IJCERT

International Journal of Computer Engineering in Research Trends

e-ISSN: 2349-7084 | Volume 13, Issue No: 3, July–September 2026, Pp: 26–34 © 2026

Original Research Article

Kumbh 2015 Telecom Dataset for Reproducible Mass Gathering Research

Lavanya Addepalli[1], Sandip Kashinath Shinde[2], Sachin Raghunath Pachorkar[3], Girish Pandit Pagare[4], Jaime Lloret[5]

[1]Instituto de Investigación para la Gestión Integrada de Zonas Costeras (IGIC), Universitat Politècnica de València, Spain; Email: phani.lav@gmail.com; ORCID: 0000-0002-2651-163X

[2]Director, Kumbhathon Innovation Foundation, India; Email: sandipkshinde@gmail.com

[3]Founding Member and Advisor, Kumbhathon Innovation Foundation, India; Email: sachinpachorkar@gmail.com

[4]Director, Kumbhathon Innovation Foundation, India; Email: gppagare@gmail.com

[5]Instituto de Investigación para la Gestión Integrada de Zonas Costeras (IGIC), Universitat Politècnica de València, Spain; Email: jlloret@dcom.upv.es; ORCID: 0000-0002-0862-0533

Corresponding author: phani.lav@gmail.com

Abstract—This paper describes telecom data from a panel of 194,520 telecom records across eight telecom operators and 1,621 operator-site pairs, collected over five nonconsecutive dates, for Kumbh 2015. An executable Python pipeline preserves 169,932 counts where observed, restores previous estimates for 715 counts estimated from a preceding three observed hours, and sets 23,873 unsupported gaps to missing. Counts are independent of the units and extremes in which they were observed. Temporal analysis reveals that two operators have constant counts within a site-day and that 99.15 percent of adjacent observed pairs for a third operator are also unchanged. The value of measurement freshness is therefore not inherent in the hourly grid. An expanding-date benchmark evaluates persistence, median hourly targets across past periods, and pooled log-ridge regression applied to 125,169 hourly target observations. Although persistence produces the lowest pooled mean absolute error for each test date, this information is at the pooled level. It cannot be used to assess operator differences or the effect of unchanged histories. The contribution is a documented event dataset, transparent preprocessing, and reference benchmarks for reuse.

Keywords—mass gatherings; telecom aggregates; data validation; missing data; forecasting

Received: 03-06-2026; Revised: 19-08-2026; Accepted: 18-09-2026; Published: 30-09-2026

DOI: 10.22362/ijcert/2026/v13/i3/v13i3003

1 Introduction

Mass gatherings create imbalances in demand for transportation, support, and temporary measures. Empirical crowd research suggests that flow transitions at the crowd level affect crowd safety [1]. Interpretation and calibration of mobile network records can be useful for population mapping [2], while for multi-operator use, explicit coverage and measurement semantics are needed [3]. Hourly subscriber identity module (SIM) totals do not reflect the level of crowd pressure or the number of unique visitors to an event in the local area. The research question is: What is the best way to document, validate, and benchmark the Kumbh 2015 panel, whereby recording the numbers and making estimates must be kept separate? Revo, re-use, re-discovery, and re-publication of data and metadata [4] are undervalued within the Findable, Accessible, Interoperable, and Reusable (FAIR) principles, as is documenting the collection, composition, and intended uses via datasheets [5]. This paper provides empirical information and an evaluation applicable in the classroom.

Firstly, the audit serves to obtain an integrated row key, completeness, and Source labels from Individually Verified Measures. Second, it describes the temporal stability and heterogeneity of operators hidden in pooled record counts. Third, an observed-only chronological benchmark offers insights into the effects of equal coverage and stable series while offering reference points for the observed time.

The contribution not only provides event-specific multi-operator data but also explicitly labels the imputation and provides a repeatable check. It is not based on any so-far unique large dataset or new forecasting algorithm, etc.

2 Related Work

The Data for Development (D4D) challenge presents call detail record (CDR)-based communication traffic and sample mobility data products [6]. These types of communication events and trajectories are not the same as those of a connected subscriber identity module total. Milan-Trentino delivers 10-minute grid-based call, short message service (SMS), and Internet data spanning 2 months, as well as weather and other urban data [7]. It enables multimodal urban analysis, whereas the current panel includes only operator-site hourly levels, observation status, and imputation labels for five mass-gathering dates. Notwithstanding difficulties in direct comparisons of measurement error, using different units and aggregations at different time levels yields distinct forecast error values.

Since 2013, the Data for Development Senegal challenge has provided hourly antenna-to-antenna traffic and sampled mobility products for 1,666 antennas [8]. NetMob23 generates datasets for 68 mobile services across 20 French metropolitan areas at 100 m spatial resolution and 15 min temporal resolution for 77 consecutive days in 2019 [9]. One of these resources has a longer history or more detailed information regarding space. Do Kumbh lack any tools other than eight-operator reporting heterogeneity, even-specific coverage, and explicitly estimated value labels? For its contribution, it does not offer increased scale or resolution, but rather auditing of this unique measurement product and an observed-only benchmark. A secondary methodological context for documenting acquisition and discriminating statistical flags from Labeled Anomalies comes from the implementation of a Smart-metering architecture [10] and online network anomaly detection [11]. Nor are they similar data, telecom, or proof related to this panel's novelty.

