The Digital Usage Divide: Evidence and New Measures from 40 Million Windows Devices (Formerly "The New Digital Divide")

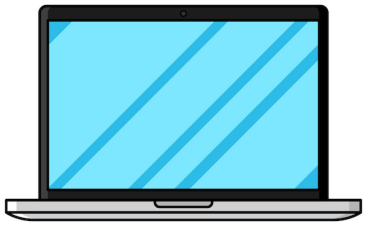
December 2025

Mayana Pereira, Microsoft, Shane Greenstein, Harvard, Raffaella Sadun, Harvard, Prasanna Tambe, U of Penn, Lucia Ronchi Darre, Microsoft, Tammy Glazer, Microsoft, Allen Kim, Microsoft, Rahul Dodhia, Microsoft, Juan Lavista Ferres, Microsoft

Free data.

- The dataset of indices is available on GitHub.
- https://github.com/microsoft/Consumption-of-Digital-Applications-Data-Set.
  - Will post link again at the end of the talk.

Overview

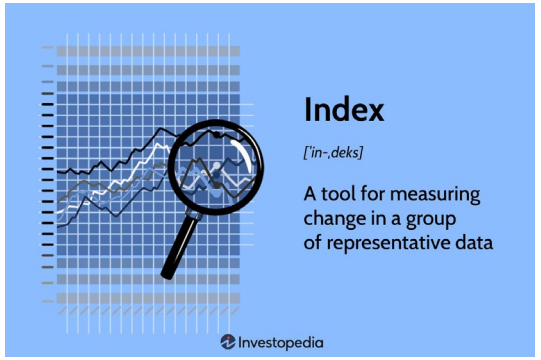


- Digital usage.
  - The ability to effectively and responsibly find, evaluate, use, create, and communicate information using digital technologies.
- Long-standing policy concerns.
  - The skills and competencies required to access, analyze, evaluate, create, and communicate information.

- Telemetry data.
  - Data collected by Microsoft in 2023 during operating system updates for more than 40 million Windows devices.
  - Minutes of usage of application categories. Across U.S. households that agreed to share this data.
- Highly sensitive.
  - How to gain insight and preserve privacy?

Index Construction Considerations: Ideals versus Pragmatics

- An ideal index of usage measures variance in usage across different geographic regions and w/generally.
  - Preserves the anonymity of each household.
  - Samples many observations for statistical precision.
  - A census is the best that can be done.
  - Benchmarks for potential comparisons over time.
- A practical index. Some lessons from other areas.
  - Computes an average or median for households within a geographic area.
  - Uses existing geographic boundaries, allowing for matching with other data (e.g., demographic info).
  - Sacrifices some details to highlight broad patterns.

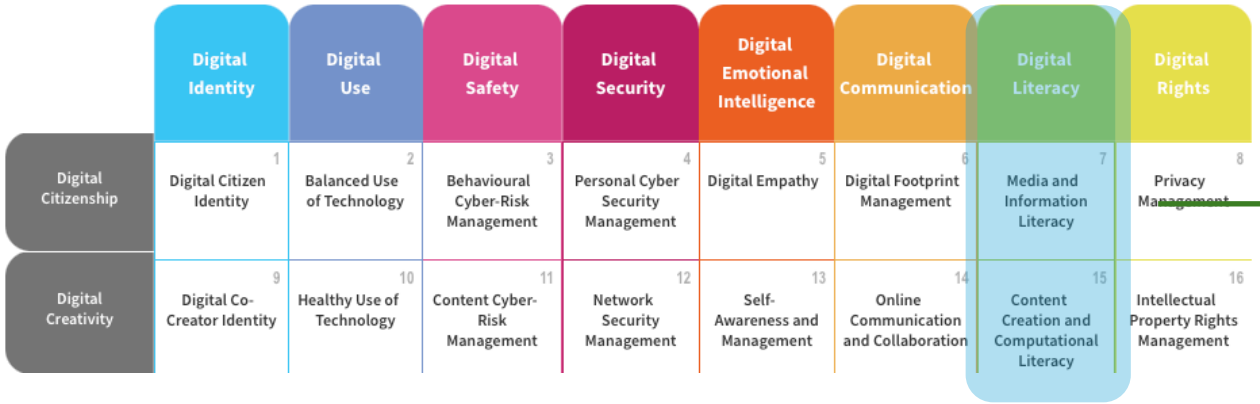


- Usage indices of the census of activity are rare.
  - Price indices are an area that frequently utilizes them.
  - We borrowed many elements from standard practices in that field.

