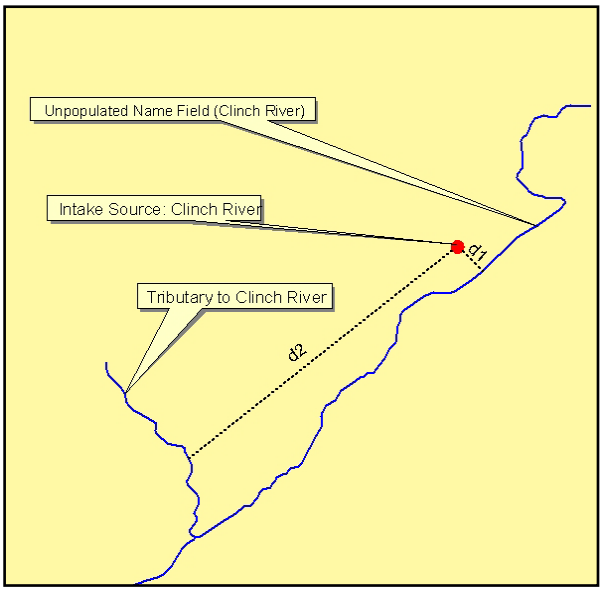


Figure 2. Hypothetical example where the relative distance between two reaches can be used to refine georeferencing model output.

Initial testing of the georeferencing algorithm and manual verification of the model results using 1775 intake locations indicated that the name matching part of the algorithm was the most critical control of the modeling process. For all intakes that produced a perfect name match, further review of the flag values during the verification suggested that the other conditional tests do not add any further value for selecting the most appropriate reach. Based on these observations, the flow of the georeferencing model was modified to streamline the algorithm (Figure 3).

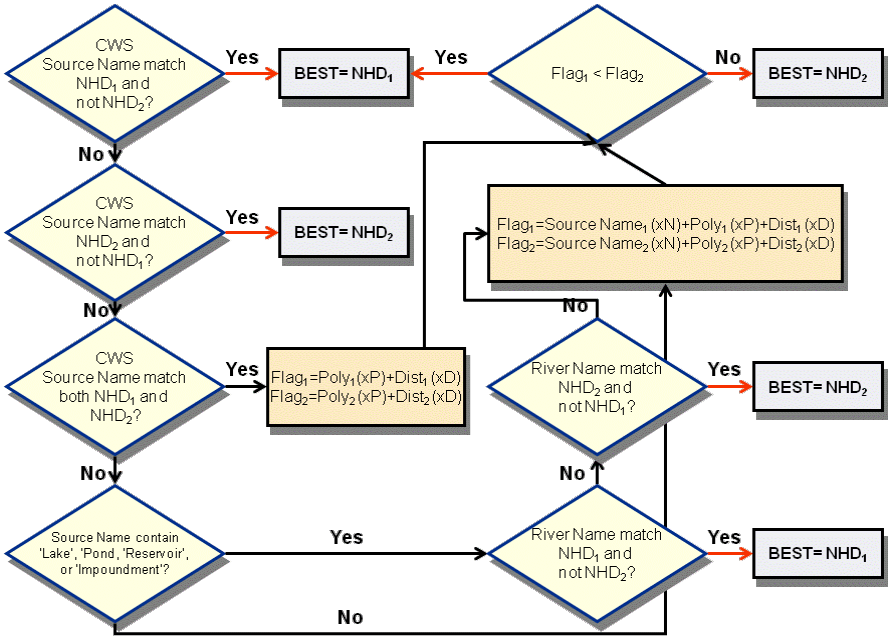


Figure 3. Flowchart describing the general functionality of the georeferencing model.

General Process of Conditional Testing

Compare CWS intake source name with the closest (NHD1) and the second closest (NHD2) NHD reach names and evaluate all flags (*x*N, *x*A, *x*D, *x*P, *x*R).

