

Equity in Resident Crowdsourcing: Measuring Under-reporting without Ground Truth Data

Zhi Liu¹ Nikhil Garg²

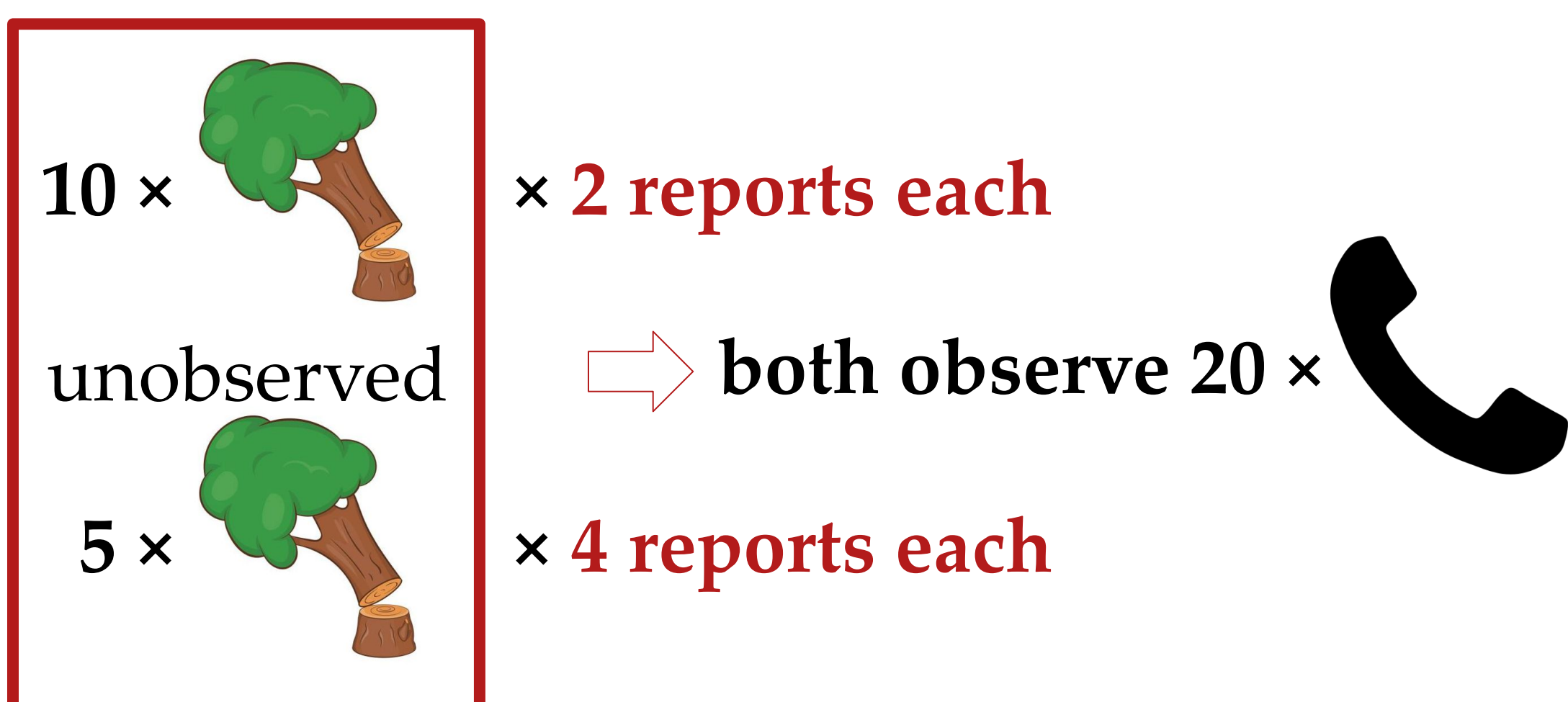
1. Operations Research & Information Engineering, Cornell University 2. Cornell Tech
with special thanks to the Urban Tech Hub @ Cornell Tech & NYC Department of Parks & Recreation

BACKGROUND

- Resident crowdsourcing systems
 - 311 systems, in NA
 - help cities learn about conditions: potholes, bedbugs, powerlines... ..and in our problem, **trees!**



- Historic **equity concerns** in reporting
 - technological disparities, awareness
- Hard to measure under-reporting**
 - observed **reports** governed by both **incidents & reporting behavior**
 - want to estimate **under-reporting**
 - no ground-truth about incidents**



