Stratifying the Digital Divide: Analysis of Socio-Economic Influences on Internet Performance

SCAN ME

Shivani Kalamadi [1], Aditya Kumar Bej [1], Sachin Kumar Singh [1], Varshika Srinivasavaradhan [2], Elizabeth Belding [2], Alexander Gamero-Garrido [1]

[1] *University of California Davis,* [2] *University of California Santa Barbara*

UC DAVIS UNIVERSITY OF CALIFORNIA

RQ1: How are crowdsourced speed tests distributed across regions, neighborhoods, and years?

RQ2: To what extent does population density confound socio-economic predictors, and how does stratifying by density change what matters?

RQ3: After controlling for density, which socio-economic factors shape internet performance, and do they differ across regions?

Motivation

The Problem

Internet quality gaps persist across U.S. neighborhoods, not just urban vs. rural.

San Francisco 2025 download speed (Mbps): Min: 46.3, Max: 602.2, **SD: 59.3**.

Coverage maps show where the internet goes. Socio-economic analysis of internet measurement shows who it serves.

Related Work

1. National/state-level studies mask neighborhood variation [1]
2. Granular single-city studies exist but findings are not transferable
3. No prior work examines temporal consistency in the U.S.
4. Linear models assume uniform socio-economic effects → our results show they are not

Why CBG Granularity?

Census Block Groups (CBG)

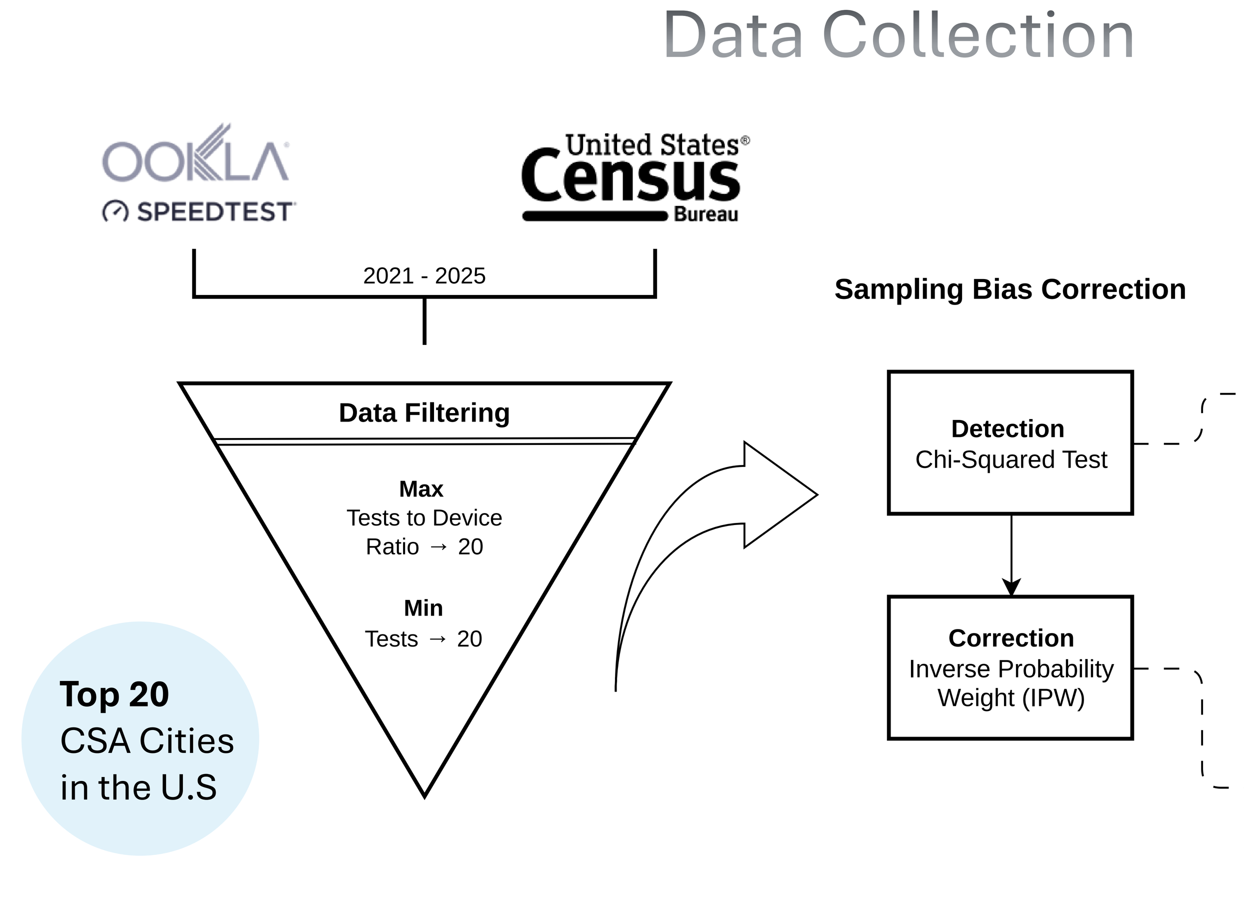
~1,500 residents * 220k CBGs is the finest census unit with demographic data that captures block-level variation