1. If source name matches the closest and not the second closest NHD reach, select the closest as the most appropriate reach for indexing.
2. If source name matches the second closest and not the closest NHD reach, select the second closest as the most appropriate reach for indexing.
3. If source name matches both the closest and the second closest NHD reaches
    a. Evaluate polygon (*x*P) and distance (*x*D) flags for both reaches.
    b. Add the numeric values to determine Flag1 and Flag2 and the reach associated with the lower flag value is selected as the most appropriate reach for indexing.
4. If source name does not match neither the closest nor the second closest NHD reaches
    a. If the source name contains the words “lake”, “pond”, “reservoir” or “impoundment”, compare source name with associated river name
        i. If river name matches the closest and not the second closest NHD reach, select the closest as the most appropriate reach for indexing.
        ii. If river name matches the second closest and not the closest NHD reach, select the second closest as the most appropriate reach for indexing.
    b. If the source name does not contain the words “lake”, “pond”, “reservoir” or “impoundment”, compare source name with associated river name
        i. Evaluate name (*x*N), polygon (*x*P), and distance (*x*D) flags for both reaches.
        ii. Add the numeric values to determine Flag1 and Flag2 and the reach associated with the lower flag value is selected as the most appropriate reach for indexing.

Once the logical association was established with a NHD reach with an intake using the general algorithm described above, the indexed intake location on the NHD is determined by:

a. Determining the closest location on the network if the logical association is with a linear reach (river, stream)
b. Determining the nearest location on the shore if the logical association is with a polygon reach (lakes, reservoirs) and then selecting the closest point on the drainage network (drains) inside the polygon reach.

2.3 Results

Out of the 6361 intake points georeferenced through the modeling process, 5084 (79.92%) were indexed to the closest NHD reach and 1277 (20.08%) were indexed to the second closest NHD reach. A total of 4207 (66.14) intakes were indexed to a linear reach (river,

streams) while 2154 (33.86%) intakes were indexed to a polygon reach (lakes, reservoirs). Detailed statistics of the georeferenced intakes are given in Table 4.

Table 4. Statistics of Georeferenced (6361) CWS Intake Locations

Georeferenced to	Linear Reach		Polygon Reach		Total
	With Name	No Name	With Name	No Name	6361
Nearest NHD reach	2461 *(48.41%)*	756 *(14.87%)*	1366 *(26.87%)*	501 *(9.85%)*	5084 *(100.00 %)*
	3217 *(63.28%)*		1867 *(36.72%)*		
2nd nearest NHD reach	990 *(77.53%)*	0 *(0.00%)*	164 *(12.84%)*	123 *(9.63%)*	1277 *(100.00 %)*
	990 *(77.53%)*		287 *(22.47%)*		

Analysis of the results from the name-matching algorithm indicated that 4496 (70.68%) intake locations had some level of match (name flags 1, 2, or 3) with the NHD reaches they were indexed to. 3244 (51%) intake locations showed a perfect name agreement between the CWS source name and indexed NHD reach name while another 729 (11.46%) locations showed a perfect name agreement between the CWS source name and the associated river name. Other details of the name-matching model are as provided in Table 5.

Table 5. Relative distributions of the CWS intakes georeferenced to the closest and the second closest NHD reach.

Name-match Score	CWS Intake georeferenced to		Total
	Nearest NHD reach	2nd nearest NHD reach	
1	2816 *(55.39%)*	428 *(33.52%)*	3244 *(51.0%)*
1A	206 *(4.05%)*	523 *(40.96%)*	729 *(11.46%)*
2	128 *(2.52%)*	13 *(1.02%)*	141 *(2.22%)*
3	329 *(6.47%)*	53 *(4.15%)*	382 *(6.01%)*
4	0 *(0.00%)*	0 *(0.00%)*	0 *(0.00%)*
5	1257 *(24.72%)*	232 *(18.17%)*	1489 *(23.41%)*
6	29 *(0.57%)*	3 *(0.23%)*	32 *(0.50%)*