The graph-neural-network imputation method in [12] fills incomplete multivariate series. For reconstruction, there should be defensible spatial relations and a referent point. The deterministic trailing-mean rule provides explicitly labeled estimates, and observed-labeled histories and targets are used for forecasting. Without known reference values, no claim is made regarding the degree of reconstruction accuracy.

3 Methodology

The methodology maps and validates the provided telecom panel, including explicitly reconstructing missing values, to make it a versioned dataset. Chronological forecasting benchmark represents re-use. Count entries remain in original subscriber identity module (SIM) units; the minimum and maximum observed counts are retained.

3.1 Problem Statement

Using the provided consolidated panel D, create a version of the panel with distinct operator/site/time keys, consistent observed counts, and different estimated and unresolved values. Do the following: If a gap is filled, the three preceding hours must be observed on the same operator-site-day; Evaluate the forecast without using future information.

Let s, d, and h be (Operator, `SITE_ID`), Event Date, and Hour (0–23). `COUNT` denotes the number of subscriber identity modules attached to the tower for the reported hour. Each subscriber identity module is treated under the stated one-person–one-phone–one-module assumption and is linked to only one tower per snapshot. Any movement within the one hour will not be counted as unique attendance, and the sums of the towers will be added to the results of the preceding hour. Let $x(s,d,h)$ be an observed total and $m(s,d,h)$ the indicator for an `OBSERVED` label.

3.2 Proposed Framework

The team reports that operators supplied telecom reports by email in 2015. The executable workflow starts from the five supplied consolidated comma-separated-value (CSV) files; it does not reconstruct those email attachments. Figure 1 links this input to Python validation, mapping, missing-value reconstruction, quality assessment, and reference forecasting. Source files are preserved, and each changed row is logged.

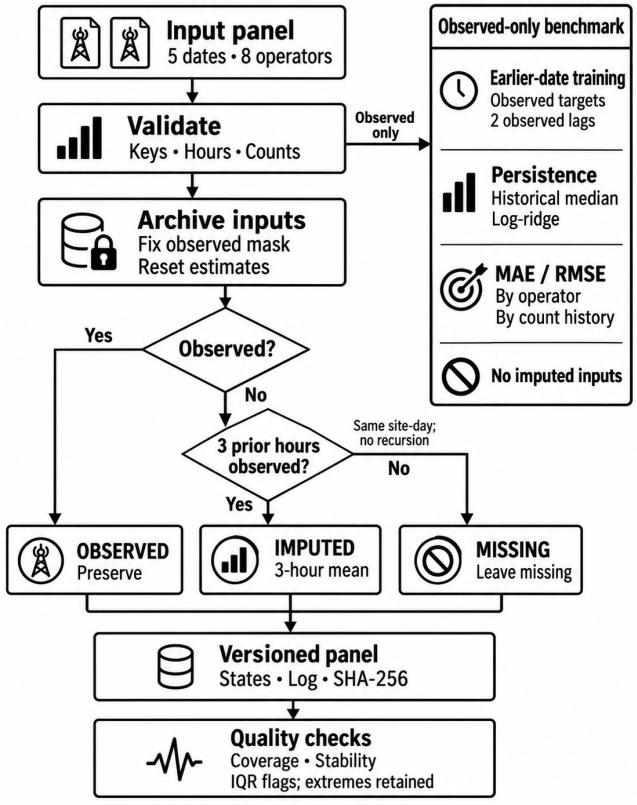

Figure 1. Executed preprocessing and reference-evaluation framework. Processing starts from the supplied consolidated panel. Unsupported gaps remain missing; observed counts are preserved.

3.2.1 *Acquisition and Data Mapping*

A row represents one observation, and columns represent variables, following tidy data principles [13]. Mapping ϕ associates each report record r with one operator-site-time key and metadata g. For the canonical panel, require $|\{k(r) : r \in D\}| = |D|$; `SITE_ID` alone is insufficient because operators can reuse identifiers. Equations (1) and (2) define the mapped record and its composite key, respectively.

$$\phi(r) = (s,d,h,x,m,g,\text{source}), \tag{1}$$

$$k(r) = (s,d,h). \tag{2}$$

3.2.2 *Cleaning and Retaining Extremes*

The metadata tuple g stores coordinate information and region/route labels. The number of count values observed must be finite, an integer, and nonnegative. Each composite key needs to be unique; the timestamps need to match the Date/Hour field; and each site-day needs to have all hours 0–23. If there are any invalid keys, times, or observed counts, processing stops, and any missing metadata is output. Zero is valid. Estimates on a fractional-imputation basis may be allowed.

For each operator o, estimate $Q_{1,o}$ and $Q_{3,o}$ from observed-labeled counts using linearly interpolated quantiles. Define the interquartile range (IQR) as $\text{IQR}_o = Q_{3,o} - Q_{1,o}$, then set $L_o = Q_{1,o} - 1.5\,\text{IQR}_o$ and $U_o = Q_{3,o} + 1.5\,\text{IQR}_o$. The

flag $a_i = \mathbf{1}\{x_i < L_o \text{ or } x_i > U_o\}$ retains x_i. These all-date descriptive fences never enter forecasting. This is the standard 1.5-interquartile-range (IQR) fence convention [14], used for inspection rather than deletion.