Why CSA Aggregation?

Combined Statistical Area (CSA)

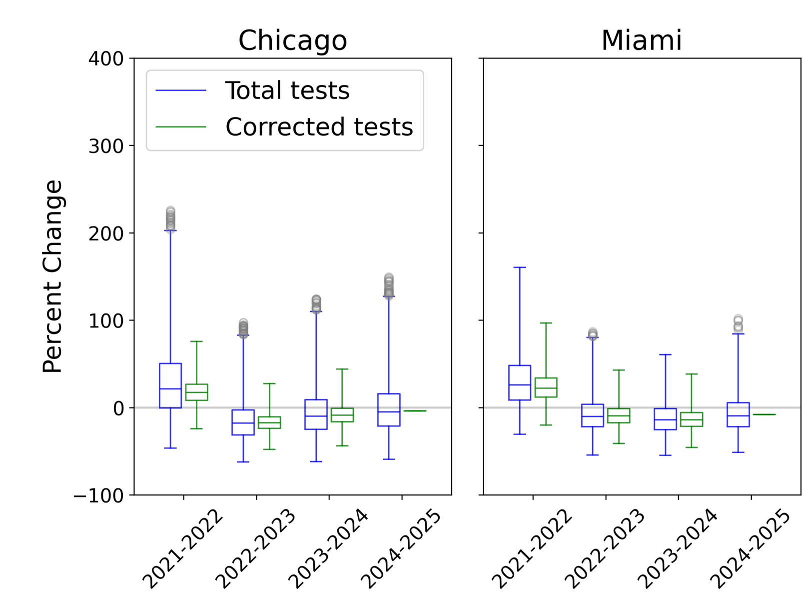
Enough CBGs per CSA for reliable density stratified modeling. Higher aggregation masks the intra-city variation.

Data Collection



$$Weight_{CBG} = \frac{population\ share_{CBG}}{test\ share_{CBG}}$$

Mean % Difference in Expected & Actual Tests	
Chicago	9%
Dallas	17%
San Francisco	28%
New York	41%
Washington DC	45%
Los Angeles	54%
Miami	119%



Sampling Bias Correction

1. Crowdsourced tests are not distributed proportionally across CBGs relative to population [2]
2. Upweights under-sampled neighborhoods, downweights over-sampled ones