7	1 *(0.02%)*	1 *(0.08%)*	2 *(0.03%)*
8	2 *(0.04%)*	0 *(0.00%)*	2 *(0.03%)*
9	57 *(1.12%)*	3 *(0.23%)*	60 *(0.94%)*
10	259 *(5.09%)*	21 *(1.64%)*	280 *(4.40%)*
Total	5084 *(100.00%)*	1277 *(100.00%)*	6361 *(100.00%)*

[Note: The name-match score is the numeric value (x) of the name flag (xN) described in the algorithm. The scores 1 and 1A represent the name-match flags for the CWS intake source name and associated river name respectively. The absence of any score (4N) is conspicuous and can be explained by the presence of a CWS source name for all intakes (thus the condition to satisfy a 4N flag never occurs).]

The results of the distance analysis part of the algorithm indicate that 5166 (81.21%) of the CWS intake locations were within the distance of 6 arc seconds of the NHD reach they were indexed to. The rest of the intakes (1195 or 18.79%) were farther than 6 arc seconds from the NHD reach they were indexed to.

2.4 Verification and Validation

The verification of algorithm output is an important step in the design of geographic applications in order to preserve the quality of stored data. Enforcing the attribute accuracy of geographic information systems should be a primary goal. However, statistically significant analysis of error often can only be achieved through manual cross-validation of data points.

To facilitate the manual verification and validation process, which is time consuming and laborious, GIS-based tool was developed. The tool allowed the user to sequentially run through a list of selected intake locations. For each intake under consideration, associated NHD data, GNIS (Geographic Name Information System) data layers, United States Geological Survey (USGS) digital topographical maps (1:24K DRG), and the attribute data for the intake were automatically overlaid for visual inspection. The following scheme was designed to record the results of the verification and validation of each point:

1. **Keep Best**: Keep the model predicted reach as the most logical reach for georeferencing the intake to.
2. **Use Alternate**: Select the alternate reach (second best prediction by the model) as the most logical reach for georeferencing the intake to.
3. **Pick New Line**: Select an entirely different linear reach as the most logical reach for georeferencing the intake to.
4. **Pick New Polygon**: Select an entirely different polygon reach as the most logical reach for georeferencing the intake to.
5. **Bad CWS Point**: Mark the intake as a problem point because of unacceptable quality of NHD and/or intake location data (as indicated with reference to the DLG).

Numbers (in parenthesis) associated with each item indicate the numeric QA/QC flag assigned to the point indicating the verification result. For all of the above choices, the user has the ability to record a comment that is linked to the intake point in the attribute table.

To develop a representative sample of all CWS points (N= 6361) by choosing a desired confidence level and a desired confidence interval for the total number of CWS points, the optimal number of points needed to be verified was determined by the following relationship (Equation 1):

$$SampleSize = \frac{Z^2 * p * (1 - p)}{c^2} \qquad \text{(Equation 1)}$$

Where:

Z = the value corresponding to the desired confidence level (1.96 for 95% confidence)

p = number of total points, and

c = confidence interval (expressed as a decimal; 0.02=±2%).

Using this relationship, it was determined that at a 95±2% confidence, for the total number of CWS points, a representative sample would include at least 1743 points. Thus by verifying 1933 out of 6361 points (30.40%), the confidence was slightly higher than 95±2%. Confidence intervals for each flag category were determined individually (all are at a 95% confidence level) (Table 6).

Table 6. General statistics of the QA/QC analysis of the CWS intake locations.

General statistics of the QA/QC analysis of the CWS intake locations

Flag	QA'd	Total Points	Percentage	Confidence
10N	280	280	100.00%	100%
9N	60	60	100.00%	100%
8N	2	2	100.00%	100%
7N	2	2	100.00%	100%
6N	32	32	100.00%	100%
5N	929	1489	62.39%	95± 2.0%
4N	0	0	N/A	N/A
3N	302	382	79.06%	95± 2.6%
2N	112	141	79.43%	95± 4.3%
1A	120	729	16.46%	95± 8.2%
1N	95	3244	2.93%	95± 9.9%

Total	**1933**	**6361**	30.40%	95±2.0%

Overall Confidence Level 95% and Confidence Interval ± 2% for all intake locations verified (N =6361; n =1933).

