Auditing for Spatial Fairness

Dimitris Sacharidis
Université Libre de Bruxelles
Belgium
dimitris.sacharidis@ulb.be

Giorgos Giannopoulos
Athena Research Center
Greece
giann@athenarc.gr

George Papastefanatos
Athena Research Center
Greece
gpapas@athenarc.gr

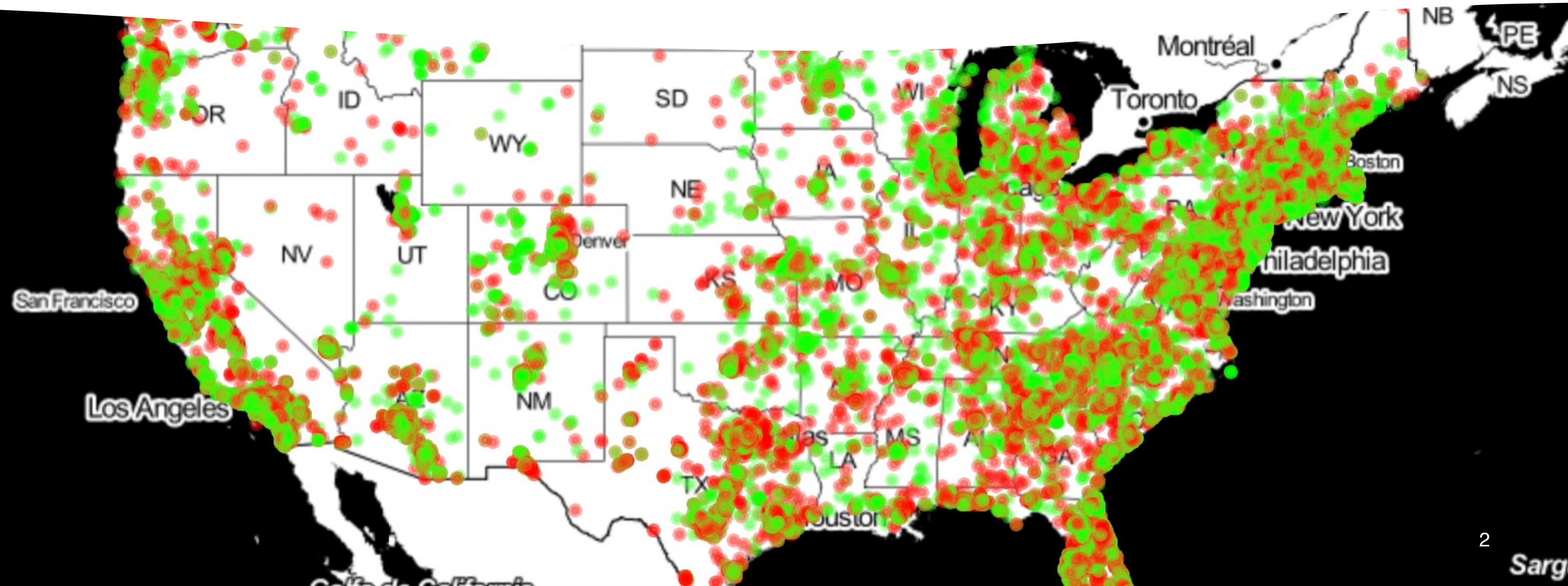
Kostas Stefanidis
Tampere University
Finland
konstantinos.stefanidis@tuni.fi

Motivation

An algorithm decides if **mortgage loan applications** are accepted.

Decisions should not depend on the **home address** of the applicant:

- To avoid **redlining** i.e., indirect discrimination based on ethnicity/race due to strong correlations with home address.
- To avoid **gentrification**, e.g., when applications in a poor urban area are systematically rejected to attract wealthier people.



Spatial Fairness – Definition

- Algorithmic Fairness: The **algorithm** (AI system, ML model, etc.) **should not discriminate** against individuals on the basis of a **protected attribute** (sex/gender, ethnicity/race, etc.)

More concretely:

- Choose a **performance measure** for the algorithm (e.g., recall)
  - Different choices result in different notions (e.g., equal opportunity)
- And require it to be statistically **independent** of the **protected attribute**.

