Data-Driven Strategies for Lead Poisoning Prevention

Subho Majumdar

University of Minnesota, School of Statistics
Literature seminar talk
October 31, 2014


Joe Brew


Alex Loewi

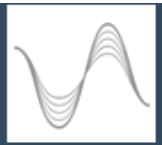

Subhabrata Majumdar


Andrew Reece

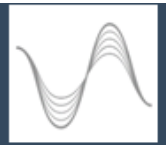

Eric Rozier

- Done as part of the **Data Science For Social Good** (DSSG) fellowship under the Univ. of Chicago.
- Project partners: **Chicago Department of Public Health**.
- **Objective**: Analyze data from CDPH as well as other sources to come up a predictive model for occurence of high blood lead level among children in chicago.
- **Statistical problem**: Predict future lead level of a child living at a certain location of the city by analyzing past city-wide data.

Table of contents



Outline



Background

- From 1995 to 2013, 298,675 Chicago children (29% of those tested) were poisoned by lead (blood lead level [BLL] >5 micrograms per deciliter).
- Although the incidence of lead poisoning has declined drastically (fewer than 3% in 2013), the consequences for those sickened are severe and life-long.
- Lead poisoning is associated with intellectual disability, systemic organ malfunction, aggression, and in severe cases, death.

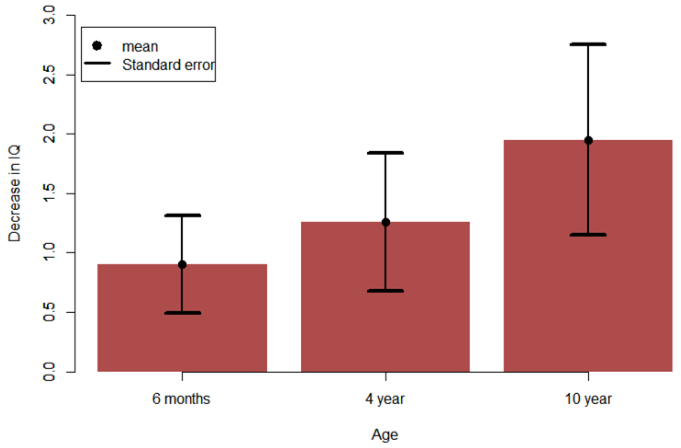
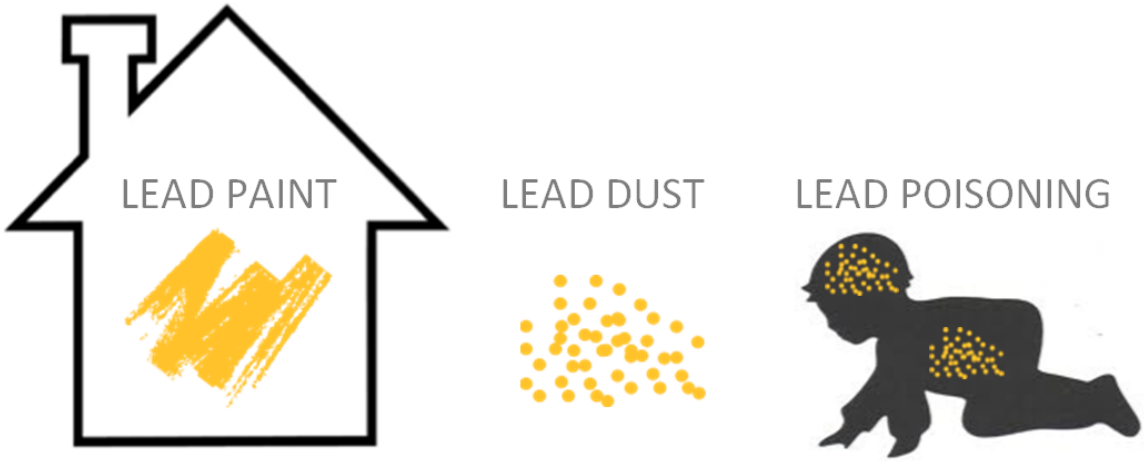