Two indices guided by the OECD/IEEE definitions for digital literacy

The Media and Information Composite Index (MCI):
General computing usage across various applications, including word processing, spreadsheets, and presentations.

The Content Creation and Computation Composite Index (CCI): specialized digital applications, such as image manipulation tools (e.g., Photoshop) or software development tools.



MCI: Individuals understand the basic structure of digital media and its impact on knowledge and information.

CCI: Individuals understand the theory of digital content creation and computational thinking and possess algorithmic literacy.

Pragmatics: 12 aggregate categories of usage split into two groupings.

Application categories

Windows devices data set

- Browser
- Communication
- Document Authoring
- Email
- Media
- Presentation
- Spreadsheet
- Creative
- Developer tools
- Security
- Cloud Storage
- Utilities

multiple indicators

Media and information literacy

Content creation and computational literacy

Extensive consulting with internal MS experts.

Media and information consumption

We split the 12 categories into two groups, resembling the IEEE standard.

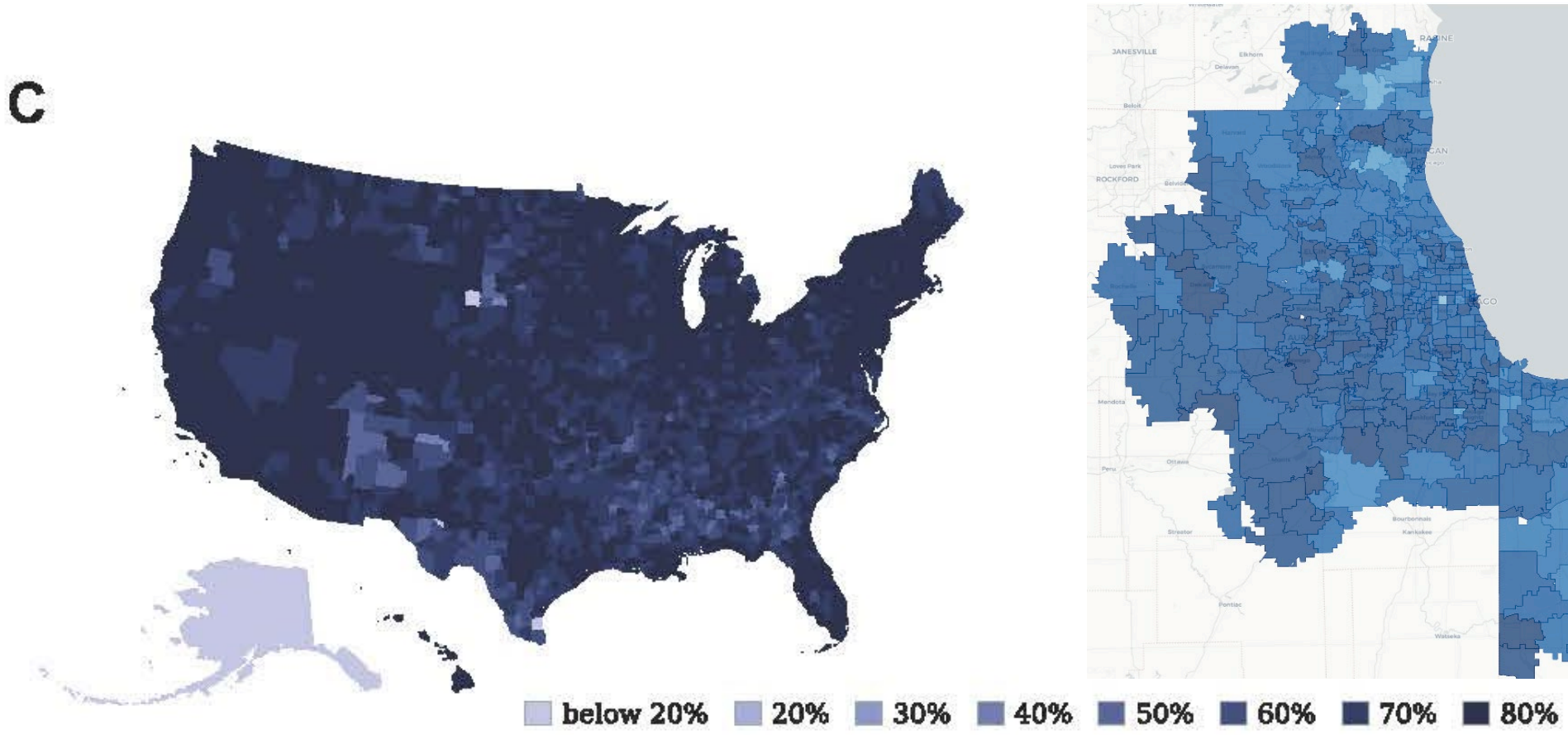
Content creation and computational consumption