All points flagged "6N" or greater were examined (376). To obtain an equal confidence level for 5N points (N=1489), 919 points were reviewed. Approximately one percent of 1N and 1A flags were checked (26 of 3244 and 4 of 729 respectively). Approximately eighty percent of the total 2N and 3N points were examined (414 of 523) and approximately 70% of those points were flagged 3N (302) and approximately 30% were flagged 2N (112). The number of 3N points to check was calculated first (60% of total) and rounded to the nearest integer, and then the number of 2N points was determined by subtracting the 3N value from the total number of 2N and 3N points desired.

Analysis of the verification and validation results (Table 7) show that majority (71.4%) of the model predicted georeferencing was correct (as indicated by QA flag 1 where the best flag assignment was considered correct). Only 17.5% of the verified intake locations were found to be problematic because of unacceptable quality of NHD and/or intake location data (as indicated with reference to the DLG). In 6.8% cases, the alternate reach (second best prediction by the model) was found to be the most logical reach for georeferencing the intake to. About 1% of the verified intake locations were manually indexed to a new NHD linear reach and about 3.3% of the verified intake locations were manually indexed to a new NHD polygon reach.

Table 7. Verification and validation results of the analysis for the intake locations.

Flag	**Keep Best (1)**		**Use Alternate (2)**		**Pick New Line (3)**		**Pick New Polygon (4)**		**Bad CWS Point (5)**		**Total QA'd**
10N	135	48.21%	15	5.36%	5	1.79%	18	6.43%	107	38.21%	**280**
9N	50	83.33%	2	3.33%	2	3.33%	1	1.67%	5	8.33%	**60**
8N	2	100.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	**2**
7N	1	50.00%	0	0.00%	0	0.00%	1	50.00%	0	0.00%	**2**
6N	23	71.88%	1	3.13%	1	3.13%	0	0.00%	7	21.88%	**32**
5N	674	72.55%	57	6.14%	8	0.86%	30	3.23%	160	17.22%	**929**
4N	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	**0**
3N	248	82.12%	9	2.98%	3	0.99%	7	2.32%	35	11.59%	**302**
2N	95	85.59%	1	0.90%	1	0.90%	4	3.60%	10	9.01%	**111**
1A	95	79.17%	10	8.33%	1	0.83%	2	1.67%	12	10.00%	**120**
1N	57	60.00%	36	37.89%	0	0.00%	0	0.00%	2	2.11%	**95**
Total	**1380**	**71.39%**	**131**	**6.78%**	**21**	**1.09%**	**63**	**3.26%**	**338**	**17.49%**	**1933**

3. Temporal misalignment: The case of population distribution

3.1 Background

High resolution population distribution data are essential for successfully addressing critical issues ranging from socio-environmental research to public health to homeland security (Dobson et al. 2000; Bhaduri et al. 2002, 2005, 2007; Chen 2002; Hay et al. 2005; Sutton et al. 2001). Commonly available population data from Census is constrained both in space and time. From a spatial perspective, Census data is limited by Census accounting units, such as blocks, and there is often great uncertainty about the spatial distribution of residents within those accounting units. From a temporal perspective, detailed Census information is only available on a decadal scale since the census was originally designed for medium to long-term solutions for social and economic planning activities over a number of years (U.S. Census Bureau, 2000). However, the pressing need for finer temporal resolution population distribution data for risk and consequence assessment of disasters, prompted the development of population distribution data at temporal scales of nighttime and daytime (Bhaduri et al. 2006, McPherson and Brown 2004, McPherson et al. 2006).