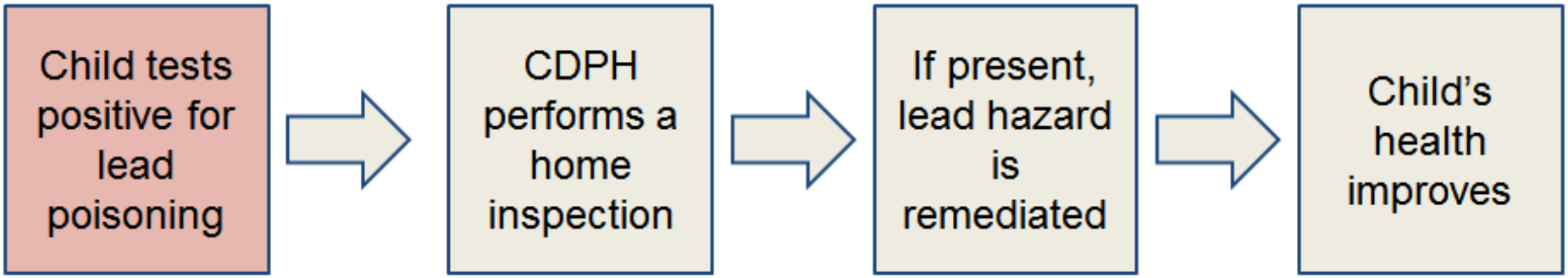
Figure : Mean decrease in IQ points for 1 mcg/dL increase in BLL (source: Mazumdar et al. Environmental Health 2011, 10:24)

The process of poisoning

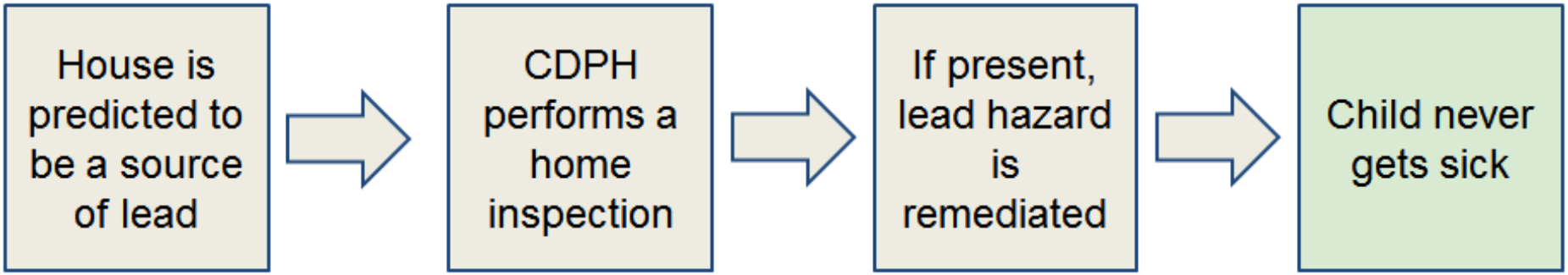


- No safe level of exposure
- permanent harmful effects

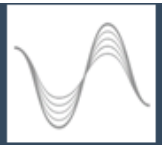
Current system



Our goal



Outline



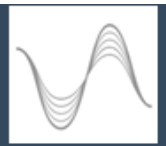
The data

- **CDPH inspections data** from 1994-2014: 80k houses
  - Interior and exterior building hazards
  - Inspection/intervention attempts
  - Successful remediation efforts
- **Cook County building assessor's records**: 800k houses
  - Building condition and structural details
  - Assessed total value
- **BLL records** for tested children in 1994-2013: 2.5 million tests
  - Recorded blood-lead level
  - Age, address at time of blood sample
- **2010 Census data**: 866 census tracts
  - Socio-demographic profiles per census tract