- Spatial Fairness: **protected attribute = location**

Fairness In Practice

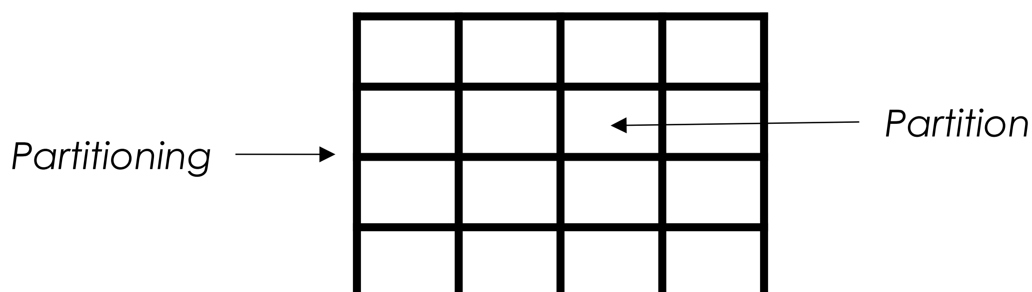
In practice, **group-comparison** test is performed.

Compare the **performance measure** across **protected groups** (individuals with same protected value).

- e.g., it's fair when recall for males = recall for females

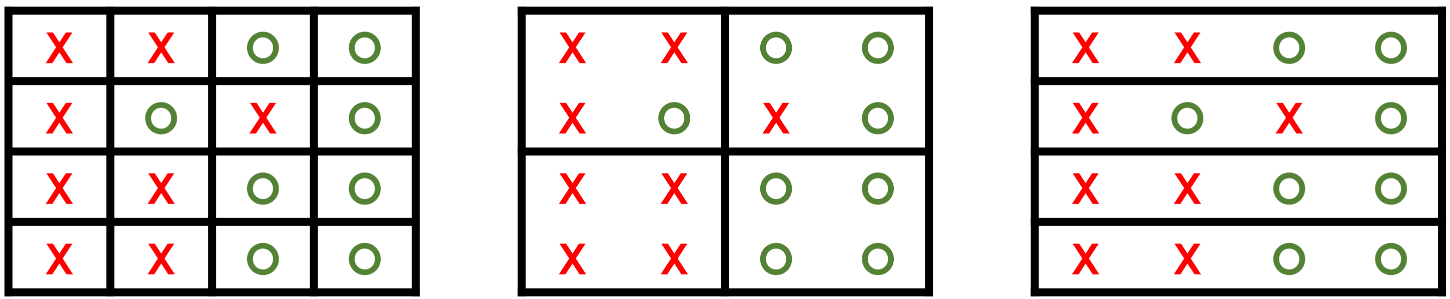
Q: How to define **groups** for the location attribute?

A: With a **partitioning** of the space in regions. (Right?)



Spatial Fairness – Challenges

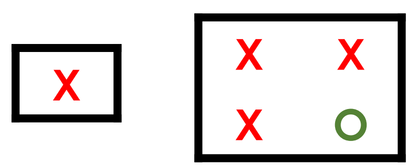
Be aware of **gerrymandering**, i.e., purposefully defining the partitioning to hide discrimination.



UNFAIR *UNFAIR* *FAIR*??

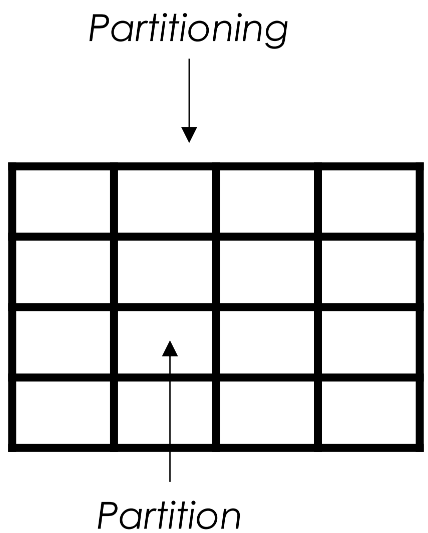
Be aware of the **modifiable areal unit problem**, i.e., statistical bias when comparing conclusions drawn from partitions of different shape and scale

Do not compare!



