Field Validation of Secondary Data Sources on Parks in the U.S.

Christopher M. Quinn[1], Kevin P. Gibbs[1], Sandy J. Slater[1], Dianne C. Barker[2]

[1] *Institute for Health Research and Policy, University of Illinois at Chicago, Chicago, IL*

[2] *Public Health Institute, Oakland, CA*

UIC INSTITUTE FOR HEALTH RESEARCH AND POLICY

bridging the gap
Research Informing Policies & Practices for Healthy Youth

INTRODUCTION

- Secondary data is frequently used to quantify and characterize access to opportunities for outdoor recreation in settings such as public parks. However, the validity of these data is often unknown.
- The purpose of this study was to assess validity of common secondary data sources on public parks overall and by urbanization.

METHODS

Field Observations:

1. A stratified random sample of census tracts in the Chicago CBSA (n=150) plus a 50 mile buffer (n=120; 50 suburban, 70 rural) was drawn. The study area included 4 states: IL, IN, WI, and MI.
2. Field observers drove every street in each tract to identify public parks in 2009, marking approximate point locations on paper maps.

Table 1. Secondary Source Feature Classes Validated

Source	Feature Classes
Tele Atlas StreetMap Premium 9.0, 2007	D83 National Park Service land D85 State or local park or forest D89 Local Park or Recreation Area
USGS GNIS 2009	Beach Forest Park Woods
Navteq Land Use A Discover Americas Region 5 Q1 2009	Beach Park (City/County) Park (State) Park/Monument (National)
Navteq Park & Recreation POI Discover Americas Region 5 Q1 2009	4493 Marina 7947 Park/ Recreation Area 9517 Campground

Validation:

1. Field observation points were digitized using ArcMap 9.3.
2. Secondary data from **Navteq** (Land Use A, Park & Rec Points of Interest), **USGS GNIS**, and **Tele Atlas** were compiled and joined to census tracts.
3. Field observations were reviewed for eligibility and web research/telephone follow-up was conducted if necessary to determine land management/ownership.
   - Golf courses, schools, fairgrounds, church property, private residential/HOA parks, and duplicates (i.e., multiple observations of the same park from different streets) were excluded.
4. Field observations were matched to the secondary data by name and location.

Statistical Analysis:

1. Agreement statistics and standard errors were calculated for each secondary data source for all observations and stratified by urbanization:
   - **Sensitivity**: proportion of field observations matched to secondary data
   - **Positive Predictive Value (PPV)**: proportion of records in the secondary data observed on the ground in sampled tracts
   - **Concordance**: proportion of park observations matched to the secondary data among all observations either observed in the field or listed in the secondary source

PRELIMINARY RESULTS

- 1018 field observations were made in 233 census tracts. After follow-up research, **788** observations in **223** census tracts were considered eligible public parks.
- Overall, **2.7** eligible parks were observed per tract (range 0 – 13). Suburban tracts (outside the CBSA) had the most parks (mean 3.92, 95% CI 3.17-4.67) compared to:
  - urban tracts (mean 2.63, 95% CI 2.19-3.06), and
  - rural tracts (mean 2.83, 95% CI 2.33-3.32)

Table 2. Sensitivity of Secondary Source Lists by Urbanization

Data Source	All Parks Observed in the field (n=788)	Urban (n=394)	Nonurban (n=394)	p-value
All Sources	.677 (.021)	.698 (.023)	.524 (.025)	<.0001
Tele Atlas	.430 (.022)	.475 (.025)	.109 (.016)	<.0001
USGS GNIS	.526 (.022)	.546 (.025)	.385 (.025)	<.0001
Navteq	.380 (.022)	.391 (.025)	.301 (.023)	.0083

Notes. Standard errors in parentheses. Observations weighted for census tract sampling probability. P-value from chi-square test for difference between list sensitivity estimates by urbanization.

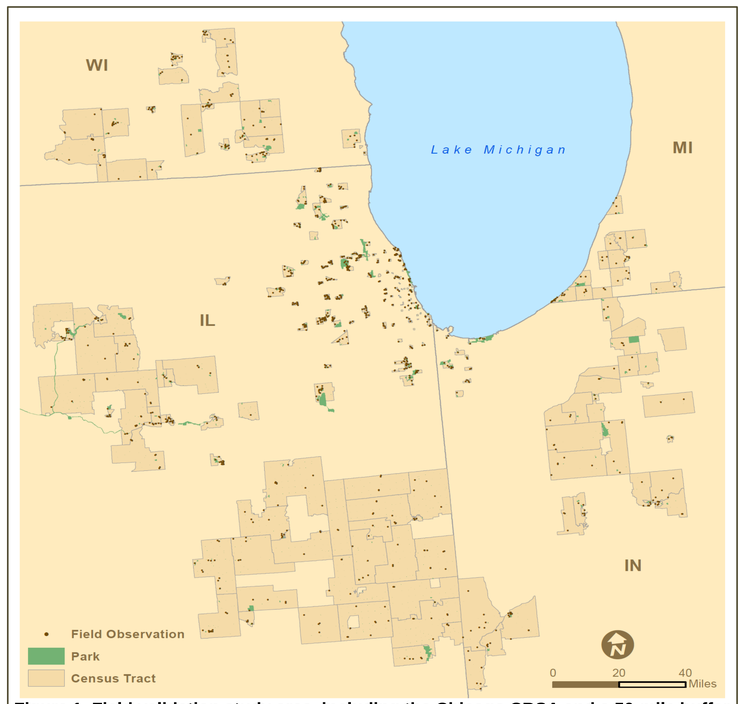

Figure 1. Field validation study area, including the Chicago CBSA and a 50-mile buffer

Table 3. Positive Predictive Value (PPV) and Concordance of Secondary Source Lists by Urbanization

Data Source	All Parks	Urban	Nonurban	p-value
PPV				
Tele Atlas (n=293)	.827 (.024)	.830 (.025)	.741 (.058)	.1203
USGS GNIS (n=446)	.713 (.025)	.741 (.027)	.424 (.037)	<.0001
Navteq (n=669)	.837 (.017)	.859 (.019)	.659 (.027)	<.0001
Concordance				
Tele Atlas (n=846)	.390 (.021)	.428 (.024)	.105 (.015)	<.0001
USGS GNIS (n=858)	.398 (.021)	.427 (.024)	.183 (.019)	<.0001
Navteq (n=1130)	.294 (.018)	.314 (.021)	.185 (.015)	<.0001

Notes. Standard errors in parentheses. Observations weighted for census tract sampling probability. P-value from chi-square test for difference between PPV and concordance estimates by urbanization.

CONCLUSIONS

Overall, available secondary data sources have slight to moderate coverage of public parks that exist on the ground. Agreement improves markedly when data sources are combined.

Significant differences in list coverage were found by urbanization, with parks in nonurban areas less likely to be listed compared to parks in urban areas.

Caution should be taken when using secondary data to identify parks, particularly in nonurban areas. Supplemental data gathering, such as web research and calls to local jurisdictions or park districts, may be necessary.

Studies of access to parks and outdoor recreation areas need to account for systematic secondary data inaccuracies in the absence of on-the-ground data collection.

Acknowledgements and Contact

•We are grateful to the Robert Wood Johnson Foundation for supporting this study.

•Corresponding author: e: cquinn5@uic.edu p: 312-413-9073

•Find out more about our research at **www.bridgingthegapresearch.org**