DESIGNSAFE-CI

A NATURAL HAZARDS ENGINEERING COMMUNITY

METRICS TO INCREASE DATA USAGE UNDERSTANDING AND TRANSPARENCY

Maria Esteva Joshua Freeze James Carson Craig Jansen

Metricas de impacto en repositorios: citas y uso

Maria Esteva, 26 de julio 2024

RDA-UNR

Citas

- https://summit.makedatacount.org/
- DataCite corpus. http://corpus.datacite.org/dashboard
- Data Citation Index
- Dimensions AI

Motivation

- Metrics are key to repositories and their communities.
- Make Data Count initiative standardized metrics across datasets and repositories; COUNTER provides their Code of Practice; DataCite provides infrastructure.
- Implementing & analyzing metrics is complex & contextual.
- Many repositores lack a method to easily compare usage over time across datasets.

What is DesignSafe?

- Cyberinfrastructure of the US National Science Foundation-funded Natural Hazard Engineering Research Infrastructure.
- A web-based research platform that provides tools to manage, analyze, and understand critical data for natural hazards research.
- Data Depot: open access natural hazards data repository where researchers curate, publish, discover, and access natural hazards research datasets.

DesignSafe Data Depot Repository Diverse User Community

- Hazards: Earthquakes, Hurricanes, Tornadoes, Storm Surges, Drought, Wildfires and more.
- Data Types: Experiments, Simulations, Field Research, Other.
- Disciplines: Engineers, Social Scientists.

DATA METRICS: DISCUSSIONS AND INITIATIVES

- Data citations remain the most valued by the research community.
- Usage metrics are seen as complementary.
  - MDC enabled reliable data usage counting
  - https://search.dataone.org/profile
  - Most viewed
- Data metrics can be "gamed." (Borgman, 2022)

  Careful assessment of how they are obtained.
- Meaningful Data Count.

  Set path to new analyses of citations and usage with DataCite metadata.

This project analyses usage within the repository's designated community, in relation to time and to contextual metadata gathered about each dataset, adding value to usage metrics.