Spatial Fairness – Prior Work

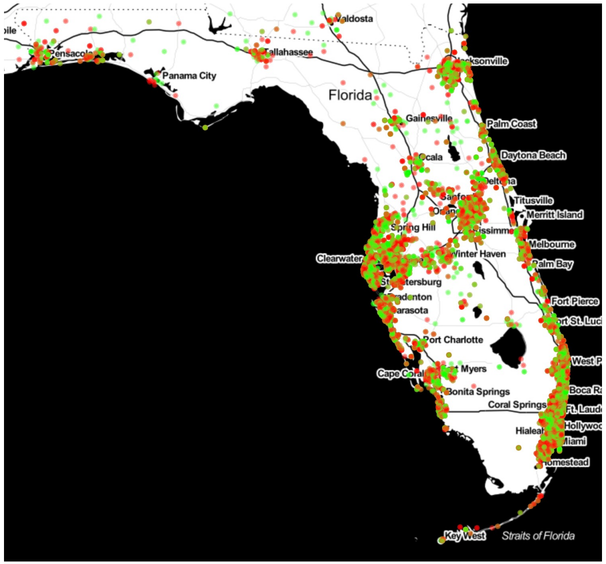
- To address these two challenges, MeanVar [AAAI 2022]
  - Considers **all possible rectangular partitionings** of the space.
  - For each partitioning, computes the **variance of the performance measure** in the partitions,
  - Finally, reports the **mean variance** across partitionings.
- low MeanVar = high fairness

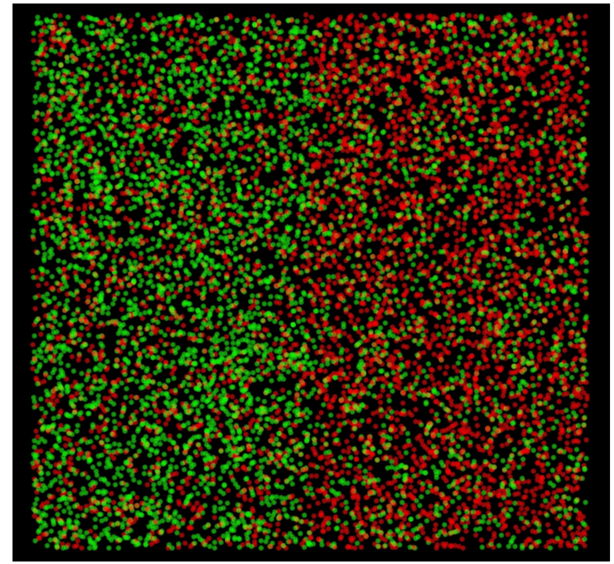


But leads to **counter-intuitive** conclusions when observations are **not regularly distributed** on a grid.

[AAAI 2022] Xie et. al. *Fairness by "Where": A Statistically-Robust and Model-Agnostic Bi-level Learning Framework. AAAI Conf.*

Spatial Fairness – Prior Work





- **Fair by Design** (non-regularly distributed points, no spatial bias)
- has higher MeanVar **0.05**
- appears **less fair**

- **UnFair by Design** (regularly distributed points, spatial bias)
- has lower MeanVar **0.04**
- appears **more fair**

Cannot answer the question: ***Is it fair?***

Spatial Fairness – Our Solution

- Design choices for spatial fairness:
  - Can **audit:** "*Is it fair*?"
  - Can **testify**: "*Where is it unfair*?"
  - Works for non-regularly distributed observations.

- No partitionings, no comparison among fixed groups.

Intuition: For **any region** of the space, the **performance measure** should be roughly **the same inside and outside the region.**

Spatial Fairness – Our Solution

- Define a statistical test to quantify which is more likely:
  - **inside = outside** (H0: spatial fairness)
  - **inside ≠ outside** (H1: spatial unfairness)

Inspired by work on spatial-scan statistics [Comm. Stat. 1997]

- Define the **likelihoods** L_0 and L_1 of hypotheses H0 and H1 given data
- **Scan the space** (i.e., visit a large number of regions) and estimate the **maximum likelihoods** L_0^{max} , L_1^{max}
- Compute the likelihood ratio **test statistic** $\tau = \frac{L_1^{max}}{L_0^{max}}$
- Determine the **p-value** of the test statistic