3.2.3 Missing-Value Reconstruction

First, archive the input counts, status, and method labels; then reset all inherited `IMPUTED` values to missing. Equation (3) preserves each observed value. Equation (4) applies only when $m(s,d,h) = 0$, $h \geq 3$, and $m(s,d,h-j) = 1$ for every $j \in \{1,2,3\}$:

$$\widetilde{x}(s,d,h) = x(s,d,h), \qquad m(s,d,h) = 1, \tag{3}$$

$$\widetilde{x}(s,d,h) = \frac{1}{3}\sum_{j=1}^{3} x(s,d,h-j), \qquad \text{eligible gap.} \tag{4}$$

Otherwise, the output is still missing. The indicator m is always for input `OBSERVED` labels, and estimates are never calculated as input to subsequent fills. There is no sharing between operator, location, and date. `OUTPUT_COUNT`, `OUTPUT_STATUS`, and `OUTPUT_METHOD` retain the values and labels from the inputs, using `OBSERVED`, `IMPUTED`, and `MISSING` states. All means are not rounded to integers or normalized. This is not a reconstruction of the original manual preparation; it is a preprocessing decision that has been implemented.

3.2.4 Forecasting as a Reuse Benchmark

The expanding-date evaluation follows the out-of-sample testing rationale in [15]. Let $e_i = \mathbf{1}\{h \geq 2\}\, m(s,d,h)\, m(s,d,h-1)\, m(s,d,h-2)$. For test date d_k, $H_k = \{i : d_i < d_k,\ m_i = 1\}$ is observed history, $T_k = \{i \in H_k : e_i = 1\}$ is ridge training data, and $V_k = \{i : d_i = d_k,\ e_i = 1\}$ is the common test set. Lags stay within a site-day; $k = 2,\ldots,5$.

Persistence predicts $\widehat{y}_i = x(s,d,h-1)$. Historical medians use H_k, matching operator-site-hour before the specified fallbacks. Ridge regression [16] uses the feature vector $z_i = [1, \ln(1+x(h-1)), \ln(1+x(h-2)), \sin(2\pi h/24), \cos(2\pi h/24)]^T$. Equation (5) defines the penalized objective, and Equation (6) converts the fitted output to subscriber identity module count units:

$$\beta^* = \arg\min_{\beta} \left\{ \sum_{i \in T_k} \left[\ln(1+x_i) - z_i^T \beta\right]^2 + \sum_{j=1}^{4} \beta_j^2 \right\}, \tag{5}$$

$$\widehat{y}_i = \exp\left(\text{clip}(z_i^T \beta^*, 0, 25)\right) - 1. \tag{6}$$

Equivalently, solve $(Z^T Z + \Lambda)\beta^* = Z^T y$, where Z stacks z_i^T, $y_i = \ln(1+x_i)$ and $\Lambda = \text{diag}(0,1,1,1,1)$. Thus, $\lambda = 1$, and the intercept is unpenalized. Predictions are returned to the original subscriber identity module units; the released data are not transformed.

3.3 Framework Algorithm

Algorithm 1 specifies validation, reconstruction, and evaluation. The observed mask is fixed before any estimate is calculated.

Source comma-separated-value (CSV) files remain unchanged; Secure Hash Algorithm 256-bit (SHA-256) hashes identify inputs and the derived panel. The change log records every replaced inherited estimate.

Algorithm 1 Versioned preprocessing and evaluation

Require: Supplied panel D; ordered event dates $d_1,\ldots,d_5$
Ensure: Processed panel P; change log A; scores S

```
Validate keys, time grid, observed counts, and labels
if any required validation fails then
    halt
P ← copy of D; archive input counts and labels
Fix mask m from the input OBSERVED labels
Reset inherited estimates in P to missing
for each operator-site-day (s,d) do
    for h = 0,...,23 do
        if m(s,d,h) = 1 then
            Preserve x(s,d,h); label OBSERVED
        else if h ≥ 3 and the three prior hours are all observed in m then
            Set count to their arithmetic mean
            Label IMPUTED
        else
            Retain missing count; label MISSING
Flag operator-specific observed IQR extremes
Write P, change log A, and input/output hashes
S ← ∅
for each chronological test date d_2,...,d_5 do
    Fit baselines using earlier observed data only
    Score common observed targets with observed lags
    Append the resulting scores to S
return P,A,S
```

3.4 Dataset Details

Table 1 describes the five daily sequences and tower-level hourly SIM totals. Person counts are an interpretation under the stated phone/subscriber identity module assumption, not independently measured attendance.