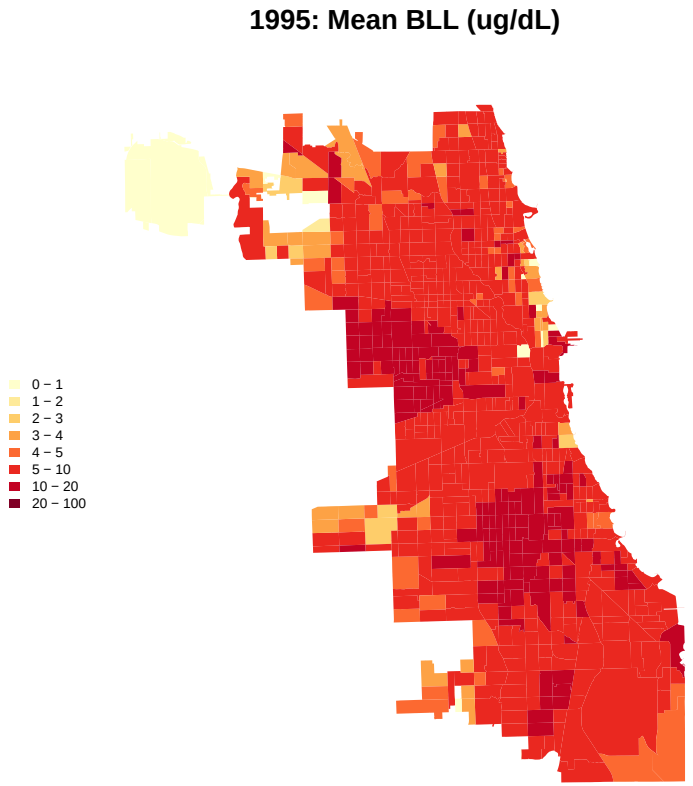
The deliverables

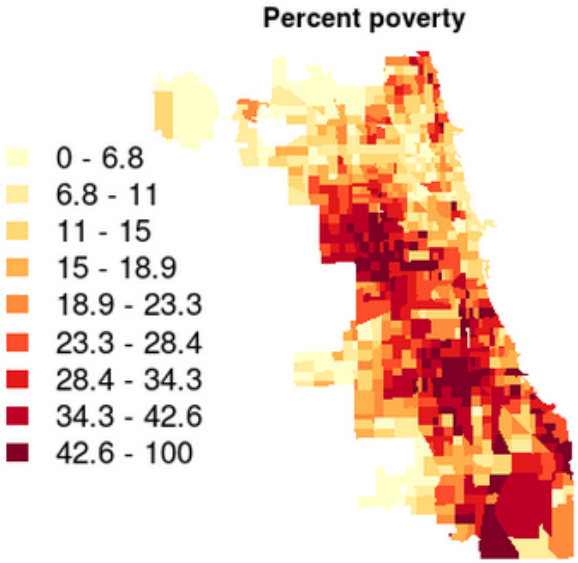
- Exploratory data analysis: visualizations, Shiny app
- Inferences: generating insights from the data
- A viable predictive model for lead poisoning: web application for health professionals
- Modeling lead poisoning through time for a specific child