Dasymetric modeling is one of the most well recognized and popular spatial modeling methods for disaggregating Census data. In dasymetric mapping, ancillary spatial data at a finer spatial resolution is utilized to augment the spatial interpolation process, and the variability and spatial discontinuity in their values enable an asymmetric and discontinuous allocation of population (Wright 1936; Langford and Unwin, 1994; Eicher and Brewer, 2001; Mennis 2003). Land cover/land use is the best example in this respect (Monmonier and Schnell, 1984; Reibel and Agrawal, 2006) where different land cover or land use categories for each cell can be used as a weighting function for population distribution such as urban areas which will have a higher weight than forested areas. LandScan USA is a high resolution (3 arc seconds or approximately 90m cells) population distribution model for the 50 U.S. states developed at Oak Ridge National Laboratory (Bhaduri et al. 2007). This model employs the principles of dasymetric mapping specifically with the National Land Cover Database (NLCD) as one of the key ancillary spatial data. For NLCD 1992 and NLCD 2001, the years denote the year of the satellite imagery used to derive the databases and not the data distribution date which was several years later. Given the assumed relationship between land cover data and population, ideally both databases need to be from the same time for the distribution model to produce optimal results. However, the process of revising LandScan USA database with yearly updates from Census and older (and static) land cover data accentuates the challenge of temporal misalignment. It is critical that the land cover data is updated and synchronized with the Census data year; otherwise the algorithm constraints will dictate that the increased population be accommodated by the existing populated cells. The following section describes how secondary databases are used to update the land cover information in an attempt to temporally align it with the population data.

3.2 Technical Approach

Census population data serve as the foundation for the LandScan USA model. The model is resolved to each census block, with the goal being to maintain the integrity of the Census Bureau data at the block level for actual Census years. An iterative methodology is used to characterize each census block using the NLCD 2001 to estimate the ratio between developed

pixels and population counts. For each census block, a histogram of land cover types is built using the corresponding NLCD extent. Using this information, each census block is assigned a model type that informs the algorithm how population should be allocated within that block.

The easily observable links to anthropogenic activity makes land cover a crucial data layer for modeling population distribution. However, using NLCD 2001 as a baseline input into the LandScan population distribution model presents immediate temporal issues:

1. NLCD 2001 represents land cover derived from satellite imagery taken in 2001 but was not processed and released by USGS until early 2007.

2. Annual county level Census estimates after 2001 were no longer temporally consistent with the land cover layer.

3. The NLCD cannot be utilized for characterization of growth areas over time.

In subsequent out-years from the 2000 U.S. Census, population estimates reported at the County level are decomposed to the spatial resolution required for LandScan USA. The U.S. Bureau of Census releases annual intercensal population estimates for the nation, states, counties, incorporated places, and minor civil divisions, but does not report disparate growth patterns at the local or census block level. In the absence of a comprehensive national census performed more frequently than on a decennial basis, other methods must be used to answer important questions. Where is the growth taking place geographically? What is the magnitude of growth in a particular place? When is the growth occurring?

Ideally, land cover change analysis would provide a better indicator of growth magnitude as well as improved spatial precision of growth areas. However, the lag time of national land cover datasets, as well as the different classification schemes used for successive products, makes a comparative analysis of land cover unreliable for determining where population should be distributed.

In the absence of current land cover data, other methods to allocate intercensal population estimates and improve the overall precision of all population distributions must be used. While simply prorating population growth across a county may be expedient, easy to calculate, and impervious to spatial data anomalies, this method also has the distinct disadvantage of discounting non-uniform growth patterns over time. Spatial data change analysis through historical census analysis or identification of road network changes provides a consistent and scalable method that avoids uniform distribution of growth, but these methods also have drawbacks. For example, while new roads are a good indicator of development, these roads may be built long before homes, businesses or other structures for human occupancy are built.

Secondary Data for Improving Population Distribution

Parcel Data

Acquiring parcel address data from commercial providers affords frequent (quarterly) updates and is therefore an excellent candidate for replacing outdated land cover in some areas. In the initial implementation of this method, over 49 million parcel address points were acquired for 536 counties. This accounted for only 17% of the total U.S. counties, yet represented 63% of the total U.S. population.

