

Finding parks: Lessons learned from collecting and cleaning parks data in eleven Texas counties

Christine M. Marx¹, Peg Allen², Andrew T. Kaczynski³, Christine M. Hoehner¹

¹Division of Public Health Sciences, Washington University School of Medicine; ²The Brown School of Social Work, Washington University in St. Louis; ³Department of Kinesiology, Kansas State University

Background

Public parks are an important resource for promoting active living for all ages.

Although research on the health benefits of parks has advanced in recent years, data quality and methodological issues are rarely discussed in detail in research publications. The National Recreation and Park Association (2009) reports that “communities are recording similar information at different scales or recording completely different information.”

This lack of standardization for recording and classifying parks is important for researchers, practitioners, community groups, and others who seek to enumerate parks for assessing the availability or quality of parks.

In addition, this information would aid in understanding the degree to which measurement error may bias research results on associations between park access and physical activity and obesity risk.

The continued lack of documentation of methods impedes the advancement of parks research and the utility of these data for practice.

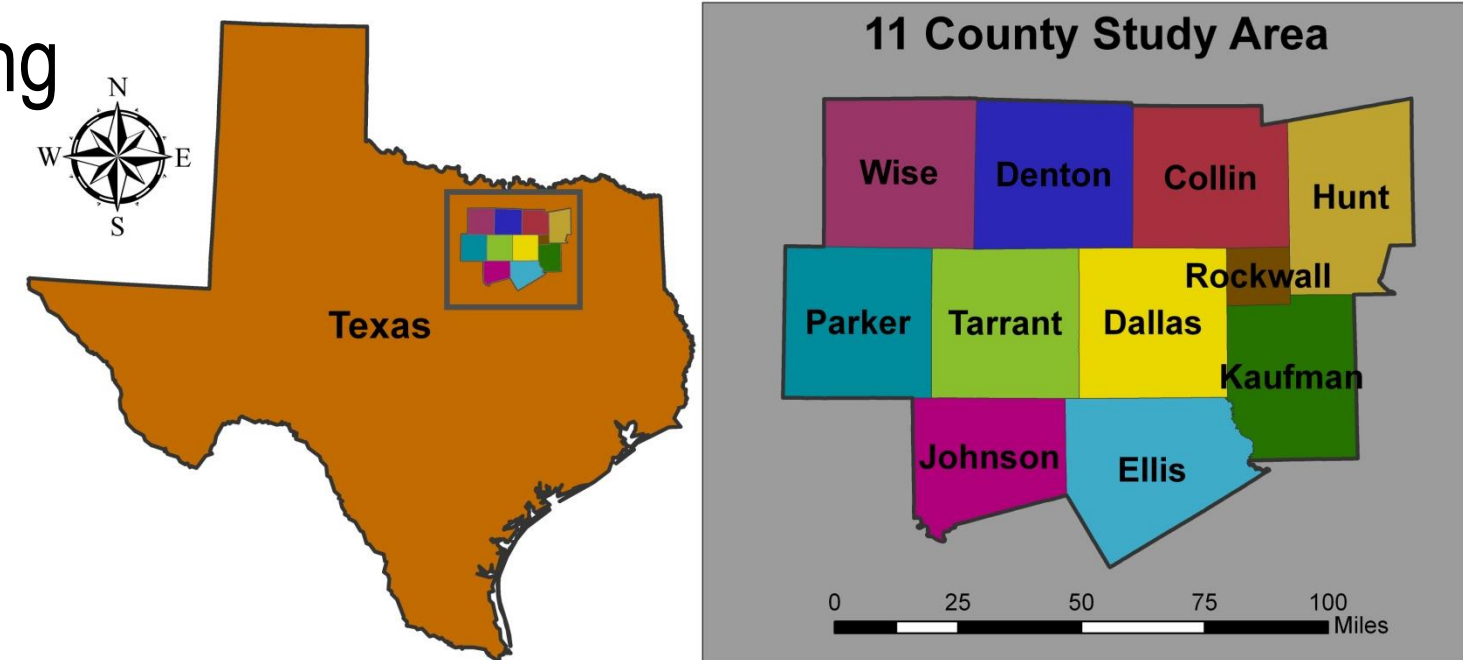
Objectives

- (1) Describe the process of enumerating parks in a large metropolitan region
- (2) Summarize major problems encountered when collecting and cleaning secondary data on park availability