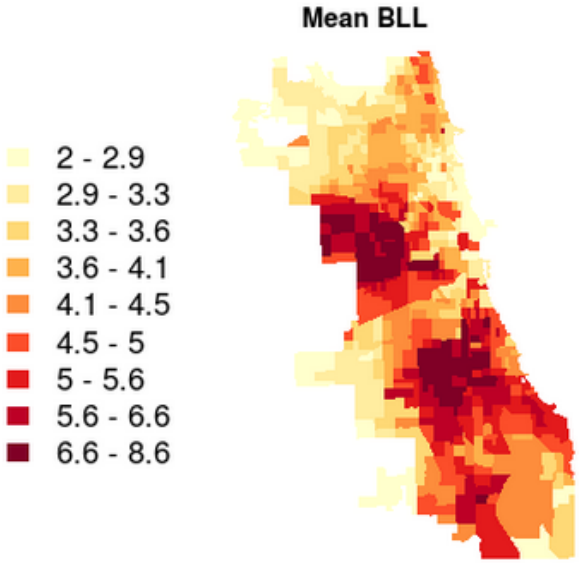
Outline

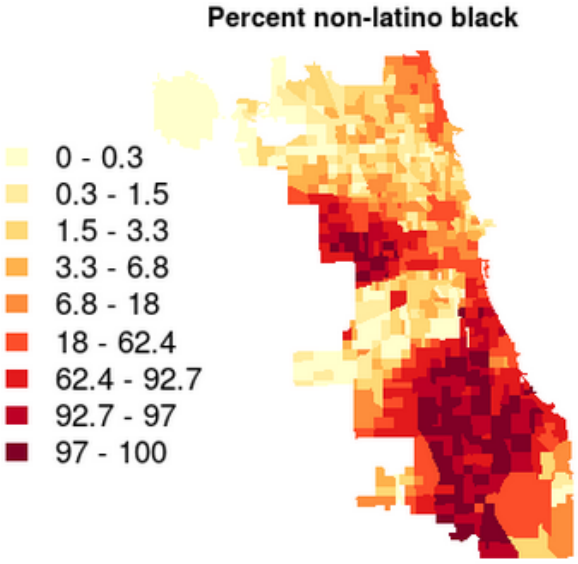


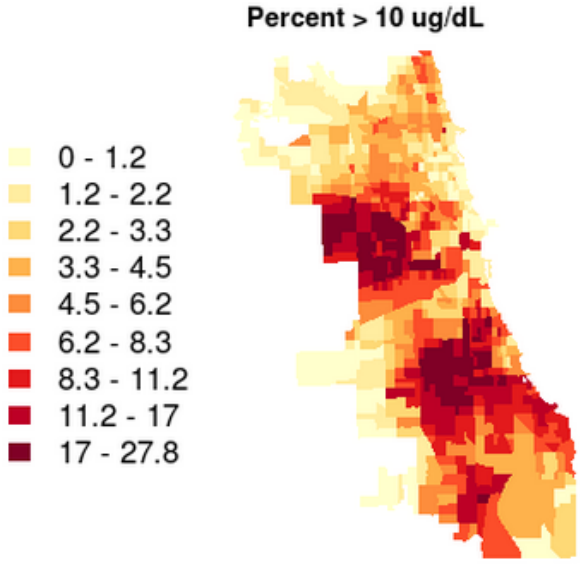
Scale of the problem




Percent poverty
0 - 6.8
6.8 - 11
11 - 15
15 - 18.9
18.9 - 23.3
23.3 - 28.4
28.4 - 34.3
34.3 - 42.6
42.6 - 100


Mean BLL
2 - 2.9
2.9 - 3.3
3.3 - 3.6
3.6 - 4.1
4.1 - 4.5
4.5 - 5
5 - 5.6
5.6 - 6.6
6.6 - 8.6

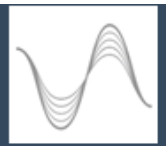

Percent non-latino black
0 - 0.3
0.3 - 1.5
1.5 - 3.3
3.3 - 6.8
6.8 - 18
18 - 62.4
62.4 - 92.7
92.7 - 97
97 - 100


Percent > 10 ug/dL
0 - 1.2
1.2 - 2.2
2.2 - 3.3
3.3 - 4.5
4.5 - 6.2
6.2 - 8.3
8.3 - 11.2
11.2 - 17
17 - 27.8

