Information integration: a case study of air quality in Chicago and St. Louis

Jooho Lee & Catherine Blake
University of Illinois at Urbana-Champaign

Background
- Research on Air Quality
  - Toxic emission can cause serious health problems (Spadaro & Rabl, 2004; Chen & Kan, 2008).
  - People with a low-income are exposed to more dangerous air pollutant threats (Grinecki, Rolin, & Boone, 2007; Torres, 2007).
  - Regulatory efforts appeared to reduce emission (Cramer, 1999; 2002; Sicksle & Shadwick, 2007).
- Data literate citizenry (Twidale, Blake, & Gant, 2013)
- Chicago and St. Louis are both participating in the US Department of Energy’s (US DOE) Clean Cities program since 1994 (Chicago, 2013; St. Louis, 2013).

Goal
- Explore the potential usage of public data related to air quality that have been provided by two different government agencies, the US Census Bureau and US EPA.
- Explore the gaps between two public data collected for different purposes.

Method
- Data and Preprocessing
  - Data Sources: AirData, US EPA; US Census; American Community Survey (ACS) 3-year estimates
  - AirData: The average annual Air Quality Index (AQI) values were calculated using the average daily values by county. Air Quality data coverage: 2000-2012
  - Census & ACS: Census coverage: 2000; ACS data coverage: 2007-2012; missing demographics values of 2001 to 2006 were estimated based on a simplified assumption that all values on demographic variables increase or decrease constantly by year.
  - Two datasets are integrated by county and year.
- Variables
  - Six response variables (AirData): Ozone, (2) Particulate Matters in (PM10), (3) Particulate Matters 2.5 (PM2.5), (4) Sulfur dioxide (SO2), (5) Carbon monoxide (CO), and (6) Nitrogen dioxide (NO2)
  - Eleven explanatory variables (Census, ACS) on population, education, and economics: (1) total population, (2) percentage of males, (3) percentage of whites, (4) percentage of blacks, (5) median age, (6) percentage of high school graduates, (7) percentage of bachelor degree and above; (8) median income, (9) income per capita, (10) percentage of people with an income below poverty level, and (11) percentage of home owners who occupy their own house.
- Analysis: Descriptive statistics (linear regression)

Results
- Air trend: Air quality has improved as the US DOE’s Clean Cities program claimed.
- Population: Population size did not correlate with increased air pollution; counties with large black populations are more polluted (see Figure 1).
- Education: Air pollutant measures were lower in areas with a higher education (see Figure 3).
- Economics: The counties with more poor people are more polluted (see Figure 2); higher percentages of home owners who live in their homes are correlated with cleaner air.

Conclusions
- Chicago’s air quality is better than St. Louis. This might indicate better financial support from the Clean Cities program or Chicago is more affluent.
- The access to both the original data and subsequent analysis will be critical to support a data literate citizenry and that tools that currently work well with individual datasets will need to support data integration. Such extensions will be necessary to close the gap between data access and the analysis necessary to answer questions in science, business, and public policy.

Acknowledgments
This poster was made possible in part by a grant from the U.S. Institute of Museum and Library Services (DE-024-2015-014) U.S. Department of Education (SLR-05-12-0054-12 Socio-technical Data Analytics (SODA)).

References

| Table 1 shows overall regression coefficients for each explanatory variable. The highest coefficient was for NOx where the level of p-value of 0.001 or below. |