Table 1. Dataset and mapping specification

Item	Specification
Input / derived output	5 CSV (19 fields) / 1 CSV (23 fields); 194,520 rows
Event dates in 2015	26, 29 Aug; 13, 18, 25 Sep
Time grid	24 hours/date; 38,904 rows/date
Operators / operator-sites	8 / 1,621
Distinct `SITE_ID` strings	1,619; use operator-site key
Observed / filled / missing	169,932 / 715 / 23,873
Identity mapping	`TELECOM` + `SITE_ID`; `SITE_NAME`, `LAC`
Time mapping	`OBSERVED_AT` ↔ `EVENT_DATE` + `HOUR`
Spatial metadata	`LATITUDE`, `LONGITUDE`, `REGION`, `TALUKA`
Route metadata	`SUB_ROUTE`, `HIGHWAY_ROUTE`
Value fields	`COUNT`; `REFERENCE_COUNT` excluded
Provenance fields	Status, source and method; archived `INPUT_COUNT`, `INPUT_STATUS`, `INPUT_METHOD`; `OUTLIER_IQR`

3.5 Experimental Setup

Mean absolute error (MAE) and root mean square error (RMSE) retain the original scale and complementary loss sensitivities [17]. For test errors $\varepsilon_i = \widehat{y}_i - x_i$ and N eligible rows, Equation (7) gives MAE and Equation (8) gives RMSE

in original subscriber identity module count units:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\varepsilon_i|, \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2}. \tag{8}$$

For any group G, observed coverage is $C(G) = \sum_{i\in G} m_i/|G|$. Temporal stability $R(G)$ is the fraction of equal adjacent counts among pairs with both labels observed, within the same site-day. A zero denominator makes either ratio undefined.

For fold k, high-count scores use $P_k = \{i \in V_k : x_i \geq Q_{0.90}(H_k)\}$; the quantile uses count values in H_k. Operator MAE divides pooled absolute-error sums by pooled eligible rows, not by the number of folds. No independent-row uncertainty claim is made. Scripts and fixed settings support computational traceability [18]. Training-only fitting addresses temporal leakage [19]. Table 2 summarises the experimental settings, and Table 3 lists the chronological training and test folds.

Table 2. Experimental settings

Setting	Specification
Inputs and targets	`OBSERVED` only; identical test rows
Historical fallback	Site median → operator-hour median → global training median
Ridge penalty/tuning	$\lambda = 1$; no test-driven tuning
High-count subset	$x \geq$ training observed 90th percentile
Operator MAE	Sum absolute errors / total eligible rows
Temporal stability	Equal adjacent/eligible observed pairs within site-day
Software	Python 3.12.14; pandas 3.0.1; NumPy 2.3.5
Reproduction outputs	Audit and baseline scripts, CSV scores, source SHA-256 hashes

Table 3. Chronological evaluation folds

Test date	Training dates	Targets
29 Aug	26 Aug	32,089
13 Sep	26, 29 Aug	31,716
18 Sep	Earlier 3 dates	30,014
25 Sep	Earlier 4 dates	31,350

For Geographic Points, it provides a longitude and latitude, but not jittered; for one site, it does not provide latitude, so the site is not included. Reporting classes are a class type of combinations associated with 100% observed, mixed, or no observed site histories. Absolute hourly changes are calculated for within-site-day pairs with both endpoints observed. Symlog only affects the axis display, not the contents of the file. An amplified error measure is expressed as the root-mean-square-error-to-mean-absolute-error ratio. The history-stratified evaluation process modifies the eligible test folds to ensure they are evaluated on an even number of observations from the previous 2 counts. Models and targets stay constant; instead, MAE is "pooled" using a weighting value for each group based on the number of applicable years in the record. There are no target values for the group membership.

The framework generates a three-state panel that is traceable, retains observed extremes, and examines 125,169 chronological targets for reuse. The imputation rule is deterministic, can be performed independently, and the evidence relates to dataset quality and/or reference performance.

4 Results

The results of the chronologically forecasting are given after the processed panel is protected of the covering, provenance, and temporal properties. Data availability, temporal stability, comparisons between operators, recent count histories, and high-count conditions all examine how data affects benchmark performance.

4.1 Coverage and Provenance

Of 194,520 rows, 169,932 (87.36%) retain observed counts, 715 (0.37%) receive new estimates, and 23,873 (12.27%) remain missing. 12.64% of the input is provided by non-observed. Nevertheless, there are significant differences among operators (Figure 2). Bharat Sanchar Nigam Limited (BSNL) has the highest percentage of non-observed input (65.32%), while Airtel, Idea, and Uninor have 0% non-observed input. There are 3 rows: observed, filled, and unresolved (Table 4).

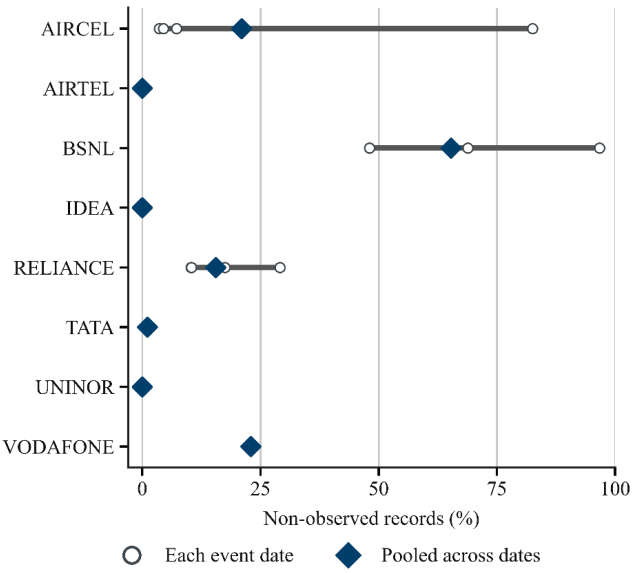

Figure 2. Operator reporting gaps. Circles show event dates, diamonds show pooled non-observed shares, and horizontal ranges span five dates. Exact observed, filled, and missing counts are reported separately in Table 4.