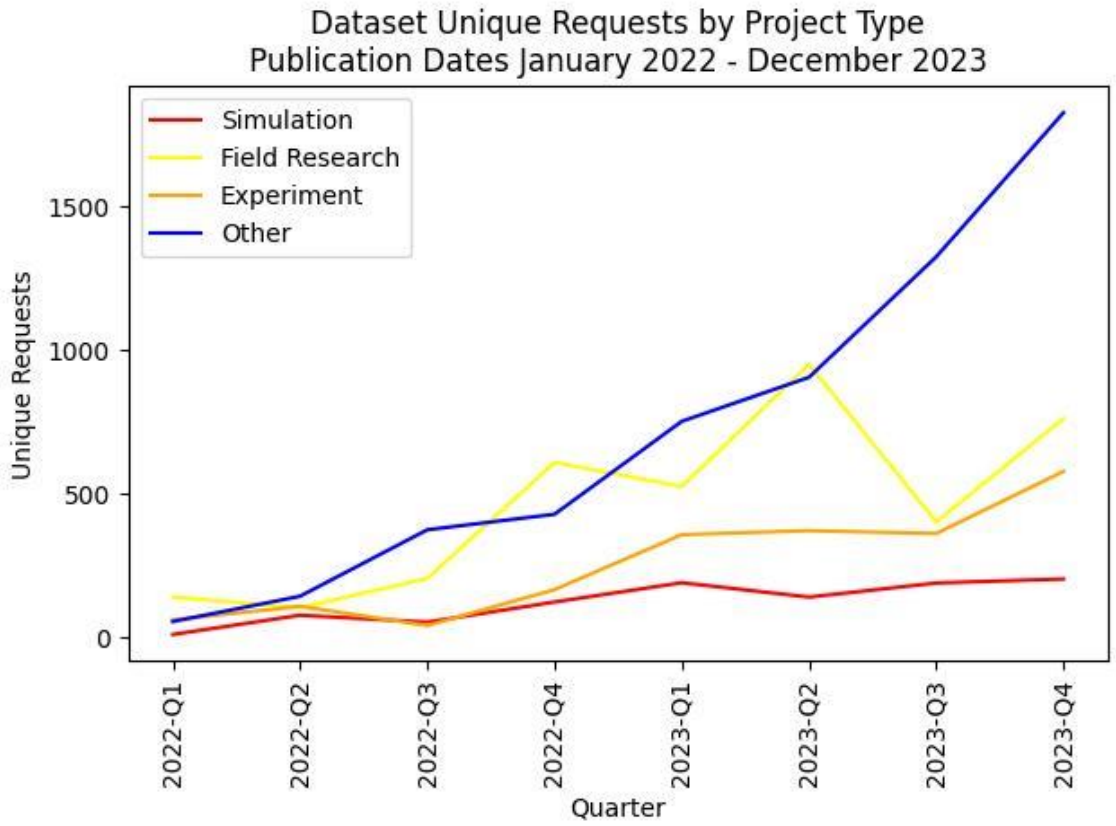
Primary Topics

- Metrics & Usage over time.
  - Usage: context.
  - Usage: benchmarking.
  - Usage: grouping analysis.
- Methods of sharing data with users.

Principal Metric: Unique Request

- Definition: one-hour sessions during which a user previewed, downloaded, or copied files associated with a DOI.
  - Clarification: "Unique Requests" & "Downloads."
    - Available in the DDR since January 2022.
  - Sources: Make Data Count, DataCite, Project Counter.

Usage by Data Type



Tenemos metadatos asociados a cada dataset – tipo de dataset y tipo de desastre natural.

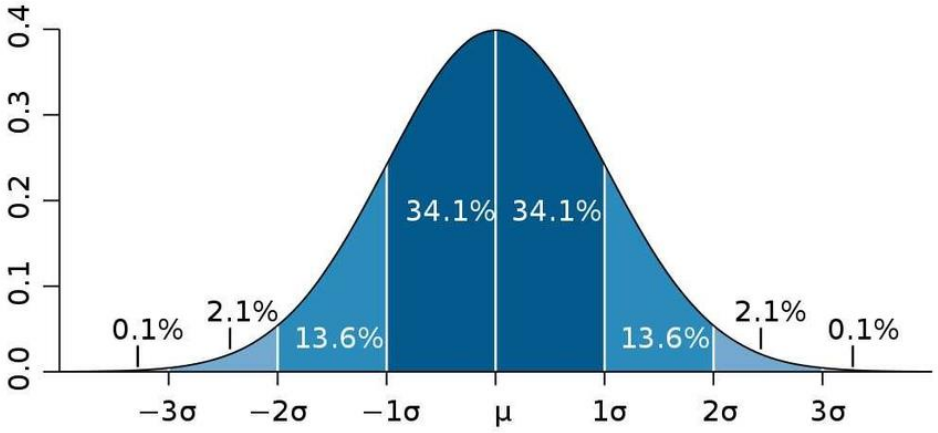
Benchmarking

- Provide a gauge to users. Punto de referencia
- Unique requests for all datasets for a specific period of time after publication.
- Define low, medium and high usage based on usage of all published datasets in the DDR.
- As more time elapses since we have UR data, we will provide benchmarking for longer time-frames and include more datasets in each time-frame.

Benchmarking

- Histogram for all datasets publicly available for at least the time-frame being benchmarked.
  - Data for very high usage is sparse.
    - Datasets with over 100 URs Defined as outliers.
      - Excluded from benchmarking analysis.
- Use tool from statistics of the “normal distribution.”
  Distribution is not quite normal, but we use the percentages for plus or minus one standard deviation from the mean to define low, medium and high.
  - Ahora log normal

