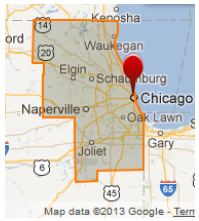
Information integration: a case study of air quality in Chicago and St. Louis

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

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Background

- Research on Air Quality
  - Population density impacts human health (Parrish & Zhu, 2009).
  - 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 (Grineski, Bolin, & Boone, 2007; Torras, 2007).
  - Regulatory efforts appeared to reduce emission (Cramer, 1999; 2002; Sickles & 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).



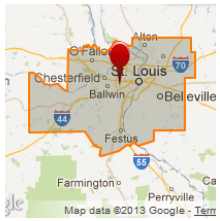


Image sources: http://www.afdc.energy.gov/cleancities/coalition/chicago
http://www.afdc.energy.gov/cleancities/coalition/st-louis

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): (1) Ozone, (2) Particulate Matters 10 (PM10), (3) Particulate Matters 2.5 (PM2.5), (4) Sulfur dioxide (SO2), (5) Carbon monoxide (CO), and (6) nitrogen dioxide (NO2)
  - Eleven exploratory 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)

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**.

Results

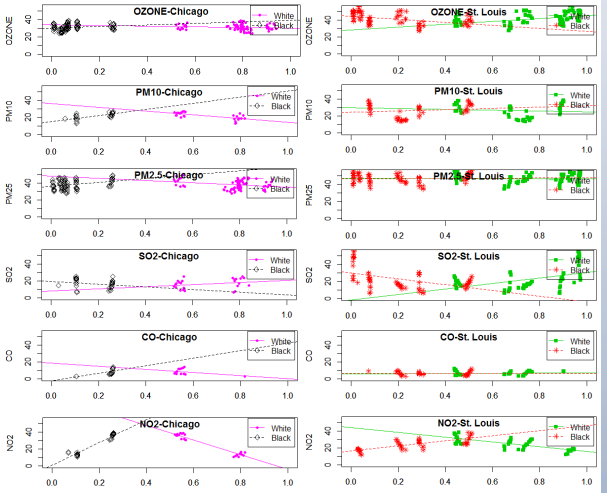


Figure 1. Air trends by percentage of Black and White population

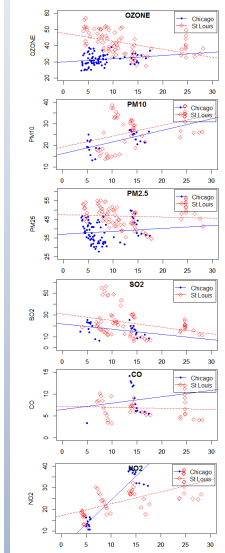


Figure 2. Poverty and air quality

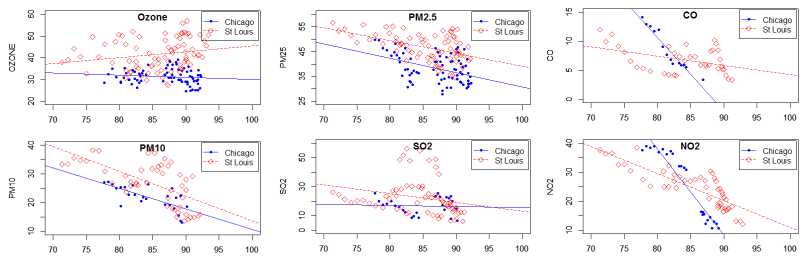


Figure 3. Percentage of the population with a high school diploma versus air quality metrics

- **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** .

	Ozone	PM10	PM2.5	SO2	CO	NO2
R2	0.6801	0.7796	0.8076	0.5511	0.8915	0.9415
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	2.07e-11	< 2.2e-16	< 2.2e-16
Year	***	*			***	
Total population	***	.	***		**	
Percentage of Male					*	**
Percentage of White	***		***			
Percentage of Black	***		***			
Median age						
High school	***	**	***	**	***	
Bachelor or above	***		***		***	
Median income	**	.	*		*	.
Income per capita	***		***			
Below poverty	.		.		*	.
House owner occupy	***	.			*	

Table 1. Summary of significant air quality correlations where ***, **, *, and . correspond to significant levels of 0.001, 0.01, 0.05 and 0.1, respectively

Table 1 shows overall regression coefficients for each explanatory variable. The highest coefficient was for NO_2 where the level of p-value of 0.001 or below.

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 is made possible in part by a grant from the U.S. Institute of Museum and Library Services (IMLS), Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12 Socio-technical Data Analytics (SODA).

References

Chen, B., & Kan, H. (2008). Air pollution and population health: A global challenge. *Environmental Health and Preventive Medicine*, 13(2), 94-101.
Chicago area clean cities coalition. (2013). *Clean cities. U.S. Department of Energy*. Retrieved from http://www. afdc.energy.gov/cleancities/coalition/chicago
Cramer, J. C. (1998). Population growth and air quality in California. *Demography*, 35(1), 45-56. Retrieved from http://www.jstor.org/stable/3004026
Cramer, J. C. (2002). Population growth and local air pollution: methods, models, and results. *Population and Development Review*, 28, Supplement, 22-52. Retrieved from http://www.jstor.org/stable/3115267
Grineski, S., Bolin, B., & Boone, C. (2007). Criteria air pollution and marginalized populations: environmental inequity in metropolitan Phoenix, Arizona. *Social Science Quarterly* (Blackwell Publishing Limited), 88(2), 535-554.
Parrish, D. D. & Zhu. T. (2009). Clean air for megacities. *Science*, 326, 674-675.
Sickles II, J.E. & Shadwick, D.S. (2007). Changes in air quality and atmospheric deposition in the Eastern United States: 1990-2004. *Journal of Geophysical Research*, 112:D17301. doi:10.1029/200
Spadaro, J. V., & Rabl, A. (2004). Pathway analysis for population-total health impacts of toxic metal emissions. *Risk Analysis*, 24(5), 1121-1141.
St. Louis area clean cities coalition. (2013). *Clean cities. U.S. Department of Energy*. Retrieved from http://www. afdc.energy.gov/cleancities/coalition/st-louis
Torras, M. (2005). Income and power inequality as determinants of environmental and health outcomes: some findings. *Social Science Quarterly* (Blackwell Publishing Limited), 861354-1376.
Twidale, M.B., Blake, C. & Gant J. (2013). Towards a data literate citizenry. *iConference 2013 Proceedings*, 247-257. doi: 10.9776/13189

ILLINOIS