A total of 715 out of 24,588 inherited estimates meet the three-observed-hour requirement (2.91%). These values are redone, while the remaining 23,873 are "no value". 435 of the new estimates are fractional. The conservative policy is to make unsupported values explicit but not to specify a level of accuracy for filled values. No ground truth was used to compare the missing values.

Table 4. Processed record states by operator

Operator	Observed	Filled	Missing
Aircel	10,615	138	2,687
Airtel	25,920	0	0
BSNL	3,829	228	6,983
Idea	25,320	0	0
Reliance	38,008	349	6,643
Tata	22,800	0	240
Uninor	18,840	0	0
Vodafone	24,600	0	7,320

4.1.1 Coverage Across Event Dates

Figure 3 maps operator-site reporting status and daily observation coverage. Table 5 reports observed, filled, and unresolved records for each date.

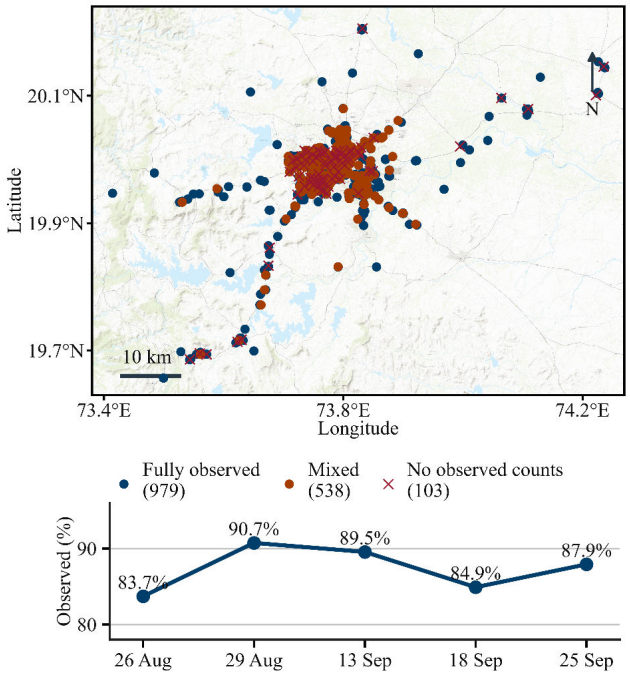

Figure 3. Geographic reporting coverage and daily observed shares. Unjittered points show 1,620 operator-sites; one site lacks latitude. Basemap: Esri World Topographic Map, retrieved 16 September 2026; not a reconstruction of 2015 geography. Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), © OpenStreetMap contributors, and the GIS User Community.

Table 5. Daily provenance composition

Date	Observed	Filled	Missing
08-26	32,563	208	6,133
08-29	35,299	87	3,518
09-13	34,837	42	4,025
09-18	33,031	378	5,495
09-25	34,202	0	4,702

The largest non-observed share occurs on 26 August (16.30%); the smallest occurs on 29 August (9.27%). On 25 September, none of the 4,702 gaps has three preceding observed inputs. These patterns do not identify the causes of gaps in original reporting. Table 6 reports structural checks and remaining metadata gaps.

Table 6. Structural audit outcomes

Check	Finding
Duplicate composite keys	0
Invalid or inconsistent times	0
Invalid observed counts	0
Fractional new estimates	435
Rows with missing latitude	120
Rows with missing `LAC`	24,240
Zero count records	2,024

These tests and others have strong internal consistency but no validation of population accuracy. Pinpoint the one site where one latitude is missing and its omission should be considered (or explicitly accounted for) in the spatial analyses.

4.2 Temporal Variation and Usable Observations

At each site-day, both Idea and Uninor have equal amounts across all adjacent observed pairs. Table 7 lists sites that lacked counts of observed labels. In contrast, the number of adjacent observed pairs with no change in Airtel is 99.15 percent (Figure 4). Stable reporting versus copying of values and/or preparation effects cannot be distinguished by the panel. It should be noted that `OBSERVED` labels on their own are not an indication of the freshness of the measurement.

For all dates, a total of 103 operator sites have no observed labeled value. They have not yet been observed, and so their counts are still missing in the reconstruction rule. Do not have a measured truth for reconstruction evaluation.

Table 7. Operator sites without any observed count

Operator	No observed history
Aircel	4
Airtel	0
BSNL	0
Idea	0
Reliance	38
Tata	0
Uninor	0
Vodafone	61
Total	103

4.2.1 Outliers in Original Count Units

The count distributions and retained extremes are plotted on a symlog-scale axis (see Figure 5), and a summary of the maxima and outlier flags is given in Table 8. The number of counters stays the same.