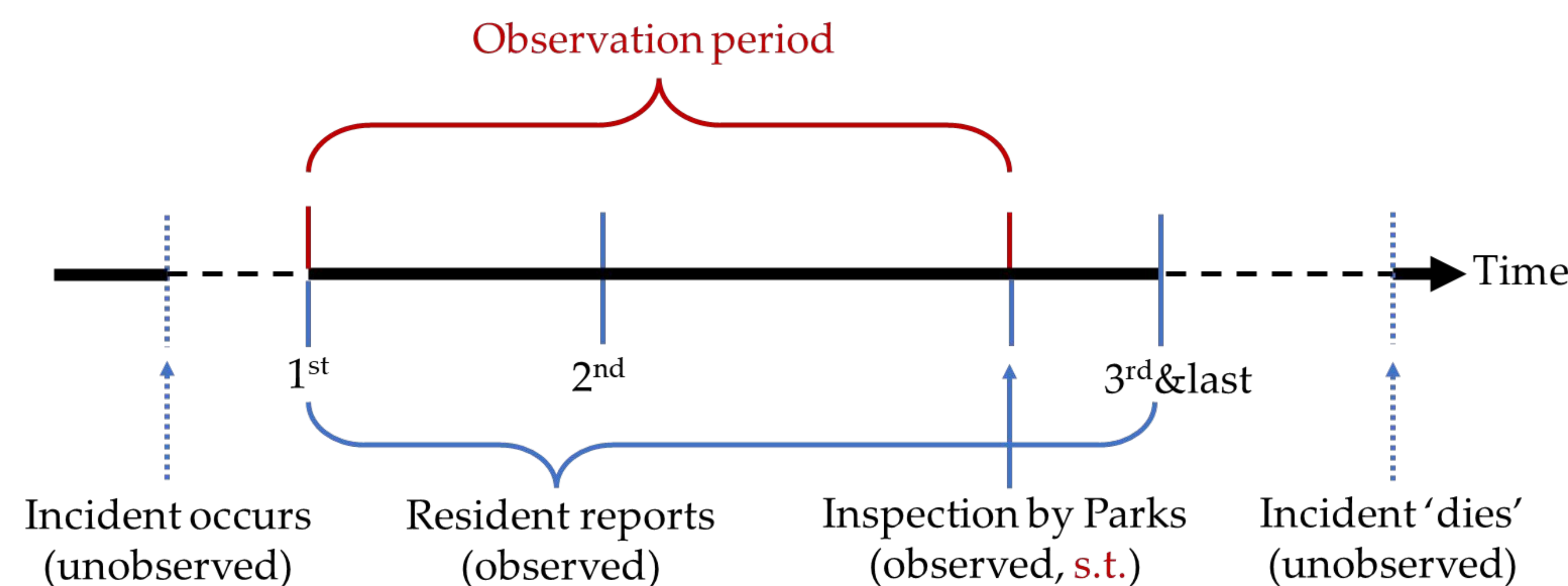
PREVIOUS WORKS

- Assume *same incident rate*
 - likely **not true**, due to varying conditions across the city
- Construct *proxies for reporting rate*
 - e.g. # reports/# trees
 - accuracy of results highly depend on accuracy of proxy, which **we cannot test**

MAIN RESULTS

- Technique which leverages **duplicate reports** to estimate reporting rate without ground-truth data, with theoretical guarantee
- Application to more than 100,000 311 reports in NYC, uncovering **substantial spatial and socio-economic disparities** in reporting

A new & general estimation technique



Model: conditional on incident happens, reports come in according to a **Poisson process** with potentially non-homogeneous rates, but incident birth and death times are unobserved
 → difficulty: **birth and death unobserved**

Theorem(restated): an **observation period** with start & end times:

- both **contained** in the period between incident birth and death;
- both **stopping times**, independent of Poisson rate;