Temporal Testing Patterns

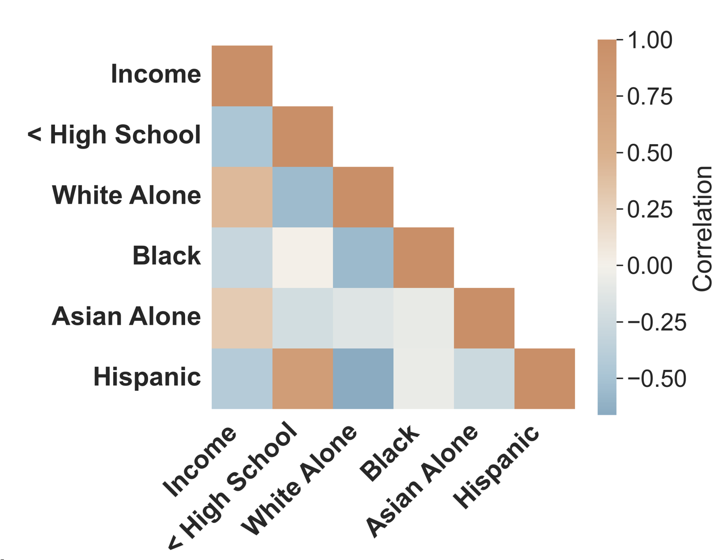
1. Year-over-year test counts vary significantly across regions
2. Consistent dip observed in 2022-2023 across all cities, followed by quieter changes through 2025.
3. Bias correction applied independently per year to account for shifting test distributions over time

→ RQ1 Takeaway

Method

Step1: Address Multicollinearity Among Census Features – varies by city

- Socio-economic features are inherently correlated
- Standard approaches such as Pearson/Spearman correlation or linear regression cannot reliably attribute independent influence to each feature
- The goal is to isolate each census feature's individual influence on network variables, one at a time.



Step2: Random Forest Regression Model

- Ensemble of decision trees that ranks features by their contribution to reducing prediction error
- Random Forest handles multicollinearity natively by selecting the most informative feature at each split
- Allows all features to compete fairly without requiring manual variable selection or dimensionality reduction.
- Hyperparameters:
  1. **Number of Trees** = **200** ∈ {100,200,500}
  2. **Max Depth** = **20** ∈ {10,20,30}
  3. **Min Split** of CBGs = **5** ∈ {2,5,10}

Step3: Improve & Validate : Permutation Model & Null Model

- Permutation Importance: **Shuffle one feature's values →** measure accuracy drop. Repeated for each census feature and network metric.
  - **Why :** Directly tests whether a feature carries genuine predictive signal above the others.

Income	Download Speed
100,000	40 Mbps
200,000	60 Mbps
300,000	100 Mbps

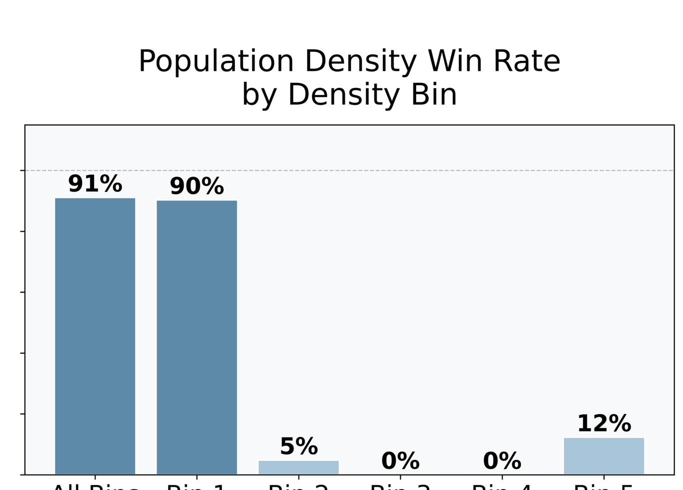
- Null Importance: **Shuffle the network variable →** measure accuracy drop. Repeated for each census feature and network metric.
  - **Why :** Establishes a chance-level baseline between Socio-economic feature and Network variable.

Income	Download Speed
100,000	40 Mbps
200,000	60 Mbps
300,000	100 Mbps

Step4: Population Density Significance & Stratification approach

- Density **dominates vast majority of region, year, and metric** as the most important feature
  - It is obscuring other socio-economic relationships when unstratified
  - Once stratified, its influence disappears from Bin 2 and above.