There are a total of 169,932 operator-labeled observed records, of which 12,404 are outside the operator fence. The largest new estimate is 2,544.33, and the observed maximum is 61,058. This plot and flag tally excludes estimates. Symlog

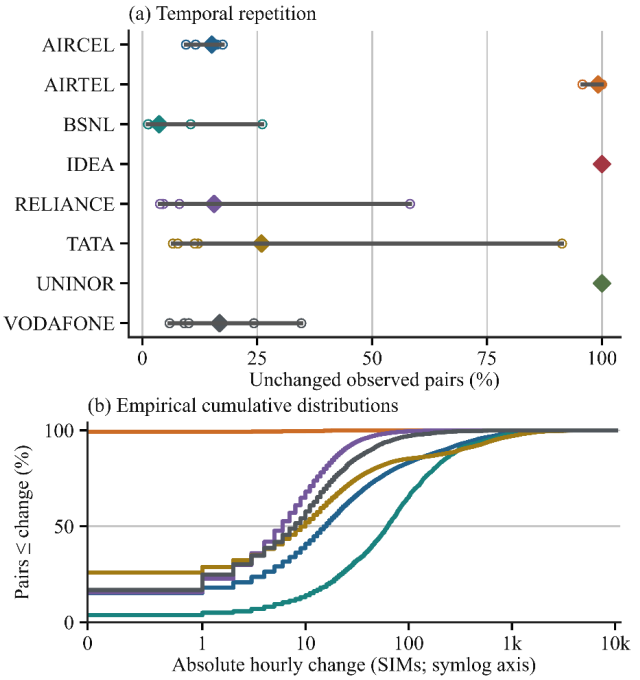

Figure 4. (a) Daily circles and pooled diamonds show temporal repetition. (b) Cumulative absolute hourly changes for 161,994 observed-labeled adjacent pairs. Colors match operators in (a); constant-series curves overlap at 100%. The change axis is symlog.

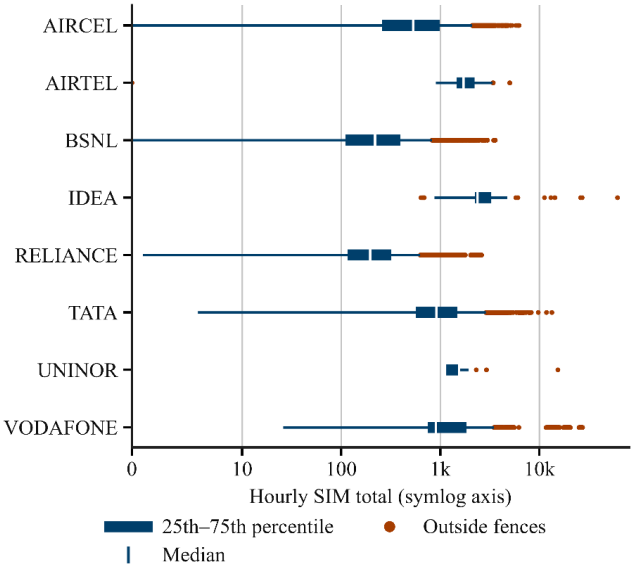

Figure 5. Observed-count distributions on a symlog axis, with interquartile intervals, medians, 1.5-IQR whiskers, and retained outliers. Repeated outliers overlap. Exact maxima and flag counts are reported separately in Table 8.

display separates small counts from the expanse of the long upper tail, without altering the stored counts.

High-count flagged may be due to a busy speed camera location, a special event takeoff, or a reporting problem. There is no way to determine which by the rule. There are repeated flags, counts repeat over hours, so it is a record and not certainly an occasion or a person. Operator-wide fences also combine sites with different typical loads.

Extreme retention is important for advancing future peak-load and service-planning studies. A value should not be

Table 8. Observed count extremes and outlier flags

Operator	Maximum	Flagged	Flag %
Aircel	6,223	565	5.32
Airtel	5,010	2,307	8.90
BSNL	3,582	292	7.63
Idea	61,058	3,720	14.69
Reliance	2,606	1,535	4.04
Tata	13,335	1,103	4.84
Uninor	15,312	720	3.82
Vodafone	27,136	2,162	8.79

corrected without validation, which is checking the data from the source report with the site context. Robustness analyses can be run on all record results, flagged subsets, and all records, without disturbing the original panel.

4.3 Chronological Reference Forecasting

The pooled MAE is lowest for Persistence at each test date (Figure 6). The outcome is only descriptive evidence: here, about these reference methods, and these types of records. All the MAE values are given in Table 9. In the lower panel, signed biases indicate the difference between underprediction and overprediction.