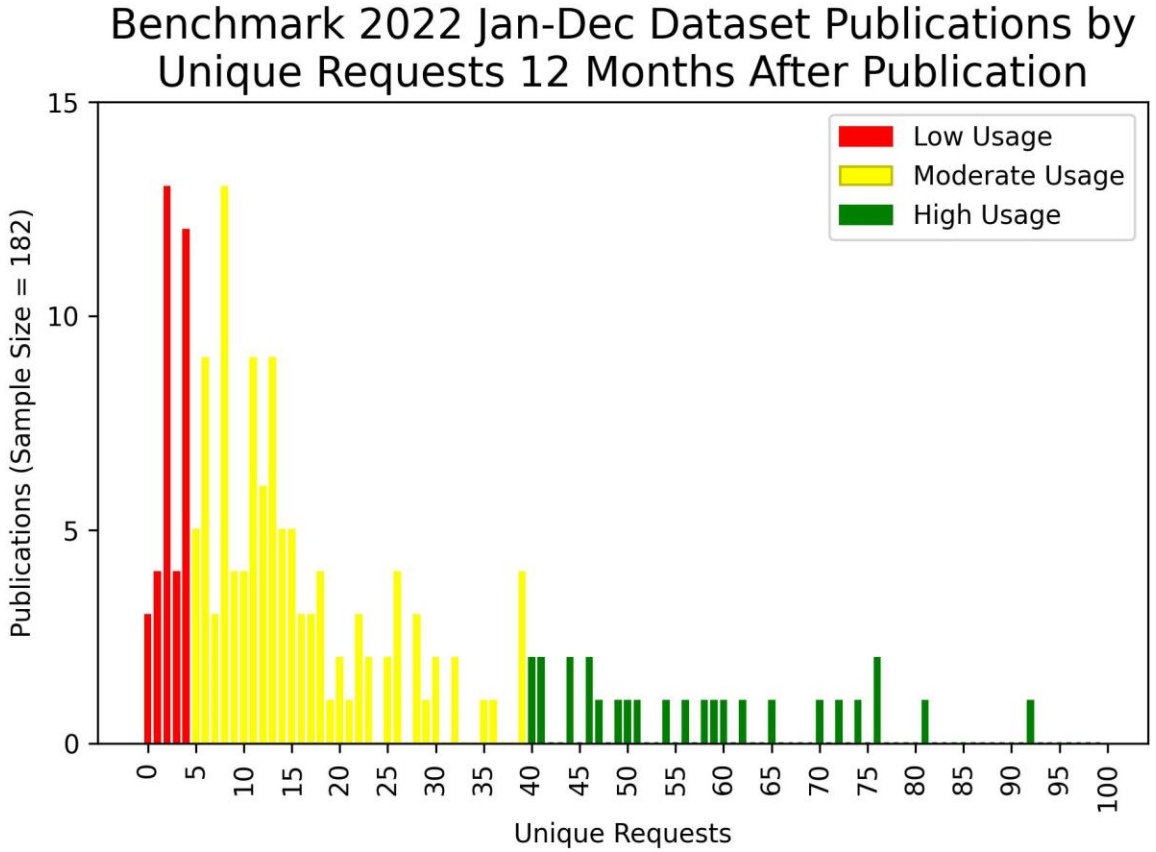
Standard Deviation



- Measure of how widely spread random data is.
- The peak is the mean of the data.
- One standard deviation (denoted by sigma: σ) about the mean includes 68.2% of the data.