Seven categories reveal the use of media, knowledge, and information, and five reveal the use of creation and computational thinking.

Two indices for each zip code.

- We make an MCI and CCI for each household. For each MCI and CCI, we calculate a weighted average of the time spent on applications, where the weights were derived from principal component analysis (PCA).
- For each index for a zip code, we calculated a weighted average of MCI and CCI across households in the zip code, where the weights represent the total time each household spends on the PC.
- We added some noise to the data, following standard practices for differential privacy. This results in two indices for 28,199 zip codes.
- What do we get? We get too much!
  - Will illustrate general patterns through the use of maps and graphs.
  - Show the US. Show the Chicago area. It is compact enough to fit into one picture, and its population demographics vary widely.

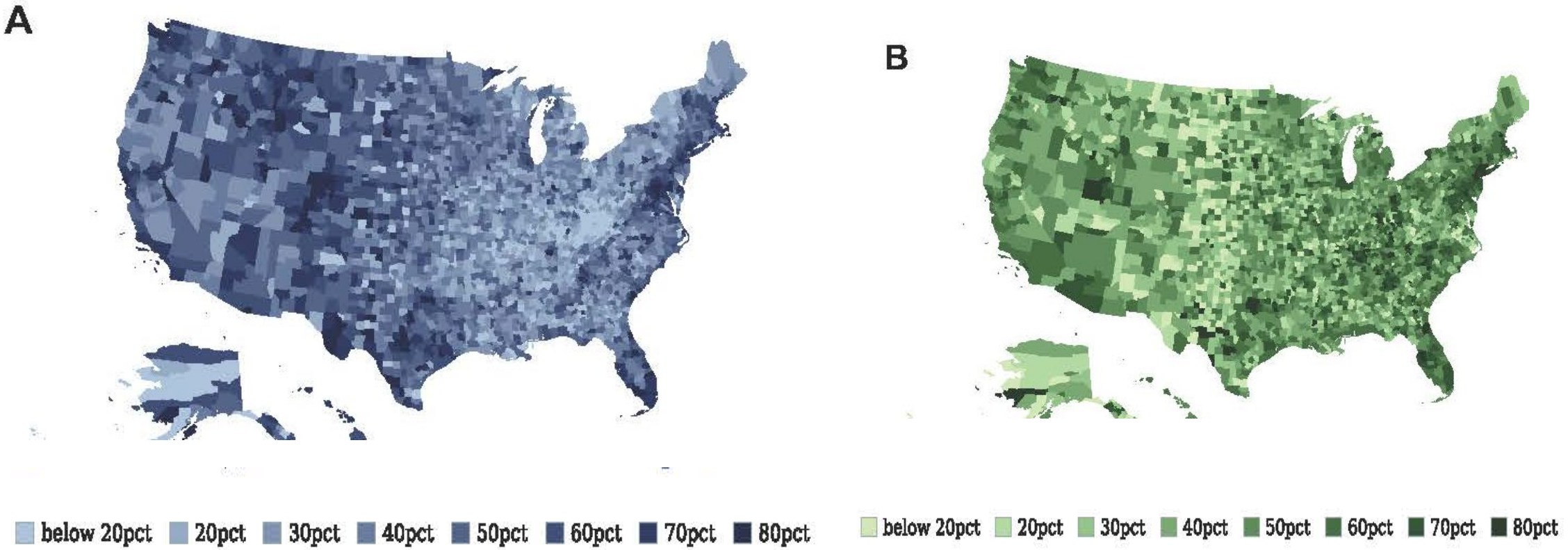
Broadband Availability: US and Chicago-Naperville-Elgin-IL-IN