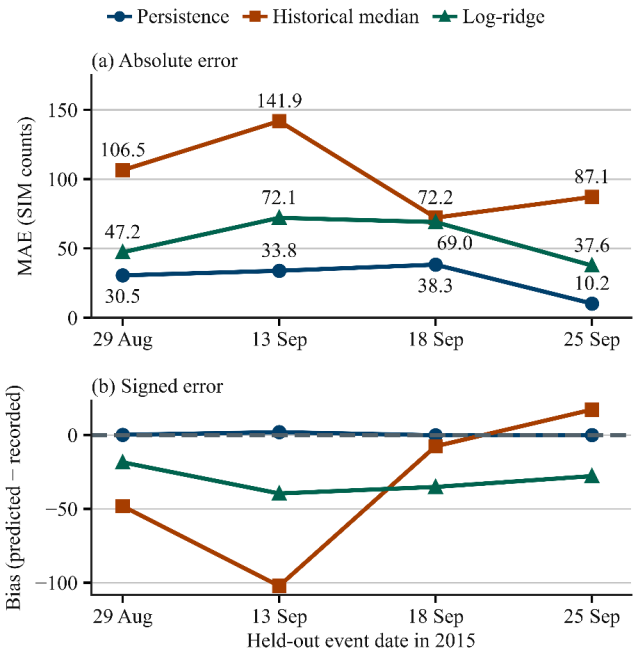

Figure 6. (a) Mean absolute error and (b) signed bias on 125,169 observed-labeled targets across four chronological folds. Negative bias indicates underprediction. Models refit on earlier dates; equal spacing denotes folds, not elapsed days.

Table 9. Mean absolute error in count units

Date	Persist.	Hist. med.	Ridge
08-29	30.54	106.50	47.24
09-13	33.85	141.86	72.07
09-18	38.27	72.19	68.99
09-25	10.20	87.13	37.65

The range of persistence MAE is from 10.20 on 25th September to 38.27 on 18th September, while the historical median spans from 72.19 to 141.86. A change in the error level does not necessarily indicate a change in the actual intensity of crowding or the quality of reporting in the absence of independent evidence. Very good persistence result – consider this temporal stability. This is tested in the comparison below, where the selection of groups is based on the history value rather than the value in the target field.

4.3.1 Large Errors and Model Objectives

Figure 7 compares RMSE and its ratio to MAE across test dates; Table 10 reports the RMSE scores.

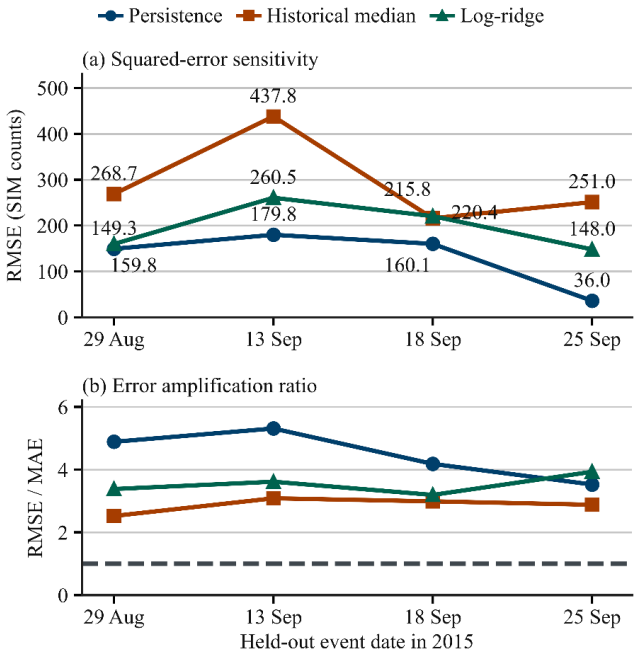

Figure 7. (a) Root mean square error and (b) root-mean-square-error-to-mean-absolute-error ratios on the same targets as Figure 6. Ratios describe unequal error magnitudes, not statistical uncertainty.

Table 10. Root mean square error in count units

Date	Persist.	Hist. med.	Ridge
08-29	149.29	268.74	159.78
09-13	179.79	437.82	260.54
09-18	160.08	215.84	220.39
09-25	35.96	251.02	148.01

The RMSE is much higher than the MAE for some method-date combinations, due to a few large errors that inflate the squared loss. Both measures then should be retained for comparison, rather than choosing one good measure, in the reference comparison. The pooled ridge model is one of the deliberately simple models, with their objective being in log space. It can compromise accuracy relative to the original scale, especially at high counts. It does not mean that all regression models or learned predictors are inappropriate, because poor pooled performance does not necessarily imply poor individual model performance. Future work involves operator-specific models, various losses, and wider cross-comparisons.

4.4 Operator Differences and Benchmark Coverage

Figure 8 shows the operator-level forecasting errors masked by the pooled ranking, and the actual MAEs for each model are shown in Table 11.

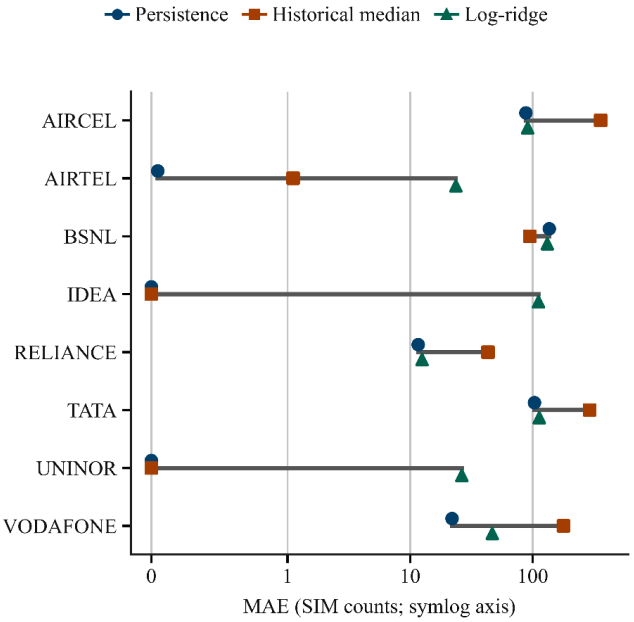

Figure 8. Operator-level pooled mean absolute error, weighted by eligible record counts across four test dates. The symlog axis resolves near-zero and large errors; vertical offsets separate coincident scores. Exact values are reported separately in Table 11.