Shiny app

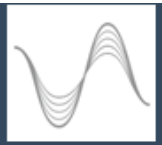
https://joebrew.shinyapps.io/cdph

Outline



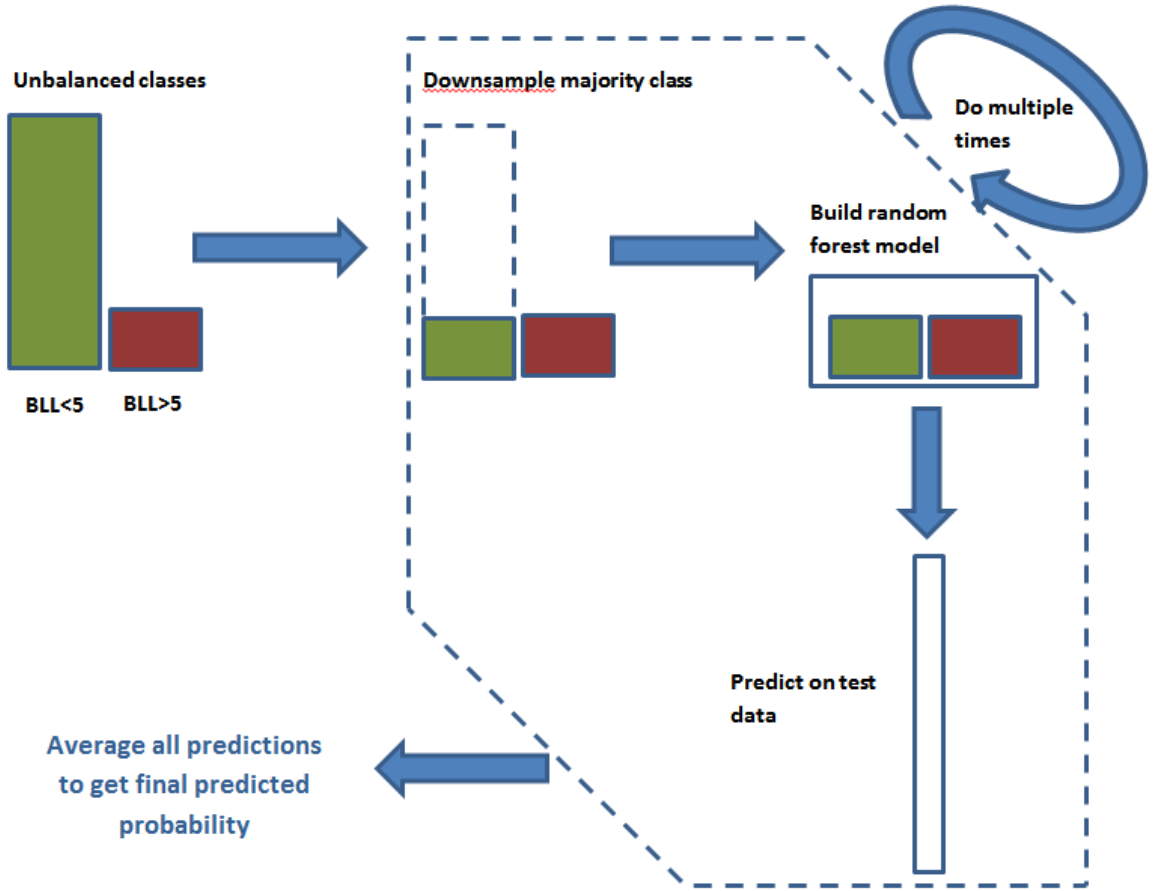
THE predictive model

- **Objective:** To predict whether a child of a certain age at certain address will have BLL>5.
- We tried different approaches, which include vanilla logistic regression, GAM, classification trees, random forests, poisson process models...
- Our final model is based on a modification of random forest.
- Feature set in final model:
  - House-level variables: location (lat/long), year built, condition(Good/so-so/bad)
  - Child-level variables: Age at testing, race, type of blood test (venous/capillary), time of year
  - Aggregate-level variables: % poverty in census tract