The parcel address points in a census block are compared to the number of housing units reported in the 2000 census as well as the number of workers assigned to the block based on the daytime population distribution algorithms. The comparison resulted in a delineation of residential census blocks indicative of "new development". Taking this forward to the residential population distribution model, these blocks receive intercensal population determined by the number of address points and the average population per household for that county. The remaining intercensal population growth is distributed to existing residential blocks. This methodology produced markedly improved spatial precision of the population distribution especially in new suburban developments and rural census blocks. For example, land cover may indicate undeveloped areas such as forest or agriculture, but correctly georeferenced parcel address point data refines the population distribution within a block.

LiDAR and Derived Products

The more recent availability of Light Detection and Ranging (LiDAR) data provided another opportunity to refine and update land cover to capture not only residential growth areas (Figure 4) but also temporal occupancy variations. The original LiDAR dataset used for this purpose provided partial coverage at 1 meter resolution for 229 counties including many major metropolitan areas. Since then additional areal coverage has been available as new data is captured or existing data is updated. Included with the reflective surface, last return, bare earth, and intensity data, very detailed building extraction features were available that included attributes of area, height, roof pitch. Using these physical building characteristics, each building feature is assigned a general land use type within the LandScan USA model. Census block type characterization is accomplished by comparing the building footprints within each census block to the number of housing units reported in the 2000 census as well as analyzing building types and the number of workers assigned to the block based on the daytime population distribution algorithms. Residential buildings indicating "new development" receive intercensal population determined by the number of buildings and the average population per household for that county.

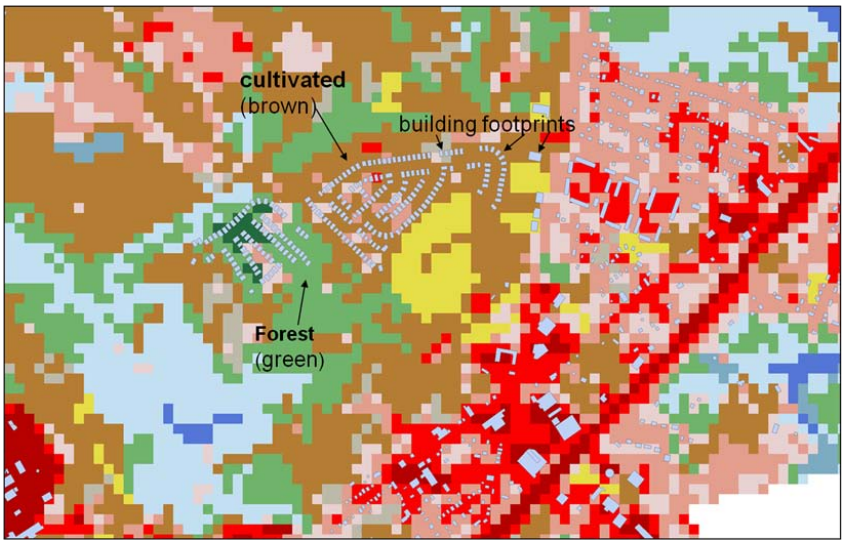
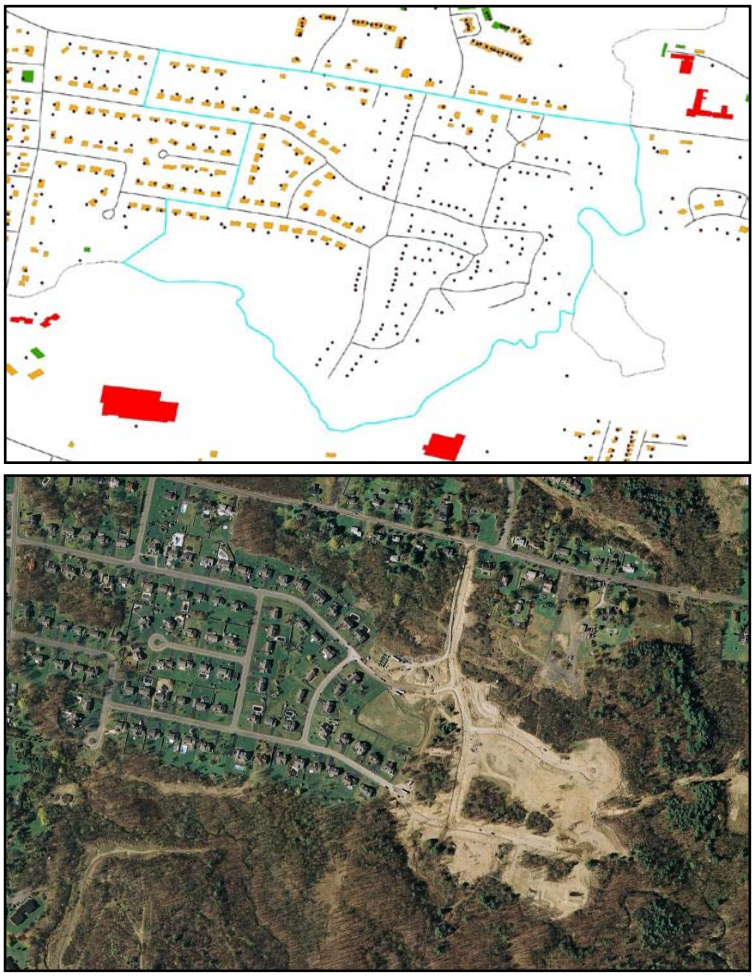