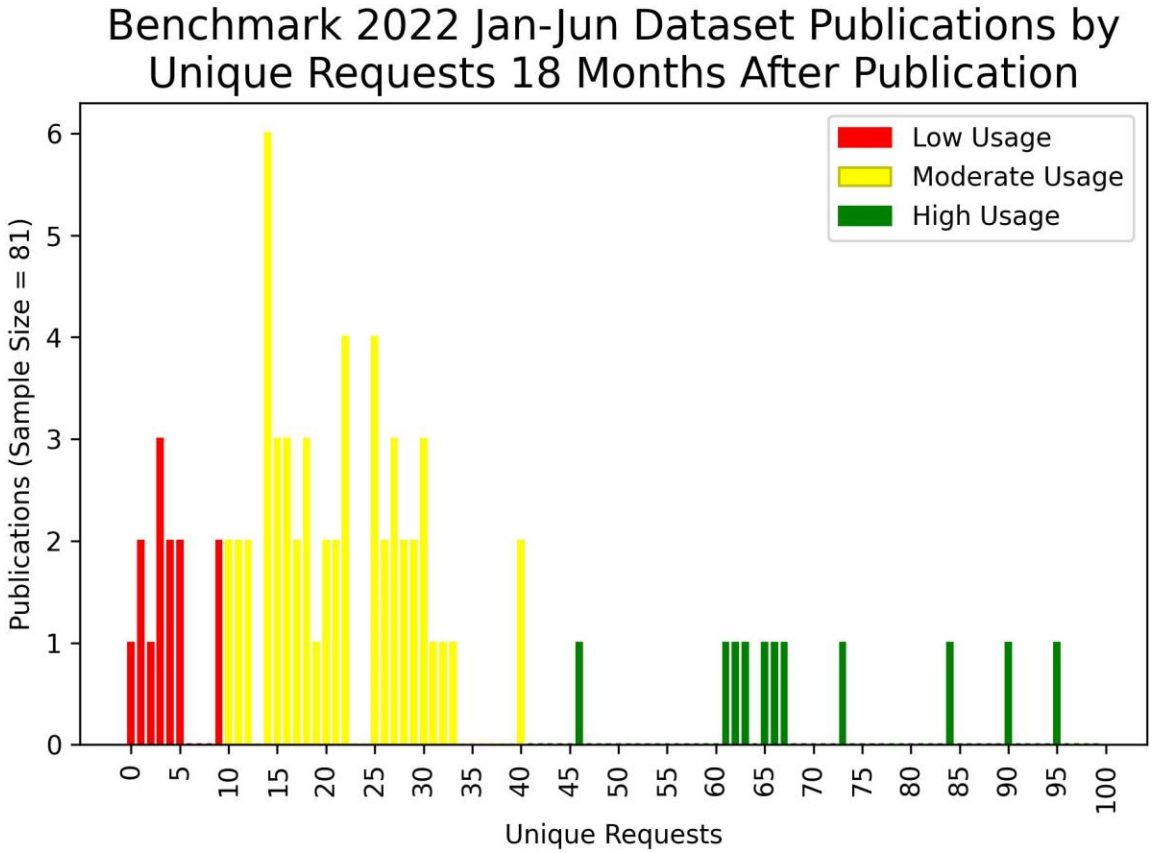
Image: Explained Sigma. MIT News, https://news.mit.edu/2012/explained-sigma-0209

Benchmarking



- Medium: usage between five and 39 URs inclusive during the 12-month period.
- Low: four or fewer URs.
- High: 40 or more URs.
- Boundaries will vary for other repositories using the same methodology.

