Quality Assessment Strategies for GIS Reference Data Used in Patient Address Geocoding

By C.A. Klaus and L.E. Carrasco, Ph.D.
North Carolina Central Cancer Registry

June 2012

The Value of GIS Reference Data QC

- Reference data propagate error into geoprocessed cases; have to test to determine amount of error
- It can be done relatively infrequently, and it frees up staff time to focus on QC of address data error from facilities and/or patients
- GIS Reference data error should be measured by source
- QC of parcels and/or address points could be time consuming, but can be leveraged by using methods outlined here

Outline

1. Background: GIS reference data and geoprocessing error
2. Availability of GIS reference data
3. NC CCR GIS reference data QA strategies
4. Method for assessing census enumeration unit (CEU) assignment accuracy
5. Conclusions

Background: How does geoprocessing error impact health analysis?

- **Can affect geocoding results and health study results when unaccounted for**
  - There are impacts in terms of health analysis capability, accessibility of patients to care, local disease rates, cluster statistics, exposure estimates and spatial weights (Jacquez 2012)
- **Geocoding error is often ignored by researchers:**
  - "..recent research initiatives continue to employ geocoded data without regard for how the accuracy can introduce possible inconsistencies or bias into the results." (Goldberg, Wilson, et. al, 2007)

Background: What Role Does GIS Reference Data Play in Geocoding Error?

- **Reference data quality vary (geographically) by data author (counties primarily, and sometimes cities) in terms of positional and attribute accuracy**
  - "… geocoding quality is very much a function of the quality and consistency of local reference data." (Zandbergen, 2008)
  - There are generally significant data quality differences between locally maintained data and data maintained at state level (Zimmerman and Li, 2010; Frizelle, B., K. Evenson et. al, 2009).

GIS Reference Data NC CCR Uses

- NC DOT street centerlines
- Local parcels, address points and centerlines
- ZIP code delineations to ‘seed’ ZIP+4 address validation
- Current county and state boundaries
- Latest census enumeration units (CEU)
- Ortho-imagery for interactive work
- TIGER 2010

GIS Reference Data QA Strategies

- **Goal:**
  - Produce estimates of error for small area analyses, and reduce error where possible
- **Resources:**
  - 2 staff that undertake GIS Reference Data QC, among other duties
  - 98 counties and 6 cities that author GIS reference data, that need positional/attribute accuracy QC for our purposes
- We chose a few QC methods from among many based on time effectiveness and other criteria

QC of CEU Positional Accuracy

- We measure agreement of census enumeration unit (CEU) assignment for a given address, as assigned by different methods:

  1. *Spatial Overlay:* CEU assignment from overlay of TIGER census polygons on address points and parcels

  Vs.

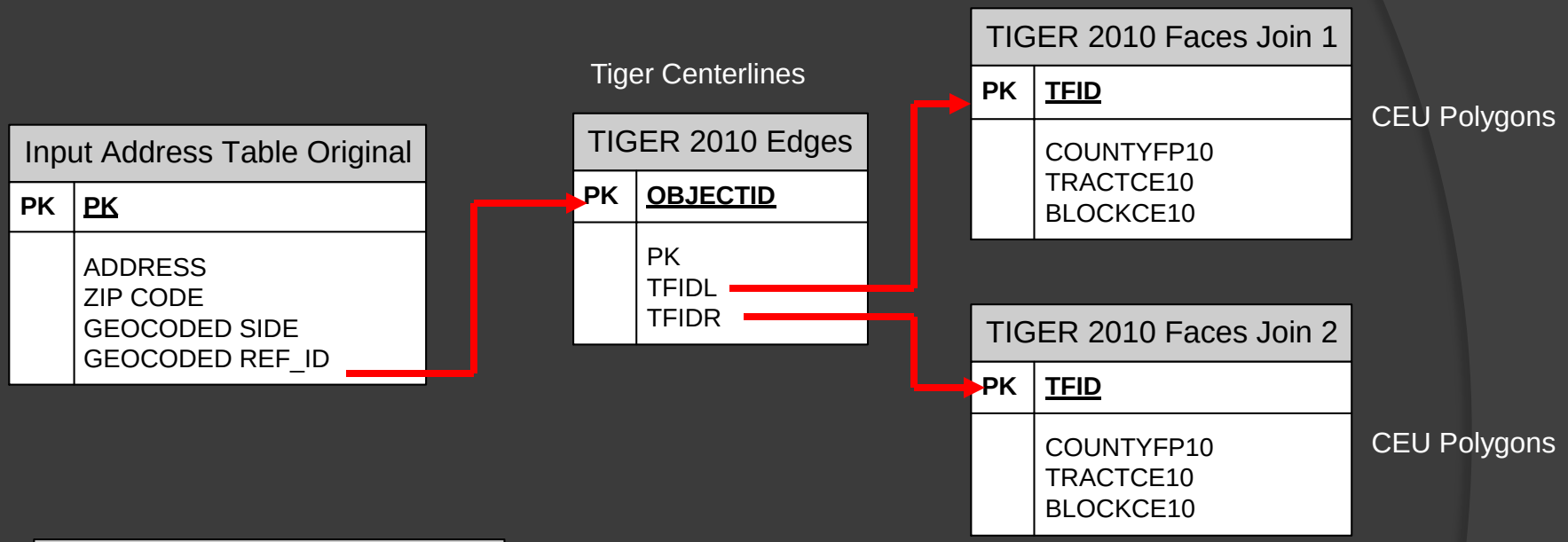
  2. *TIGER Join:* CEU assignment by geocoding address points and parcels against TIGER centerlines (roads).