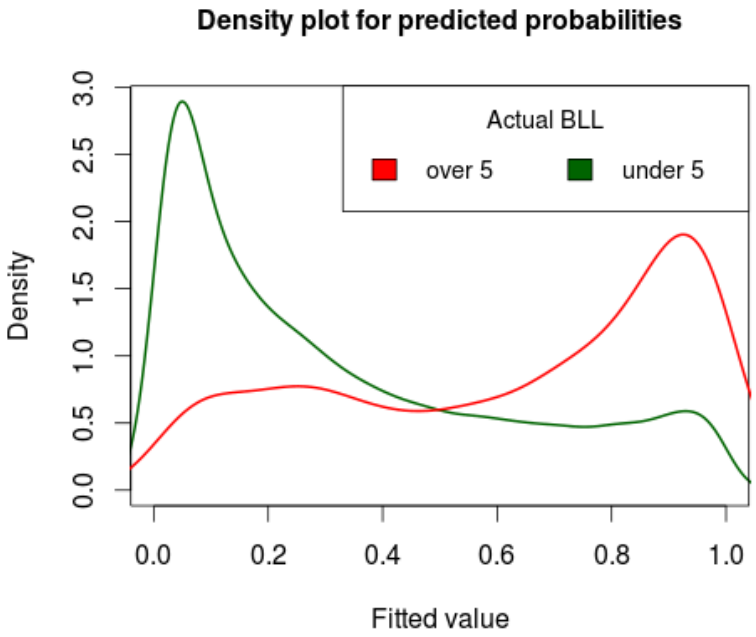


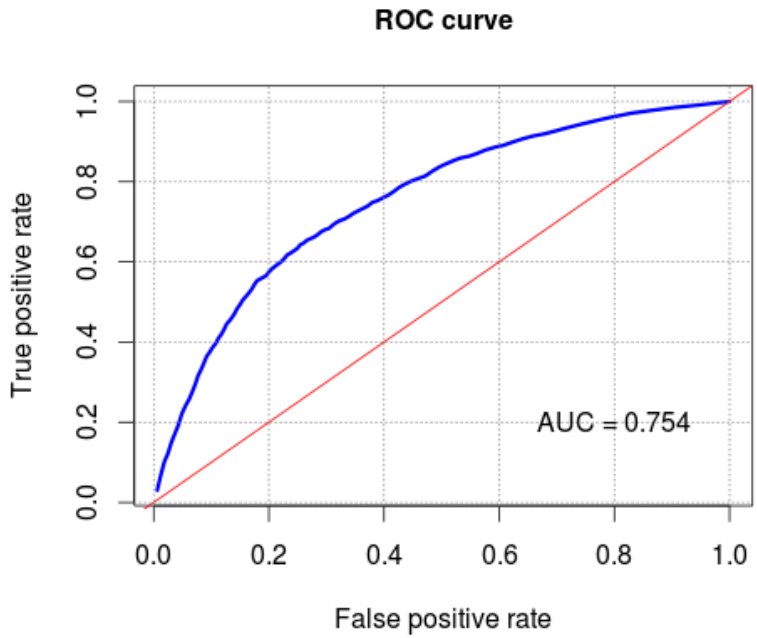
The `daggrForest` algorithm

Due to rare occurrence of high BLL the data is unbalanced, so vanilla random forest performs badly. We do multiple downsampling of majority class and subsequent model averaging to tackle this.