Benchmarking

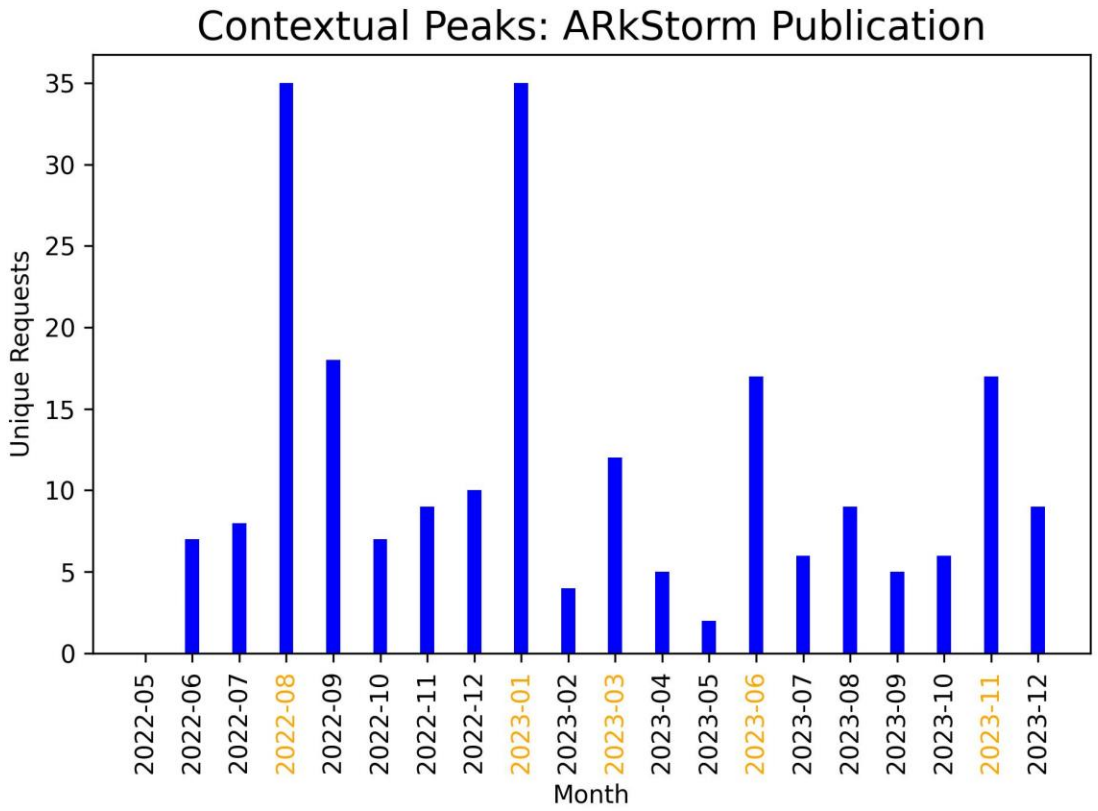


- Medium: usage between ten and 40 URs inclusive during the 18-month period.
- Low: nine or fewer URs.
- High: over 40 URs.
- Boundaries will vary for other repositories using the same methodology.

Benchmarking

- As more time elapses since for which we have UR data, we will provide benchmarking for 12, 18 and 24 months after publication.
- This will also permit us to include more datasets for each set of benchmarks, so boundaries between categories may shift over time.

Benchmarking



Outliers: Exceptionally High Usage

- Look for explanations..
- Analysis of the most used dataset
- External events: featured in two New York Times articles..
- Won DesignSafe dataset award
- Spikes in usage may also appear due to significant natural hazard event.