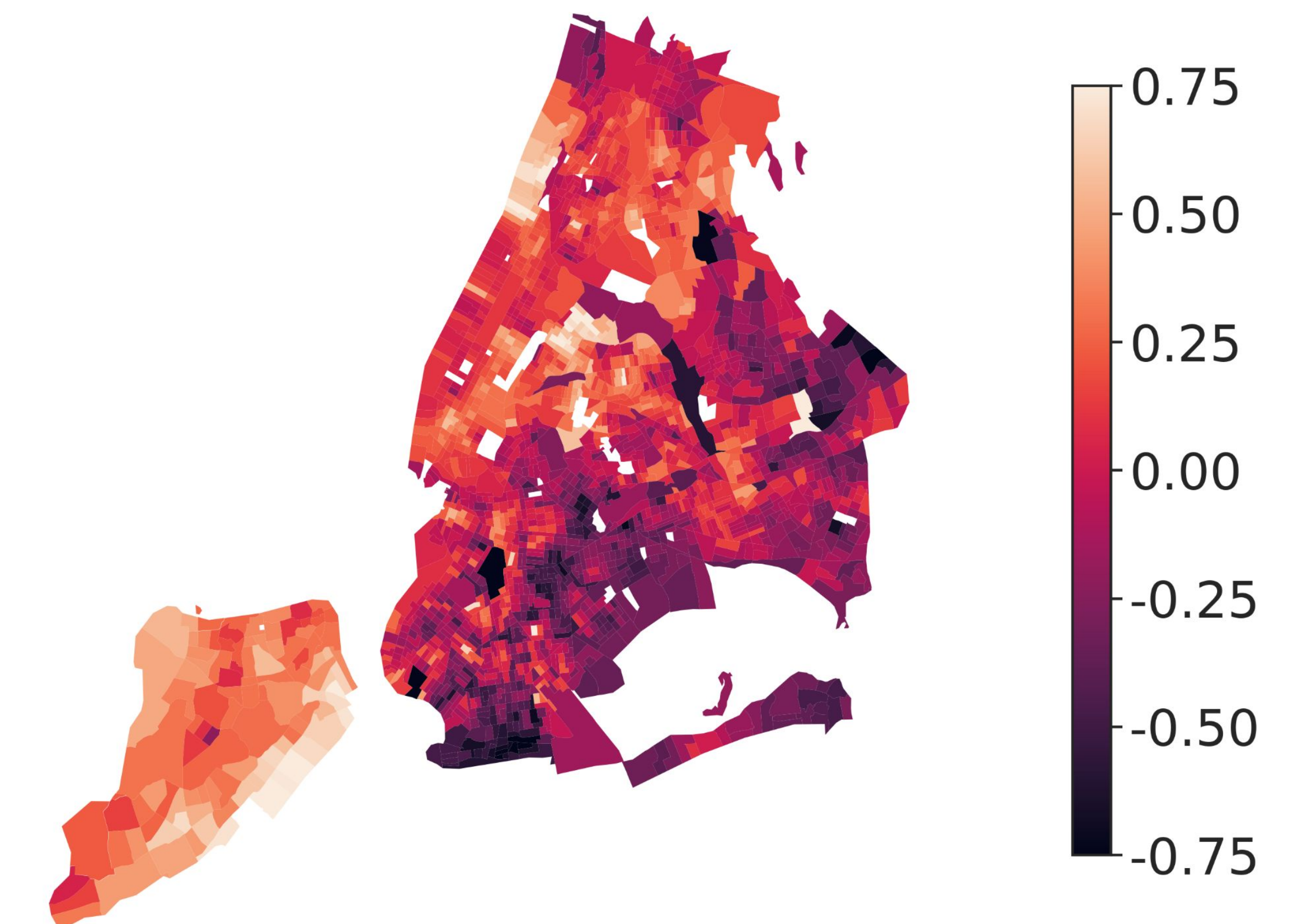
 then estimating the reporting rate in this observation period with the **duplicate reports** ⇔ estimating the overall reporting rate

Implementation on NYC Parks data

- Identifying observation period** from data:
- start time: time of first report
 - end time: **min**{inspection, fixed interval (100 days) from start}
- Bayesian Zero-inflated Poisson regression** model for estimation
- covariates: incident category, inspection outcome, tree info, location, demographics of reporter

IMPLICATIONS

Fig: Census tract spatial coefficients in NYC



Even **after correcting for incident category**:

- substantial **spatial disparities** across census tracts
- socio-economic disparities** across different groups

Tab: Implied reporting delay from estimated rates

Incident	Manhattan	Queens
Hazard, High risk, Poor tree	2.5 days	4.7 days
Root/Sewer/Sidewalk, Fair tree	112 days	209 days

Difference in **contextualized reporting delay**

- direct effort to push 311 system for more utilization
- provide foundation for downstream operations

ONGOING WORK

Deployable algorithm for inspection scheduling

- **efficiency**: inspect urgent incidents quickly
- **equity**: balance the birth-to-inspection time