Methods

With approximately 6.5 million people in over 200 municipalities, the Dallas-Fort Worth-Arlington, TX metropolitan statistical area is the fourth largest in the US. The research setting, defined by participants’ locations for a larger study, covers the 11 core counties (approximately 8,700 square miles) of this region.

The metropolitan planning organization (MPO) for the DFW region hosts a GIS data clearinghouse, offering free GIS data to the public.



Methods (continued)

For this study, we obtained a GIS shapefile of parks for 2007, compiled from the multiple municipalities in the area. This was clipped to the 11-county study area, then cleaned, remotely, by:

- identifying overlapping polygons
- identifying irregular or small polygons (<1,000 square feet)
- querying names of parks
- identifying municipalities with unusually broad inclusion criteria
- creating categories for inclusion/exclusion
- using aerial photography & street centerline shapefiles layered with the parks file for reference
- using websites for verification.

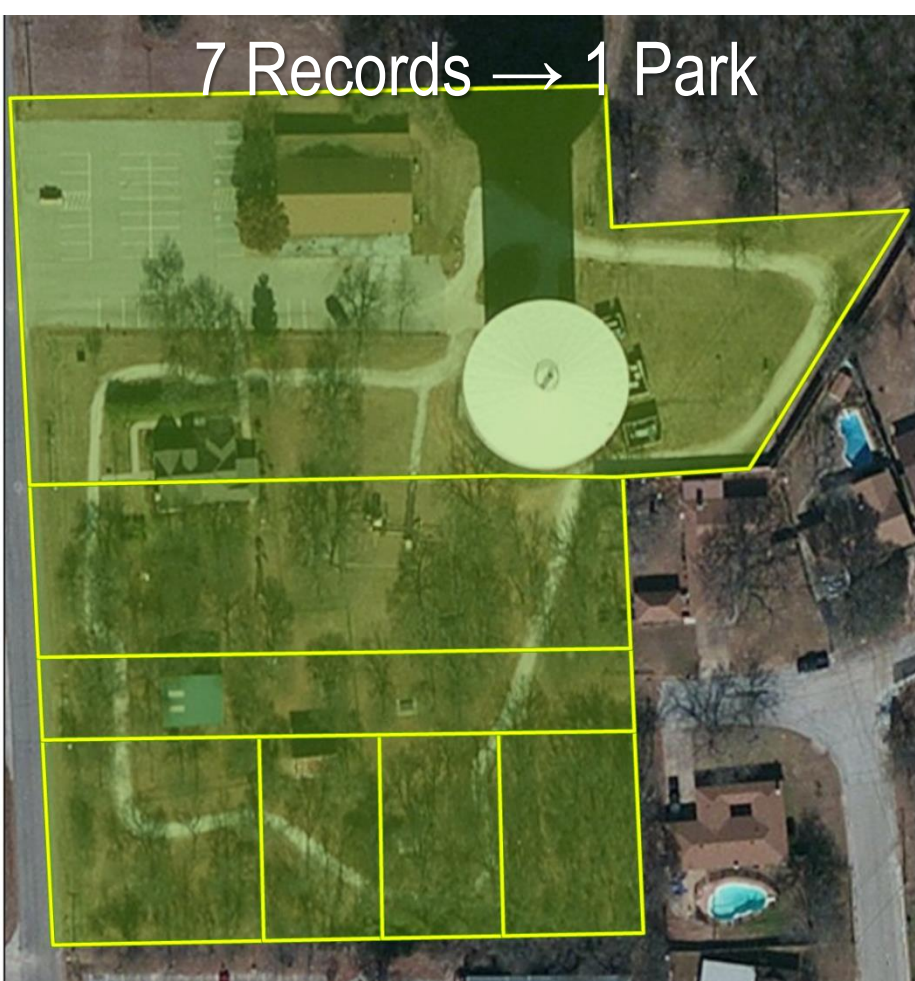
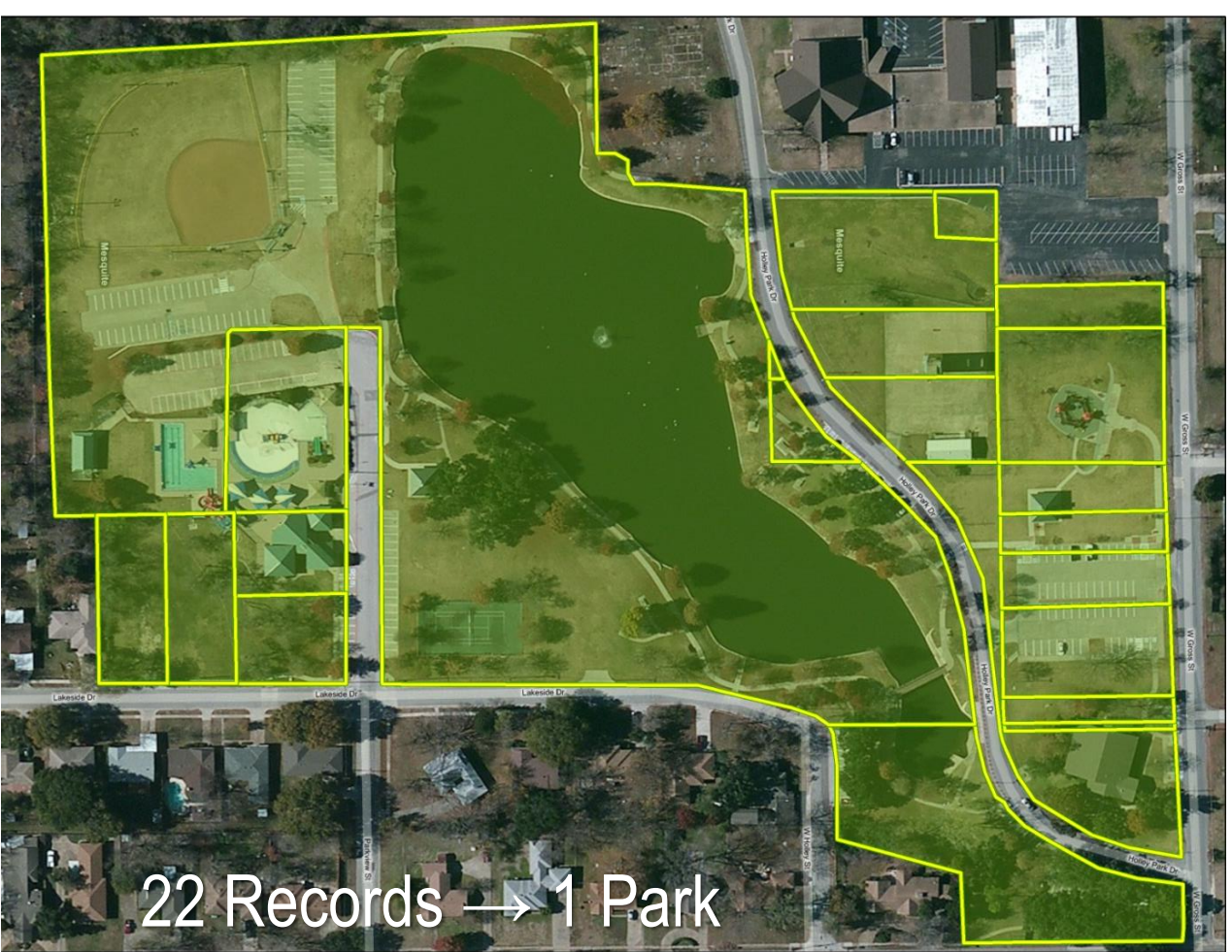
Summary of Results

This process illustrated two common problems associated with use of existing GIS parks data.