  → RQ2 Takeaway
- Partition CBGs into **5 equal-sized quantile bins** (Bin 1 = lowest → Bin 5 = highest density) and re-run the full model within each stratum
  - Each bin has sufficient data points.



Results

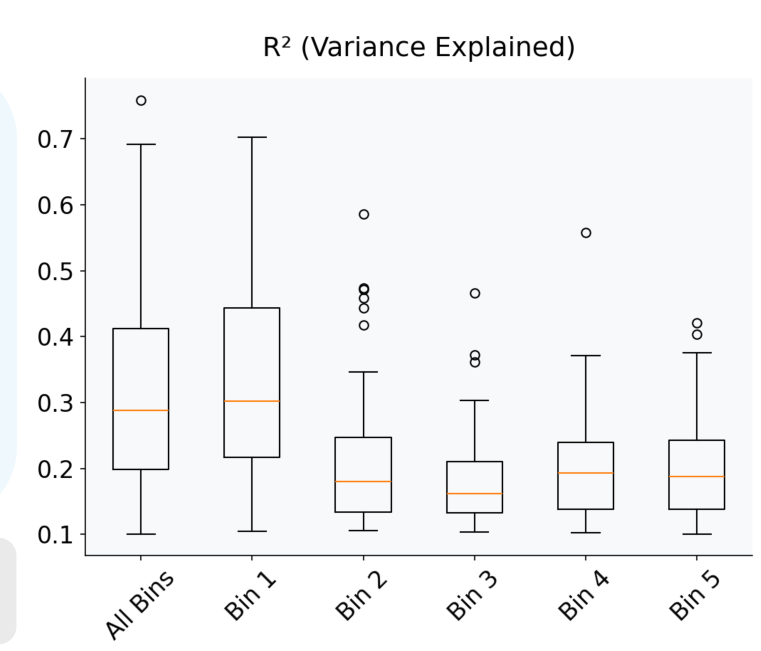
Filter out Noisy results

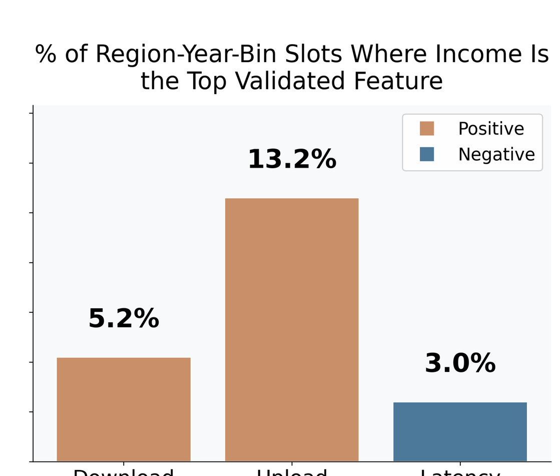
1. Bin's model must **explain ≥10% of variance** to be trusted at all (R^2)
2. A feature's **permutation importance** must itself be **≥ 0.10**
3. Must beat the **null baseline by at least 0.10**

4.7% features correlations remain important.

R^2 Values are ~30% without stratification and for lowest density bins and 15%-19% for the rest → Meaning socio-economic features explain little about the differences in bandwidth after Bin 2.

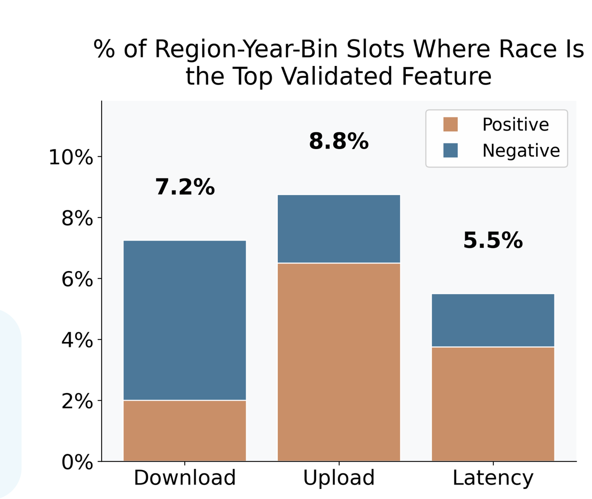
Since population density dominates the unstratified and lowest-density (Bin 1) results, rest of our analysis focuses on Bin 2- Bin 5.