- CEU positional accuracy depends on quality of locally maintained GIS reference data and Census Bureau linework.

Assigning CEU Via TIGER Table JOIN

- CEUs “should be assigned using a look-up table that links the address to the street segment in the TIGER file that contains the census (enumeration unit) of that street segment.”… CEUs ”should not be based on point-in-polygon procedures” (Rushton et. al, 2006)
- Census Bureau has similar position (US Census Bureau, 2010)

Address Table Joined with TIGER Data



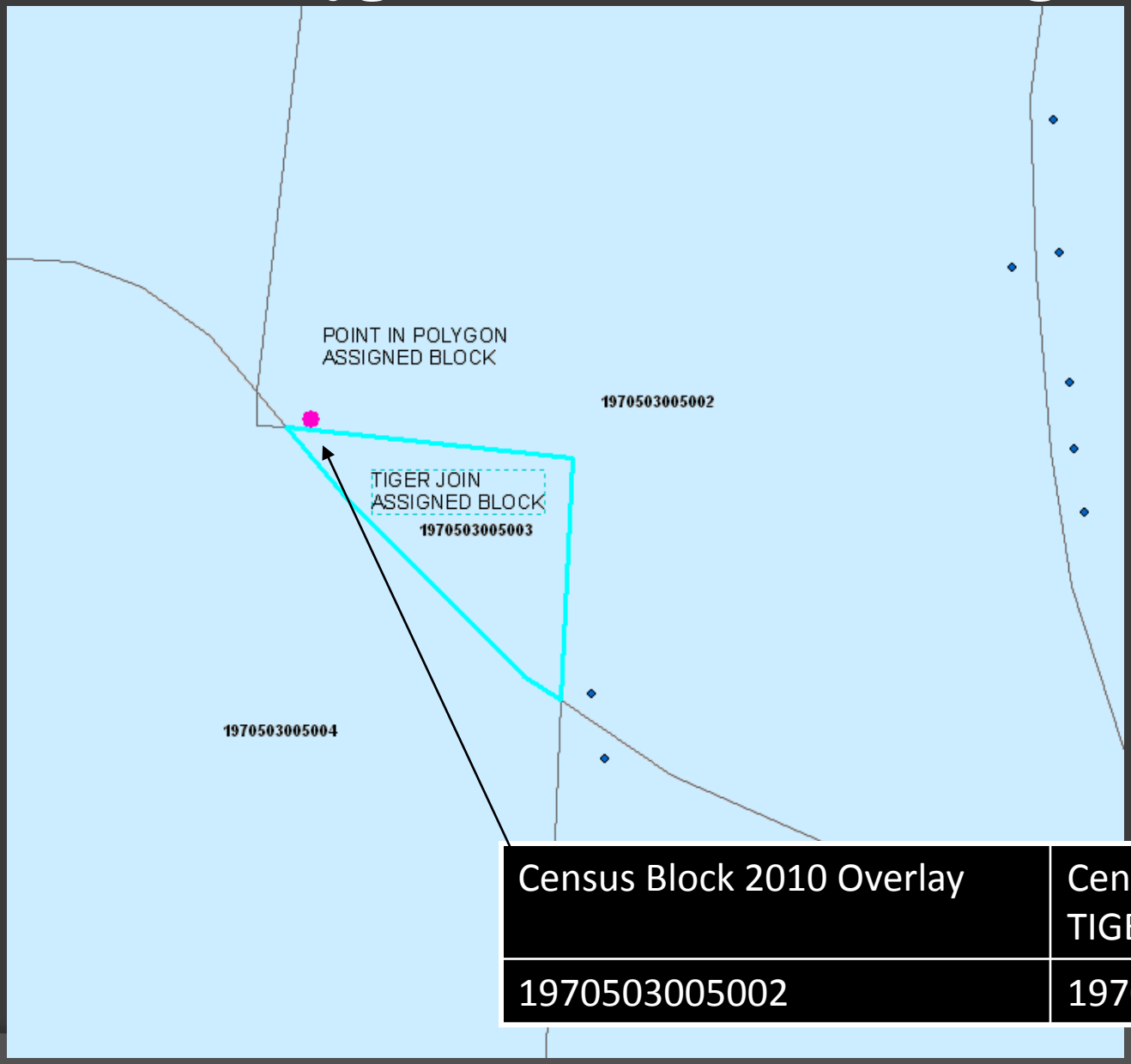
Input Address Table Processed
ADDRESS
ZIP CODE
GEOCODED REF_ID
GEOCODED SIDE
CENSUS BLOCK 2010 OVERLAY
CENSUS BLOCK 2010 TIGER JOIN