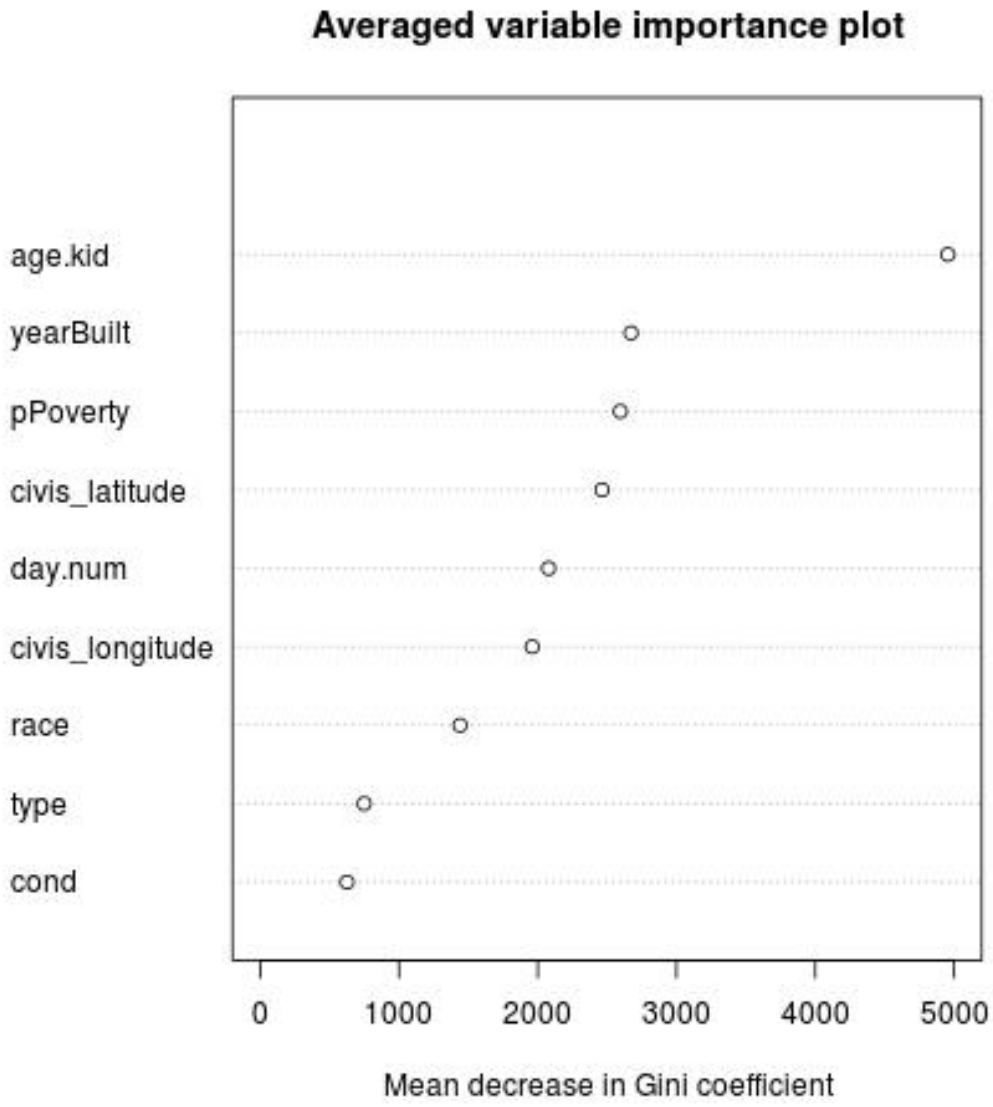
Model performance





- Training data: all samples from children aged below 5 yrs collected in 2002-2011.
- Test data: all samples from children aged below 5 yrs collected in 2012
- Only the highest blood test of a child, and tests below 20 mcg/dL are considered.
- Correct prediction of $\sim$70%

Variable importances



Improvements

- Current predictions are obtained assuming there is a 2 year old black child in each house, and the tests are taken in summer.
- Modeling brings the number of 'high-risk' ($P(BLL > 5) > 0.5$) buildings targeted for inspection down from 200,000 (Pre-1978 houses) to approximately 42,000. But that number still remains too high for feasibility.
- Our next step is to incorporate birth data so that we can target only homes with a child at the age of greatest risk (approximately 2 years old).
- Doing so is expected to bring our number of 'high-risk' homes down to fewer than 500.