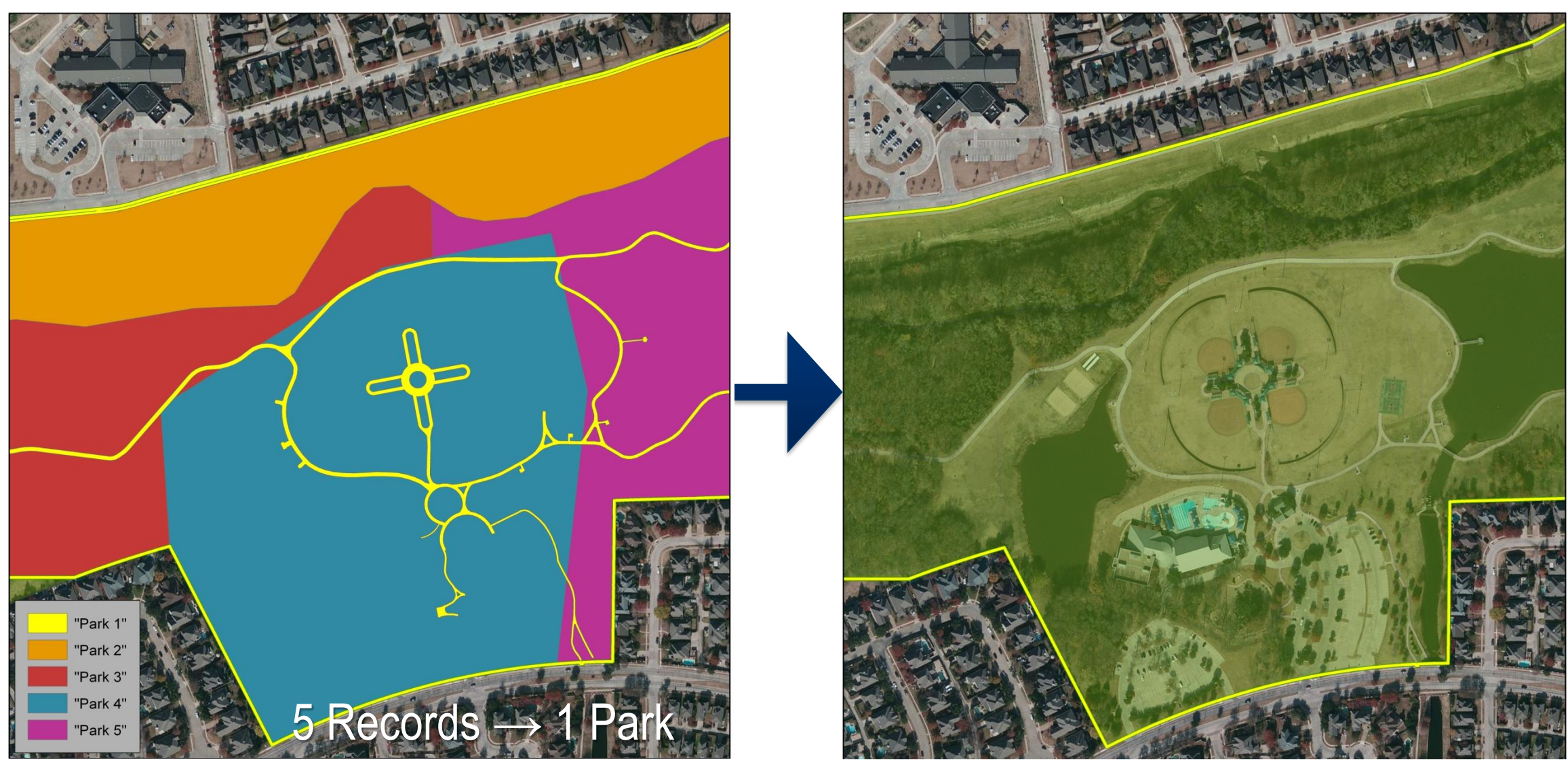
- Despite the availability of secondary data, enumerating parks across large geographic areas can be **extremely time-consuming and labor-intensive** if the data were not verified and standardized across municipalities and counties.
- There is **little consensus on what constitutes a “park.”** The types of land uses included in the parks files from the MPO varied widely across the municipalities from which these data were collected.