Discordance = Records for Which Overlay Census Block <> Join Census Block

HOWEVER: There are valid reasons for CCRs to assign CEUs by **BOTH** join and spatial overlay!

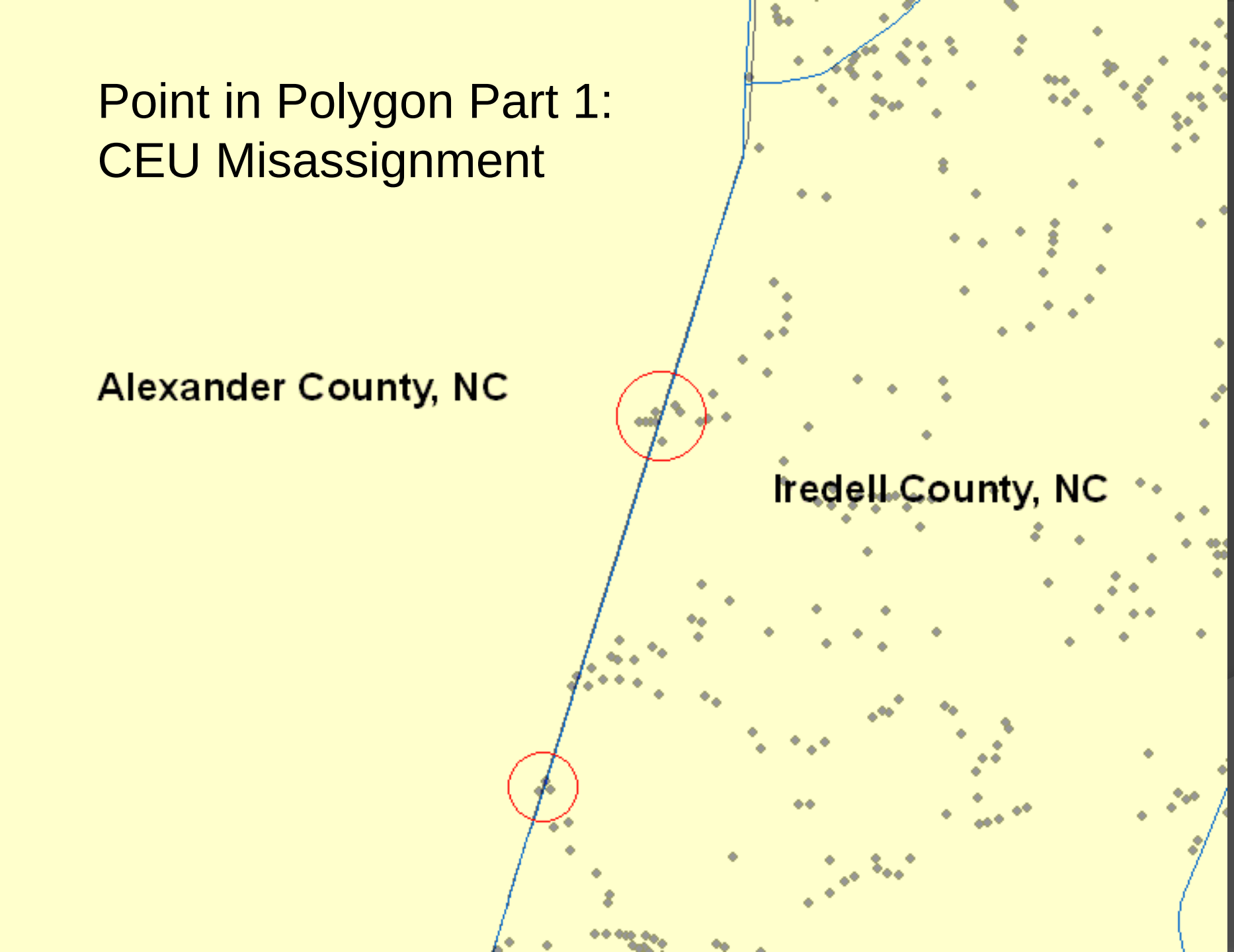
- Not all addresses will batch geocode, or geocode at all, to TIGER centerlines
- Some of these addresses will only geocode to TIGER interactively (time consuming!)