Bin Analysis: Usage over Time

- Key for data policies and sustainability strategies.
- Questions
  - When does usage peak?
  - How long does usage persist?
  - Do the answers vary by data type or domain?
- Criteria devised to classify usage patterns
  - Timing of initial peak usage.
  - Length of usage tail after the initial peak.

Bin Analysis: Hypotheses

- Most datasets would have an early peak with little sustained usage.
- Only a few high-profile datasets would see continued usage.
- Do the answers vary by data type or domain?

Bin Analysis: Methodology

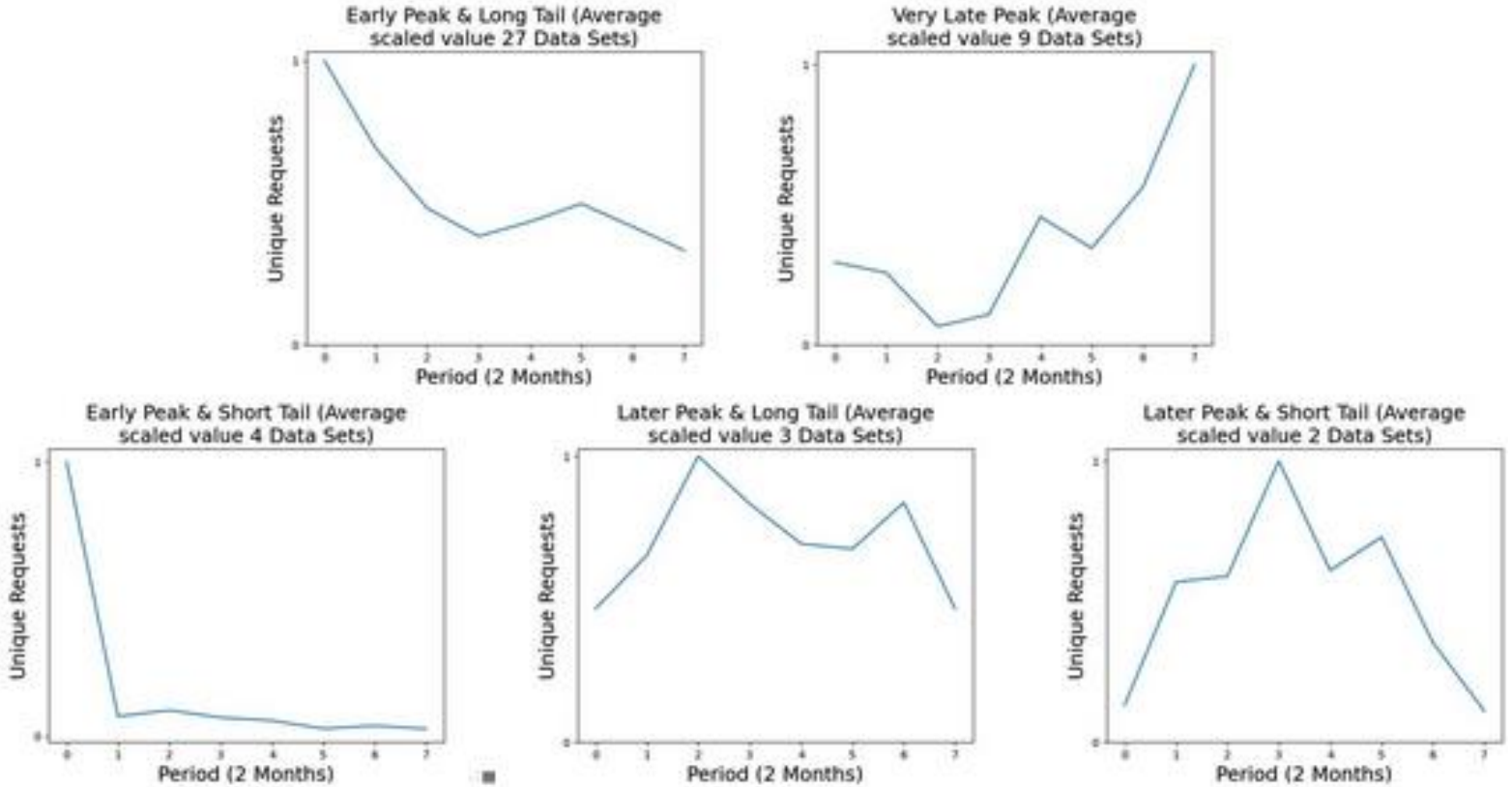
- Cohort: DDR datasets published January-June 2022.
  - Study period: 15 months following publication.
    - Exclude datasets with low (< 15 UR) usage.
- Create two-month periods to smooth high monthly variability.