Results

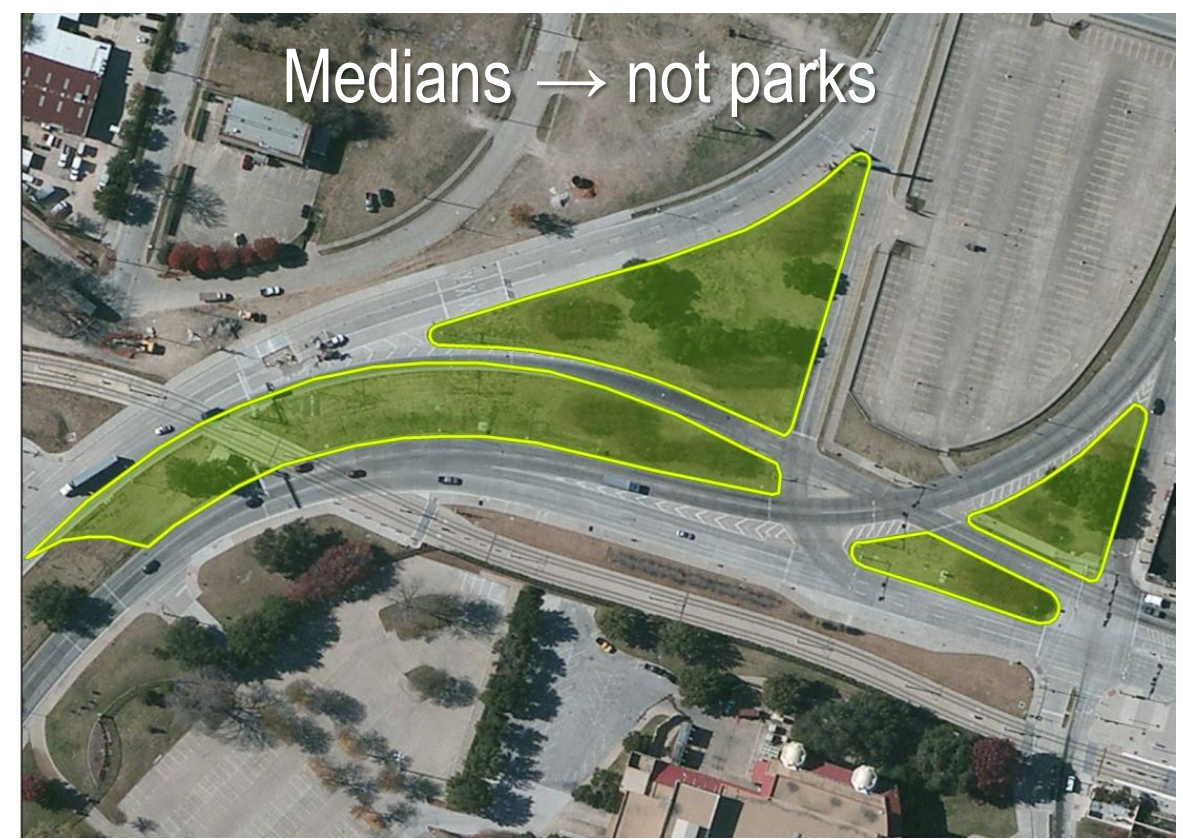
- The starting dataset for the 11 counties had 2,812 records.
- 19 percent of these (n=532) were dissolved into other features because they were multiple records representing a single park.