Point in Polygon Part 1: CEU Misassignment



Census Block 2010 Overlay	Census Block 2010 TIGER Table Join
1970503005002	197050300503

Point in Polygon Part 1: CEU Misassignment



CEU Overlay and Join/Overlay Discordance

- Overlay discordance: Two non-TIGER reference data disagree on CEU for a given address/ZIP assigned by polygon overlay
- Join/Overlay discordance: CEU assigned through TIGER table join differs from CEU assigned by point in polygon overlay, for a given address/ZIP
- Rates of discordance vary by CEU type – block, block group, tract, county, and by census year

Point in Polygon Part 2: Imperfect Overlay

- "Imperfect" overlay: a point is assigned more than one CEU
- The record count of overlaid features is greater after overlay (e.g., spatial join)
- Rates of imperfect overlay are very low, ranging from 0.0015% to 0.023% of all address points or parcel centroids by NC region

How do NC GIS Reference Data measure on CEU Join/Overlay Discordance?

- Geocoded these reference data against 2010 TIGER roads
- Measured 2010 block and block group discordance rates by:
  - Reference data type (address point, parcels and centerlines)
  - Author
- Measured level of ZIP+4 validation, geocoding match %, controlling for ZIP+4 validation false positives

2011 Address Point CEU Join/Overlay Discordance Rates with TIGER Edges, for 5 NC Counties

	Address Point ZIP+4 Validation Rate	Address Point % batch geocoded to TIGER*	Address Point % Discord Census Block 2010	Address Point % Discord Census Block Group 2010	1990-2009 NC Cancer Cases Geocoded to 2009 AP, Discord CBG 2010
Range	85.83% - 88.31%	42.88% - 86.6%	0.08% - 0.5%	0.002%-0.31%	0.10%-0.46%

If >1% error then check data, by source, for non random spatial patterns of CEU discordance

* Geocoded with ArcGIS 9.3.1, no ties allowed, minimum match score 100

2011 Parcel Centroid CEU Join/Overlay Discordance Rates with TIGER Edges, for 4 NC Counties*

	Parcel Centroid ZIP+4 Validation Rate May 2012	Parcel Centroid % batch matched to TIGER 2010	Parcel Centroid % Discord Census Block 2010	Parcel Centroid % Discord Census Block Group 2010	1990-2009 NC Cancer Cases Geocoded to 2009 PC, Discord CBG 2010
Range	37.9% - 67.94%	28.3% - 55.77%	0% - 0.52%	0% - 0.19%	0.24% - 12.26%

If >1% error then check data, by source, for non random spatial patterns of CEU discordance