Improvements

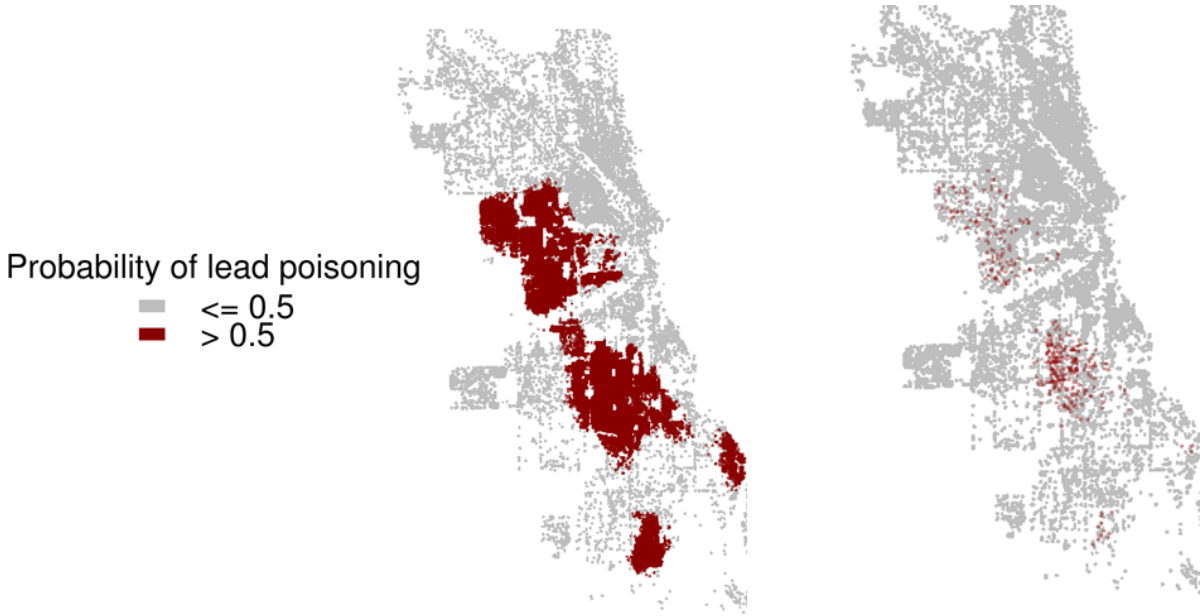


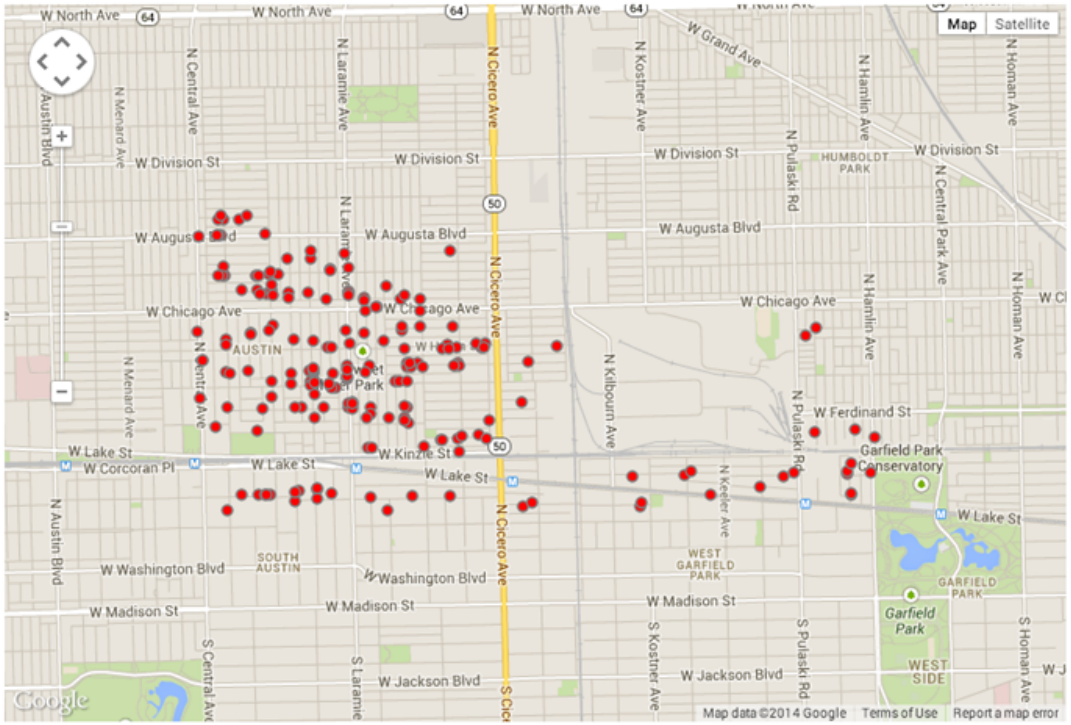
Figure : Buildings targeted for inspection using current model (left) and projected model with birth data incorporated (right). Projected model randomly puts children of specific age and race in each house. Simulation done based on 2010 census data

The application interface

City of Chicago Lead Inspections Data Portal

Home About Learn Contact

Circles represent buildings which match search criteria (click to view search criteria)



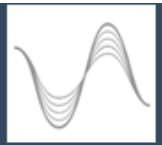
Search

Building Data | Control Panel

Address	Code	Risk	Insp	Comply	Tests
616 N XXXXXXXX	●	55.9%	✓	✓	3
5223 W XXXX	●	55.8%	✓	✓	5
539 N XXXXXX	●	55.6%	✓	✓	5
5235 W XXXX	●	55.4%	✓	✓	3
840 N XXXXXXXX	●	55.4%	✓	✓	1
556 N XXXXXXX	●	55.3%	✓	✓	6
514 N XXXXXX	●	55.3%	✓	✓	4
422 N XXXXXXXXXX	●	55.2%	✓	✓	2
5410 W XXXXXX	●	54.9%	✓	✓	7
723 N XXXXXXXX	●	54.9%	✓	✓	4
5430 W XXXX	●	54.8%	✓	✓	6
5044 W XXXX	●	54.7%	✓	✓	2
539 N XXXXXXX	●	54.5%	✓	✓	16
5430 W XXXXXX	●	54.5%	✓	✓	8
4307 W XXXXXXX	●	54.5%	✓	✓	3

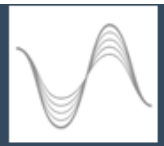
Query returned 200 records.

Outline



Acknowledgments

- My team: Joe Brew (Univ. of Florida), Alex Loewi (Carnegie Mellon), Andrew Reece (Harvard), and the mentor Eric Rozier (Univ. of Cincinnati).
- Rayid Ghani, Matt Gee and everyone else at DSSG.
- Dootika Vats, for letting me know about the fellowship.
- Abhishek Nandy, for listening to my practice talk.



THANK YOU!