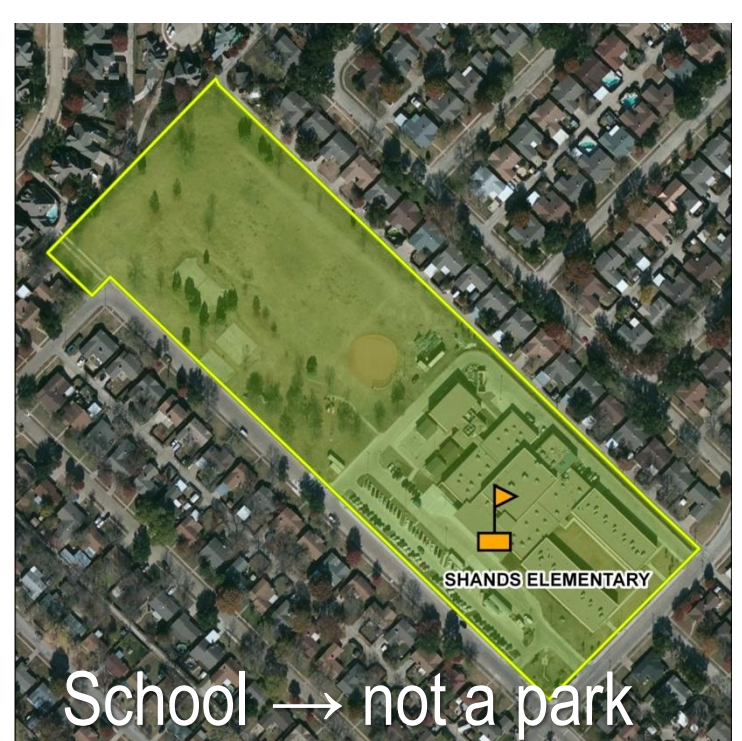
Results (continued)



- 238 features were deleted because they were not parks (e.g., cemeteries, medians, park maintenance headquarters, historical sites, mobile home “parks,” zoos, and country clubs), or because of categorical errors (e.g., future parks, errors in geometry).



- 132 physical-activity related sites (e.g., trails, private facilities, recreation center buildings, schools, and stand-alone tennis and swimming facilities) were deleted because the polygons were not parks and were collected from other data sources.



Results (continued)

- 21 parks that were not found in the 2007 shapefile but were included in an archival version of the file (year 2000) were added to the more recent database.
- 67 parks were digitized and added to the file, based on reviews of municipalities’ park websites.
- The final dataset contained 1,998 records.

Conclusions

Having accurate information on park locations is important for researchers, practitioners, and others who promote active living among youth and adults.

While secondary data are an important data source for mapping park availability, particularly across large geographic areas where primary data collection is impractical, secondary data may not be adequate for research purposes without considerable verification and modification.

In fact, had researchers used the original dataset without extensive cleaning, park availability and opportunities for physical activity would have been substantially overestimated.

For future studies involving parks, researchers should:

- move towards agreement on a definition of a park, in the context of active living research;
- encourage community partners to adopt standard definitions of parks (example: NRPA), not only for research, but also to identify inequities in service and for maintenance of facilities;
- continue to conduct micro-scale research to refine the definition and qualities of parks that most support physical activity.

Finally, researchers using GIS data on parks should not underestimate the potential gap between a working GIS file and a research-ready database, nor the amount of time and effort closing this gap requires.

Support

American Cancer Society Mentored Research Scholar Grant (MRSG-07-016-01-CPPB)

Contact

For more information, contact Christine Marx at marxc@wustl.edu