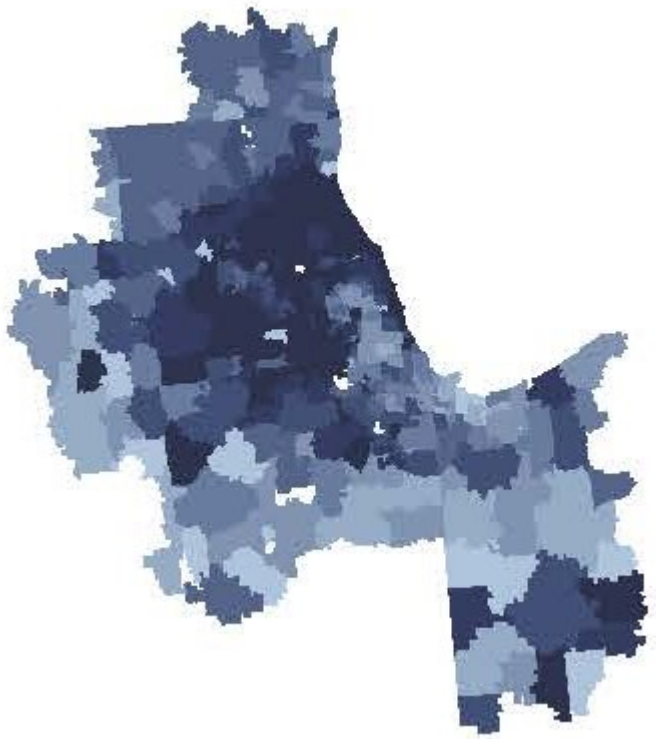
There is some variation in broadband availability at the zip code, but (as you will soon see) not enough variation to explain the variation in MCI/CCI.

MCI and CCI

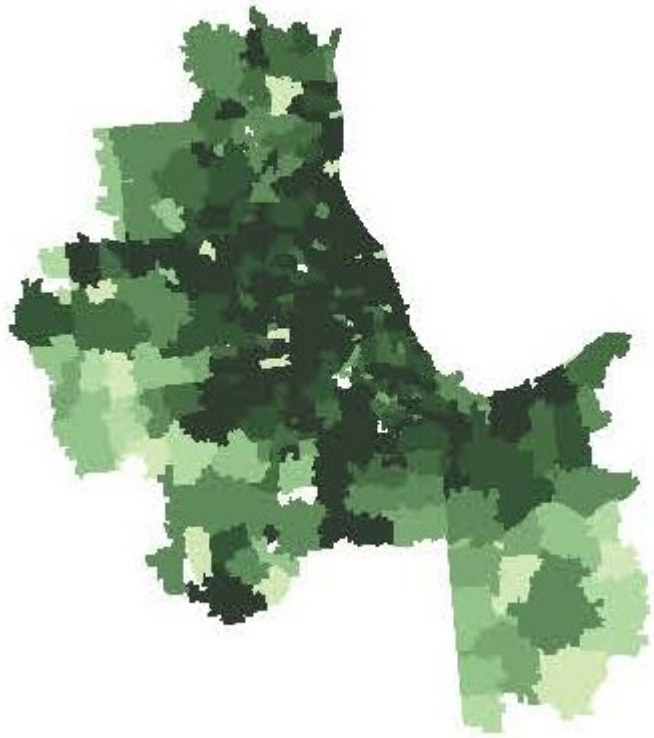
- Variation across the US
  - MCI and CCI are correlated but not perfectly.
  - Not correlated with broadband availability
  - Less visually obvious but also present: an urban-rural divide (more to come soon).