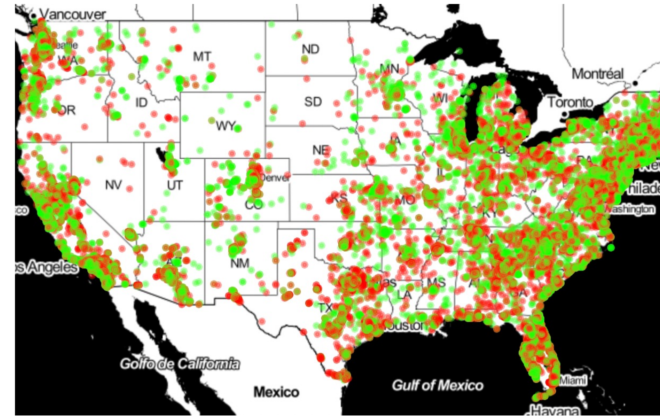
TO AUDIT: If p-value below a **significance level** α , then it's **spatially fair.**

TO TESTIFY: Return all scanned regions with p-value above α.

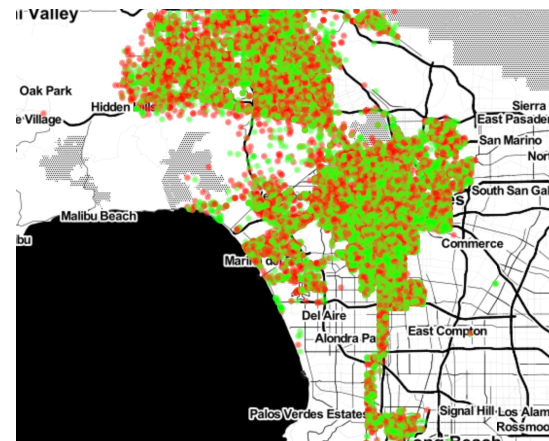
[Comm. Stat. 1997] M. Kulldorff. *A spatial scan statistic. Comm. Stat. Journal 26, 6*

Evaluation – Datasets

- **LAR**: mortgage Loan Application Register data for Bank of America in 2021 in US
  - 200K **loan applications**, 50K locations
  - **green**: loan approved 120K
  - **red**: loan rejected 80K

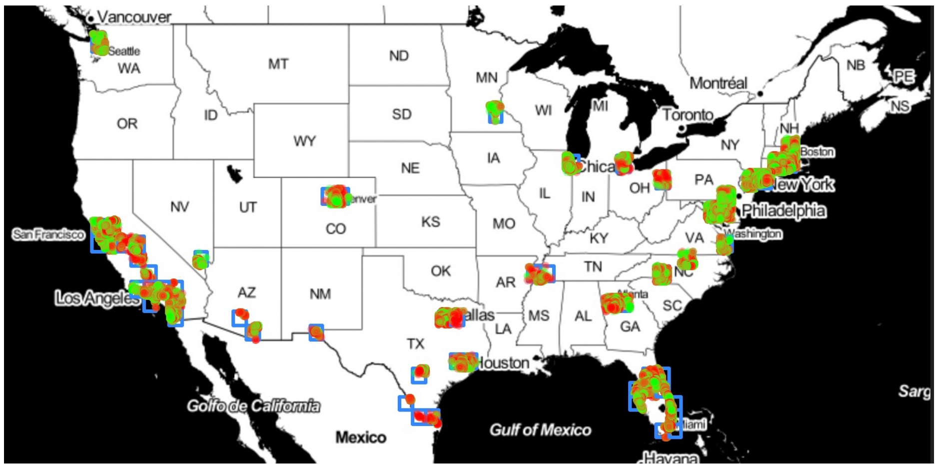


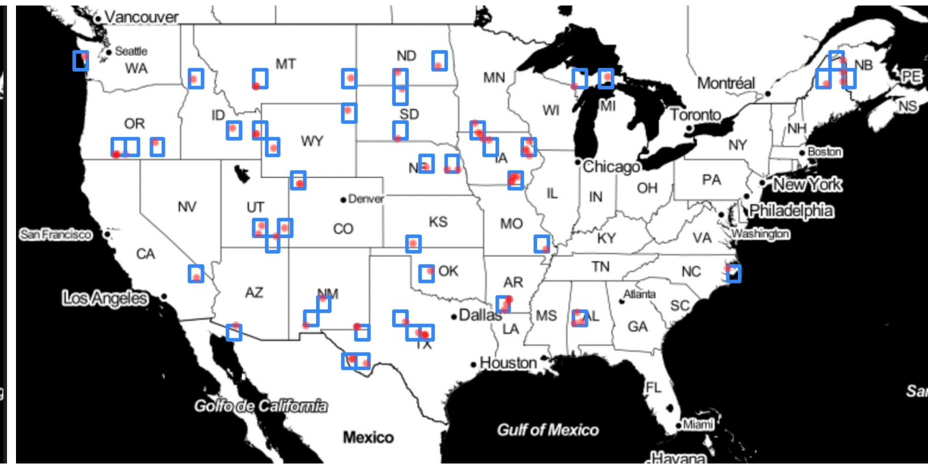
- **Crime**: crime incidents for 2010-2019 in Los Angeles
  - Test set of 60K **serious crimes**
  - A random forest classifier predicts serious crimes (recall/tpr = 0.58)
  - **green**: true positives 35K
  - **red**: false negatives 25K