* Includes only counties which have no address points, only parcels or CL

NC CCR Cancer Cases, 1990-2009, Geocoded Against 2011 Non-TIGER Centerlines: Street Centerline CEU Join/Overlay Discordance Measures for 4 NC Counties

	# of NC CCR cases that batch matched to CL, grouped by address/ZIP	# of NC CCR cases that batch matched to CL, that also matched to TIGER	Join/ Overlay Discord CB 2010	Join/ Overlay Discord CBG 2010	CBG2010 Discord from vendor maintained centerlines	CBG 2010 Discord from NC state maintained centerlines
Range	2,132 – 14,552	1,801 – 12,216	1.1% - 5.05%	0.33% - 0.94%	3.77% - 9.55%	1.53% - 4.05%

Summary: CBG 2010 Join/Overlay Discordance for NC CCR Cancer Case Addresses, 1990-2009, by Data Authorship Type

Type	Average	Standard Deviation	Range	# of cases used in sample	# of counties used in sample
Locally Authored Parcels	0.72%	0.75%	0.2% - 2.06%	11,221	6
Locally Authored Address Points	1.03%	1.98%	0.07 - 12.26%	376,066	84
Locally Authored Centerlines	0.61%	0.27%	0.33 - 0.72%	40,036	4
State Maintained Centerlines	3.26%	3.18%	0.63 - 19.63%	482,412	99
Vendor Maintained Centerlines	7.31%	6.93%	1.31 – 41.17%	332,860	100

Further CEU join/overlay observations

- County CEU join/overlay discordance: We found ~2000 address points (of ~4.6 million) across the state that were assigned incorrect county based on spatial overlay
- Address Points: in 18 NC counties the number of address points was reduced from 2009 to 2011, often for quality reasons
- Census linework – less positionally accurate in 1990 than in 2000 and 2010, with a big impact on CEU assignment through spatial overlay
- CEU join/discordance decreases with scale. Thus, rates of CEU join/overlay discordance are greatest for blocks, and least for counties.

Conclusions

- The value of GIS reference data QC:
  A. It accounts for some error in geoprocessing
  B. It can be done relatively infrequently, and it frees up staff time to focus on QC of address data error from facilities and/or patients
- GIS Reference data error should be measured on an authorship basis
- Time for QC/value add to parcels and/or address points is not insignificant, but can be leveraged by using methods outlined here

Acknowledgments

- Thanks to Bob Borchers of WI CCR for his advice and suggestions on TIGER table join
- Thanks to Charles Rudder for his help processing data for this study
- The North Carolina Central Cancer Registry acknowledges the Centers for Disease Control and Prevention for its financial support under cooperative agreement NC U58 DP000832-05.
- The content of this presentation is solely the responsibility of the authors and does not necessarily represent the official views of the Centers for Disease Control and Prevention

Questions or Comments

Contact Information:

Christian.Klaus@dhhs.nc.gov

Luis.Carrasco@dhhs.nc.gov

North Carolina Central Cancer Registry

References

- Frizzelle, B., K. Evenson, et al. (2009). "The importance of accurate road data for spatial applications in public health: customizing a road network." International Journal of Health Geographics **8**(1): 24.
- Goldberg, D.W. and Wilson, J.P. et. al, 2008. An effective and efficient approach for manually improving geocoded data. International Journal of Health Geographics 7(60)
- Jacquez, G.M., 2012. A Research Agenda: Does Geocoding Positional Error Matter in Health GIS studies? Spatial and Spatio-Temporal Epidemiology 3(1)
- Rushton, G., M. Armstrong, et al. (2006). Geocoding in cancer research -- A review. American Journal of Preventive Medicine 30(2): S16-S24.
- US Dept of Commerce Geography Division of US Census Bureau. 2010 Census TIGER/Line Shapefiles Technical Documentation, 2011.
- US Bureau of the Census. Cartographic Boundary Files. Washington, DC: US Bureau of the Census, 2012; Available at www.census.gov/geo/www/cob/scale.html. Accessed February 16, 2012
- Zandbergen, PA. A comparison of address point, parcel and street geocoding techniques. Computers, Environment and Urban Systems 2008; 32:214-232.
- Zimmerman, D. and J. Li (2010). "The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies." International Journal of Health Geographics 9(1): 10.