MCI & CCI for Chicago-Naperville-Elgin-IL-IN



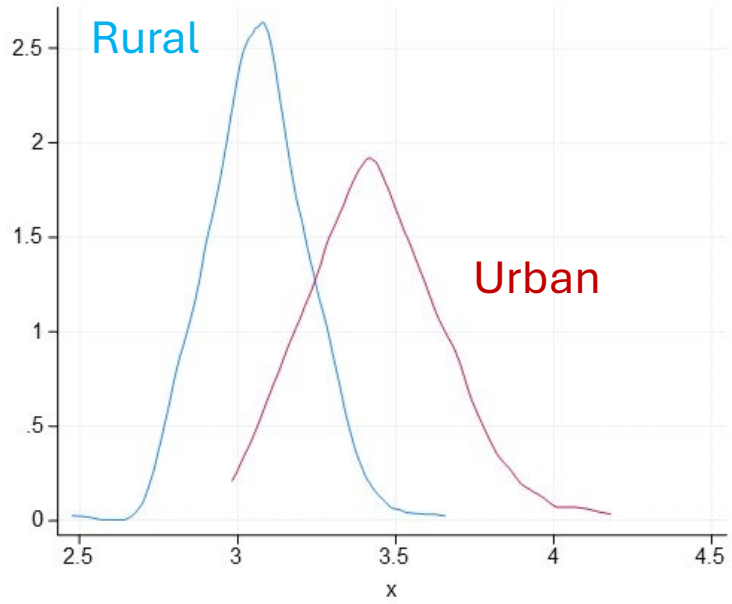
- Variation within a single urban area.
  - MCI and CCI correlated but not perfectly.
  - MCI and CCI are uncorrelated with broadband availability.



below 20pct 20pct 30pct 40pct 50pct 60pct 70pct 80pct

below 20pct 20pct 30pct 40pct 50pct 60pct 70pct 80pct

Statistical urban-rural divide



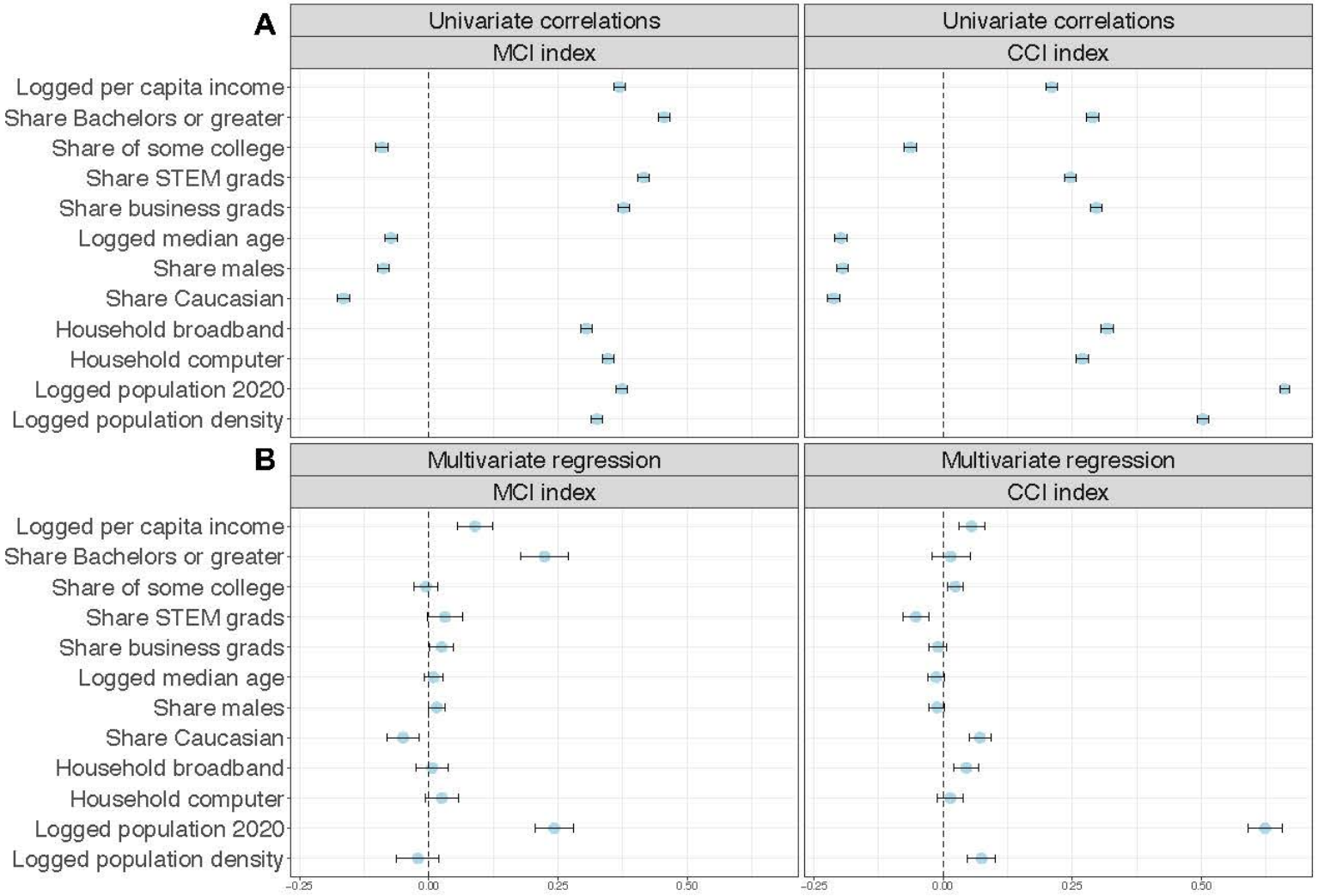
- Variation across the US
  - MCI and CCI are correlated but not perfectly.
  - Not correlated with broadband
- Urban-rural divide
  - Average urban MCI (CCI) is 0.19 (0.29), w/s.d. 0.87 (0.74)
  - Average rural MIC (CCI) is -0.27(-0.41) w/s.d. 1.05 (1.06)
- Translation: The level that marks the upper quartile for MCI (CCI) in urban zip codes is equivalent to the 37th (53.5th) percentile for rural households.
  - Implies extremely high (low) occurs in urban (rural)
  - Higher rural variance may also reflect greater heterogeneity in what it means to be rural, including vacation areas.