Evaluation – Results on LAR

For fair comparison with MeanVar, our approach only scans the regions from a partitioning





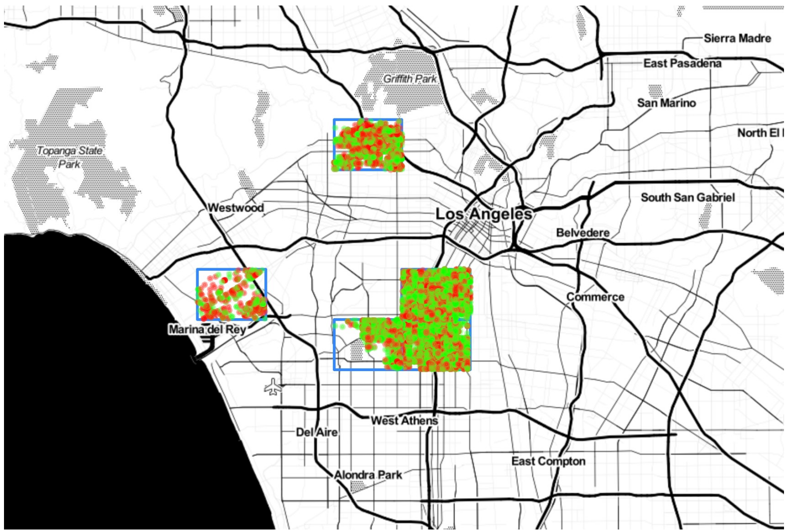
Our approach

- Declares **unfairness** and identifies 59 statistically significant unfair regions.
- **Dense** regions with **small deviations** from the performance measure mean.

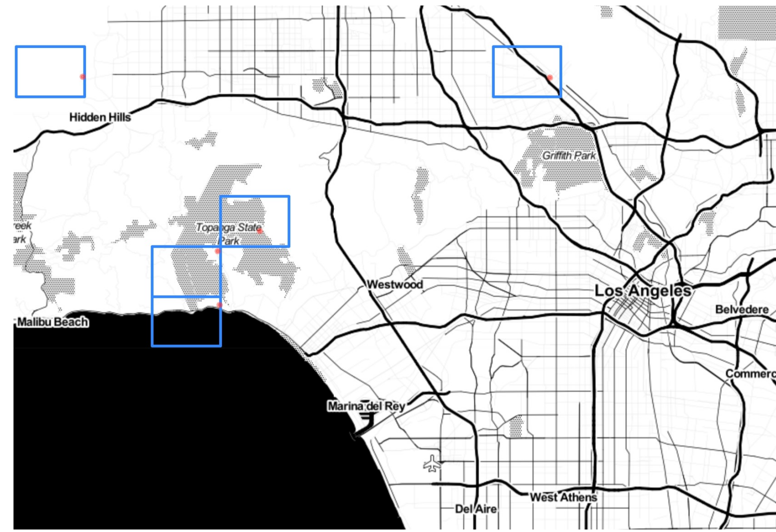
MeanVar

- Top-50 regions with highest contribution to MeanVar.
- **Sparse** regions, but **large deviations** from performance measure mean.

Evaluation – Results on Crime



Our approach
- Declares **unfairness** and identifies 5 statistically significant unfair regions.
- **Dense** regions with **small deviations** from the performance measure mean.



MeanVar
- Top-5 regions with highest contribution to MeanVar.
- **Sparse** regions, but **large deviations** from performance measure mean.

Auditing for Spatial Fairness

https://arxiv.org/abs/2302.12333

https://github.com/dsachar/AuditSpatialFairness