Figure 4. Building footprints highlight temporal misalignment of NLCD 2001 and population.

3.3 Validation and Verification

The addition of each of the aforementioned secondary data sources to the LandScan USA model has resulted in a significant step forward in adjusting for temporal misalignment between the Census annual population estimates and NLCD 2001 data. However, the introduction of this spatial data into the model can also introduce new concerns. For example, Figure 5 shows an instance of a census block where population estimates developed using parcel address data and estimates developed using LiDAR data both indicate new growth. However, the estimates are inconsistent and further validation must be done using high resolution imagery.

Secondary data is used in the LandScan USA model to improve the precision of all population distributions, not only to identify intercensal growth. The target Census date used in the LandScan USA model is known to be misaligned with the NLCD data, but there are also instances where it is unknown whether discrepancies between the land cover type and the census information is due to a temporal issue, inaccurate land cover data, scale factors, or all of the above. These disparities must be accounted for in the model by including adjustments for known error parameters in the database.



Census (2000) :

Population = 78

Housing Units = 25

Parcel Addresses (2007):

Population Estimate = 486

Address Points = 205

LiDAR (2007):

Population Estimate = 215

Residential Units = 64

Figure 5. Population from Census 2000, and estimates using parcel address points from 2007 and LiDAR from 2007 are validated against high resolution imagery.

References

Muller, J.C., Weibel, R., Lagrange, J.P., and Salge, F. Generalization: state of the art and issues. In: J. C. Muller, J. P. Lagrange, and R. Weibel (eds), *GIS and Generalization: methodology and practice*, Taylor & Francis, London, UK. pp. 3-18.

Bhaduri, B. 2008. Population Distribution During the Day. In: Shekhar S. and Hui X (eds), *Encyclopedia of GIS*, Berlin, DE: Springer-Verlag 1377p.

Bhaduri B., E. Bright, P. Coleman, and M. Urban. 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal* 69:103-117.

Bhaduri, B., E. Bright, P. Coleman. 2005. Development of a High Resolution Population Dynamics Model. ((Paper presented at Geocomputation 2005, Ann Arbor, Michigan); http://www.geocomputation.org/2005/Abstracts/Bhaduri.pdf

Bhaduri, B., E. Bright, P. Coleman, and J. Dobson. 2002. LandScan: Locating people is what matters. *Geoinformatics* 5(2): 34-37.

Buttenfield B. P. and R. McMaster (eds.). 1991. *Map Generalization: Making Rules for Knowledge Representation*, London: Langham. 245p.