- Variation within an urban area.
  - Missed w/larger geographic boundaries.
  - MCI and CCI correlated but not perfectly.
  - MCI and CCI are uncorrelated with broadband.

Analysis of demographics

- Perform two statistical exercises.
  - Univariate correlations and (comparable) regression coefficients
  - Regression with transformed variables to mean zero and standard deviation of one.
- Exogenous variables and controls
  - Log per capita income
  - Share bachelor's or greater
  - Share of some college
  - Share STEM grads
  - Share business grads
  - Log median age
  - Share males
  - Share Caucasian
  - Household broadband availability
  - Household computer adoption
  - Log population in 2020
  - Log population density
  - State fixed effects

Univariate correlations are misleading.
Univariate and multivariate estimates differ.

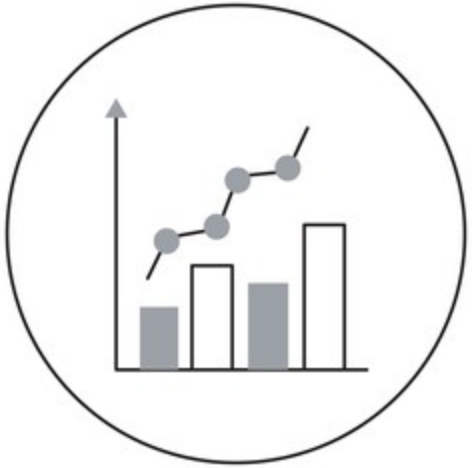


MCI: The regression reveals that the primary drivers are income, the share of bachelor's degrees, and population levels (a proxy for urban amenities).

CCI: The regression analysis reveals that the primary drivers are income, the share of Caucasians, and population levels, in particular.