Bin Analysis: Methodology

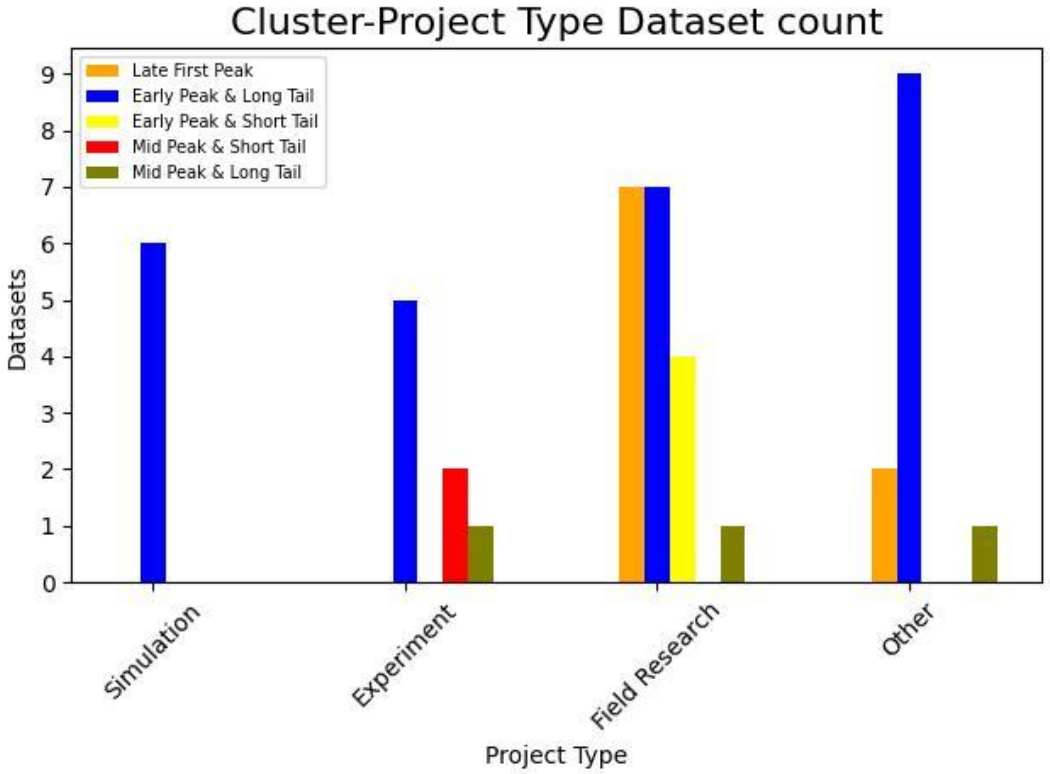
Bin parameter definitions

- Early Peak: occurs in the first two periods.
- Long Tail: significant usage in the third or later period after the first peak
- Significant: 20% of the peak height, at least 2.
- Bins: early-long, early-short, later-long, later-short.
- Fifth bin: accommodate peaks in last three periods.

Bin Analysis: Overall Results

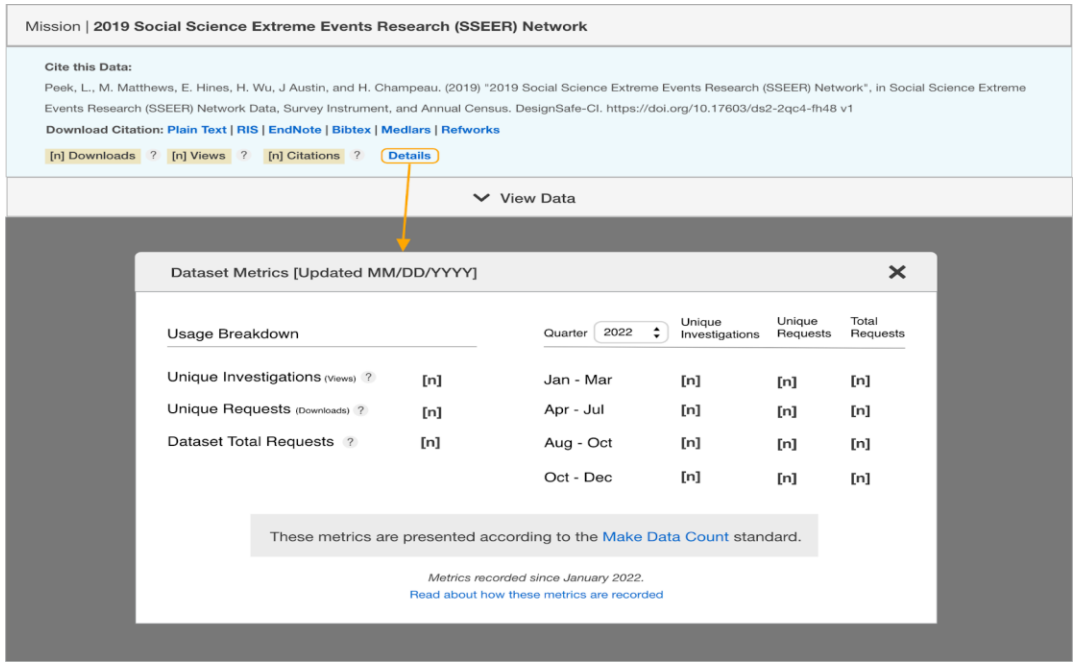


Bin Analysis: Dataset Divisions



- Dataset type, Natural Hazard, Discipline.
- Dataset Type: Simulation, Experiment, Field Research, Other.
- Early peak-long tail balanced between dataset types.
- Other ways of dividing datasets:
  - Wind & Quake balanced.
  - Engineering > Social Science.

Communicating Metrics



- Sharing metrics builds trust with users.
- Modal window to share View & Download figures.
- Defines metrics.
- Presented over time.
- Link to Users Guide for details.
- Plan to make benchmarks available as well.

Conclusions & Looking Forward

- Communities influence usage patterns.
  - Influence of external factors.
  - Most datasets had long tails.
- May add bin classification criteria.
- Continue seeking user feedback.
- https://en.wikipedia.org/wiki/Log-normal_distribution
  - Percentiles - comparacion

Metricas de Uso

- https://en.wikipedia.org/wiki/Log-normal_distribution

- https://search.dataone.org/profile

- https://zenodo.org/communities/meaningfuldatacounts/records?q=&l=list&p=1&s=10&sort=newest
- Most viewed https://zenodo.org/search?q=&l=list&p=1&s=10&sort=mostviewed