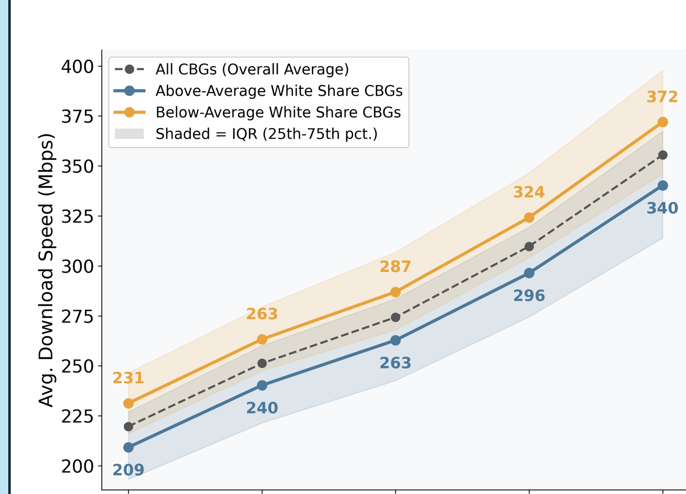




← Income is the strongest correlate of upload speeds, and less with download speeds and latency.

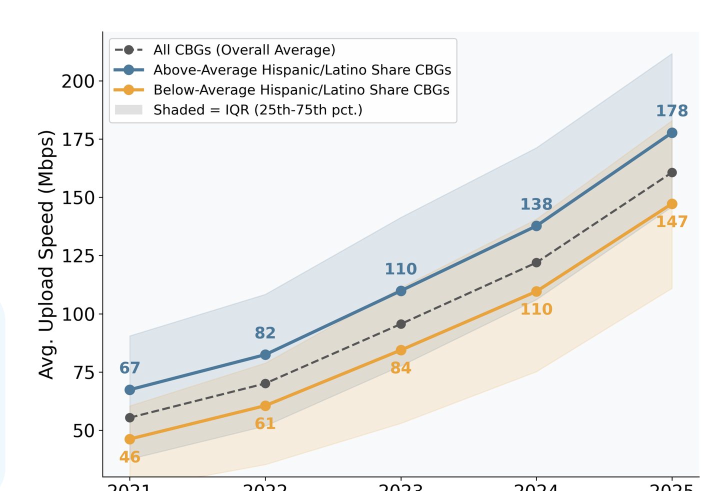
Race is another important socio-economic feature but has different correlations in different regions against network metrics. →





← New York - All 5 years **White Population Negative** correlation with download speeds

Miami - All 5 years. → **Hispanic Population Positive** correlation with upload speeds



→ RQ3 Takeaway: No single socio-economic factor explains internet performance uniformly. Both download speeds and upload speeds have varying correlations with socio-economic features across cities. **Internet inequality is locally configured.**

References:

1. U. Paul, J. Liu, V. Adarsh, M. Gu, A. Gupta, and E. Belding, "Characterizing performance inequity across U.S. Ookla Speedtest users," *arXiv preprint arXiv:2110.12038, 2021.*
2. H. Lee, U. Paul, A. Gupta, E. Belding, and M. Gu, "Analyzing disparity and temporal progression of internet quality through crowdsourced measurements with bias-correction," *arXiv preprint arXiv:2310.16136,2023.*

Ongoing Work: (1) Adding infrastructure and funding context to each of these cities **(2)** Case studies of neighborhood level differences **(3)** Extending methodology to other countries