Takeaways

- Variation in MCI and CCI indices of usage across the US
  - A noticeable urban-rural difference in average usage.
- Variation in usage within many metropolitan areas.
  - Not solely explained by the available broadband.
- Demographic determinants correlate with usage
  - Univariate correlations are misleading. We see differences between univariate and multivariate inferences.
  - MCI explained by education and income, and urban location
  - CCI explained by urban locations



Free dataset:
https://github.com/microsoft/Consumption-of-Digital-Applications-Data-Set.

Thanks

Thanks for your attention.

Extra slides

Association between category and application

Feature	Description	Applications
Media and Information Composite Index (MCI) features		
ss	Minutes of activity in software for visually organizing and analyzing tabular data	Excel, WPS Spreadsheets
dw	Minutes of activity in software for creating written content, such as reports and letters	Adobe Acrobat, Word, Wordpad
em	Minutes of activity in software for sending and receiving electronic messages	Thunderbird, Outlook, Mailbird
pt	Minutes of activity in software for organizing creating visual slideshows	PowerPoint, Canva, Prezi
md	Minutes of activity in software for playing and editing multimedia content	VLC player, Spotify, Windows Media Player
bw	Minutes of activity in software for browsing the internet	Firefox, Chrome, Opera, Edge
cc	Minutes of activity in software for communicating with others remotely	Zoom, Discord, Teams, Telegram
Content creation and computation composite index (CCI) features		
dv	Minutes of activity in software for creating, testing, and debugging code	Unity engine, Visual Studio, IntelliJ IDE
cd	Minutes of activity in software for remote file storage	One Drive, Dropbox, iCloud, Google drive
ut	Minutes of activity in software for system management and optimization	Remote desktop, printer management
sd	Minutes of activity in software that protects systems from cyberthreats	McAfee agent, Norton Security, Bitdefender
ct	Minutes of activity in software for creating and editing digital media	Adobe Photoshop, Virtual DJ, Blender

Summary of the procedure in algebra

Index calculation For each device, we compute the weighted sum of the time spent across different applications. The *Media and Information composite index* MCI is defined as:

$$MCI = \sum W_i * (X_i), i = [ss, dw, em, pt, md, bw, cc]. \quad (1)$$

where W_i are the weights used to aggregate across different digital applications. For each device in the data, we collect the indicators described in Table 1. The *Content creation and computational composite index* CCI is defined as:

$$CCI = \sum W_i * (X_i), i = [dv, cd, ut, sd, ct]. \quad (2)$$

```
pca = dp.sklearn.PCA(
epsilon=0.5,
row_norm=(13*np.sqrt(7)),
n_samples=num_rows,
n_features=7
)
```

MCI		CCI	
ss	0.211	dv	0.138
dw	0.263	cd	0.209
em	0.132	ut	0.293
pt	0.097	sd	0.186
md	0.082	ct	0.170
bw	0.145		
cc	0.067		

Zip code level aggregations For each zip code, we compute the weighted average of indices MCI and CCI. Let k be the total number of devices in postal code. We denote by T_{m_j} the total time device j spends in media and information applications, i.e, $T_{m_j} = ss + dw + em + pt + md + bw + cc$. Analogously, we denote by T_{c_j} the total time device j spends in content creation and computational applications, i.e, $T_{c_j} = dv + cd + ut + sd + ct$. The MCI and CCI zip code-level average are computed as follows:

$$\overline{\text{MCI}} = \frac{\sum_1^k (MCI_j \cdot T_{m_j})}{\sum_1^k T_{m_j}} \quad (3)$$

and,

$$\overline{\text{CCI}} = \frac{\sum_1^k (CCI_j \cdot T_{c_j})}{\sum_1^k T_{c_j}}. \quad (4)$$

```
def laplace_noise(agg, eps, sentvt):

    agg = agg.astype(np.float64, copy=False)
    # call the constructor to produce the measurement `base_lap`
    base_lap_vec =  make_base_laplace(
    scale= eps, D="VectorDomain<AtomDomain<float>>"
    )

    # invoke the measurement on some aggregate x, to sample Laplace(x, 1.) noise
    aggregated = agg
    # we must know the sensitivity of `aggregated` to determine epsilon
    sensitivity = sentvt
    epsilon =  base_lap_vec.map(d_in=sensitivity)
    output = base_lap_vec(aggregated)
    return output
```