Chen, K. 2002. An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing* 23(1): 37-48.

Dobson, J., E. Bright, P. Coleman, and B. Bhaduri. 2003. LandScan: a global population database for estimating population at risk. In: Mesev V (ed), *Remotely Sense Cities*, United Kingdom: Taylor and Francis. pp. 267-279.

Focazio, M. J., A. H. Welch, S. A. Watkins, D. R. Helsel, and M. A. Horn. 2000. A Retrospective Analysis on the Occurrence of Arsenic in Ground-Water Resources of the United States and Limitations in Drinking-Water-Supply Characterizations. USGS Water-Resources Investigations Report 99–4279

Goodchild, M. F. 2001. Models of Scale and Scales of Modeling. In: Tate N and Atkinson P (eds), *Modeling Scale in Geographical Information Science.* John Wiley & Sons 277 p.

Goodchild, M., L. Anselin, and U. Deichmann. 1993. A framework for the areal interpolation of socioeconomic data. *Environment and Planning*, 25: 383-397.

Goodchild, M., and N. Lam. 1980. Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing* 1: 297-312.

Hay, S. I., A. M, Noor, A. Nelson, and A.J. Tatem. 2005. The accuracy of human population maps for public health application. *Tropical Medicine and International Health* 10: 1073-86.

McPherson, T., and M. Brown. 2004. Estimating daytime and nighttime population distributions in U.S. cities for emergency response activities. Symposium on Planning, Nowcasting, and Forecasting in the Urban Zone. Paper presented at the *American Meteorological Society Annual Meeting*, Washington, D.C.

McPherson, T.N., J.F. Rush, H. Khalsa, A.Ivey, and M.J. Brown. 2006. A day-night population exchange model for better exposure and consequence management assessments Paper presented at the *6th Annual Meeting of the Urban Environment American Meteorological Society*, Atlanta.

Reibel, M. and A. Agrawal. 2006. Areal interpolation of population counts using pre-classified land cover data. Paper presented at the *2006 Population Association of America Annual Meeting.*

Simley, J.D. and Carswell Jr., W.J. 2009. The National Map—Hydrography: U.S. Geological Survey Fact Sheet 2009-3054, 4 p.

Sutton, P., D. Roberts, C.D. Elvidge, and K. Baugh. 2001. Census from heaven: An estimate of the global human population using night-time satellite imagery, *International Journal of Remote Sensing* 22: 3061-76.

Tobler, W. and H. Moellering 1972. Geographical variances. *Cartography and Geographic Information Systems* 20(3): 96-106.

Tobler, W. 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association.* 74 (367): 519-530.

U.S. Census Bureau, Population Division, Journey to Work and Migration Statistics Branch (2000). *Census 2000* PHC-T-40, Estimated daytime population and employment-residence ratios: Technical notes on the estimated daytime population. Retrieved January 4, 2011, from (http://www.census.gov/population/www/socdemo/daytime/daytimepoptechnotes.html)

Woodstock, C.E. and A.H. Strahler. 1987. The factor scale in remote sensing. *Remote Sensing of Environment* 31: 311-32.

Acknowledgements

The authors would like to acknowledge the ongoing financial support for the development of LandScan and LandScan USA models and databases from the Department of Defense and the Department of Homeland Security. Funding for the drinking water intake study was funded by the Office of Pesticide Programs, U.S. Environmental Protection Agency. Data assistance from the National Geospatial Intelligence Agency and the U.S. Geological Survey are particularly acknowledged. Both studies benefited from significant contributions from some of the best and brightest researchers, whose assistance in data search, acquisition, and validation have been invaluable. Such contributions from Joel Blomquist, Marilee Horn from the USGS, and Tommy Dewald and Todd Dabolt from the USEPA are thankfully acknowledged. We would also like to thank several other members of the Geographic Information Science and Technology group for their periodic insights and assistance with this work. This paper has been authored by employees the U.S. Federal Government and of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.