Table 11. Operator-level pooled mean absolute error in count units

Operator	Persistence	Hist. median	Log-ridge
Aircel	87.97	358.93	90.86
Airtel	0.05	1.11	23.57
BSNL	136.25	94.83	131.44
Idea	0.00	0.00	111.27
Reliance	11.63	43.29	12.48
Tata	103.34	292.11	113.22
Uninor	0.00	0.00	26.37
Vodafone	21.99	178.17	46.78

Hiding behind the pooled rankings is an important exception. The BSNL historical median MAE is 94.83, whereas the persistence MAE is 136.25, and the ridge is 131.44. Even the 1,927 eligible targets are limited across the folds, as BSNL is also limited to only a limited set of observed points. Eligible Idea and Uninor targets have zero MAE for persistence and historical medians, while Airtel persistence MAE is approximately 0.047. These scores are not yet ideal for crowd forecasting. Table 12 is a table without target information, in which targets are divided based on whether the last two observed counts have been equal (unchanged history) or unequal (changing history).

Unchanged histories cover 58,109 targets (46.42%); changing histories cover 67,060 (53.58%). Pooled persistence MAE rises from 5.91 to 47.40, an 8.03-fold difference. Persistence remains best in both groups, while ridge MAE is

Table 12. Mean absolute error by preceding count history

History	Persistence	Hist. median	Log-ridge
Unchanged	5.91	24.01	56.71
Changing	47.40	170.29	56.03

similar across groups. The association is descriptive and mixes operators and count scales. Equal preceding counts need not imply an unchanged target or a constant full-day series.

4.4.1 High-Count Performance

Figure 9 compares high-count forecasting errors. Table 13 reports the training-derived thresholds, subset sizes, and shares of eligible test targets.

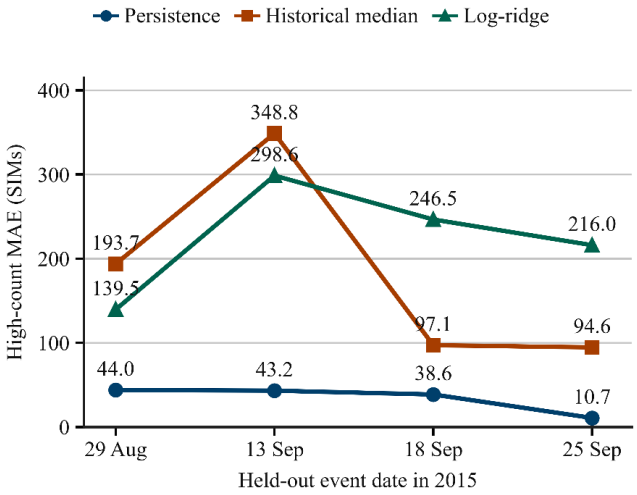

Figure 9. High-count mean absolute error across chronological test dates. The 13,250 selected targets represent statistical high-count conditions, not physical safety limits. Training thresholds, target counts, and test shares are reported separately in Table 13.

Table 13. High-count subset definition

Test date	Threshold	Targets	Test share (%)
08-29	2956.0	3,341	10.4
09-13	2956.0	3,547	11.2
09-18	3006.2	3,327	11.1
09-25	3024.0	3,035	9.7

The highest MAE in the persistence is 43.99, and the lowest is 10.74. The historical baseline is 348.75 on 13 September, and the ridge is 298.65. The results of these observations will make it easier to report demanding count regimes separately; however, the global threshold might still apply to operators using different count scales. The comparison is based only on the history of the observations, not on the original missing values. There is no assumption of statistical independence among site-hour rows, and four test dates are not sufficient to draw any conclusions about future Kumbh editions. The empirical output provides a transparent baseline that describes the operating conditions.

5 Discussion and Reuse

5.1 Implications for Dataset Release

Reuse demands an operator-site key, three output states, separate daily sequences, source checksums, and executable processing rules. Archived input fields differentiate between input (supplied) estimates and the current output estimates. While testing reconstruction against known targets is possible by synthetic masking, the causes of historical omissions cannot be determined via this method. An individual's mobility trace can still be identified [20]. This panel provides counts only and does not represent any subscriber trajectory. It does not mean it is formally done for disclosure-risk assessment.

5.2 Implications for Kumbh and Other Events

The panel can perform eventuality tests and suggest information-gathering for Kumbh 2027 and other occasions. All estimates of population, route capacity, or demand for service need to be updated and be independent measurements of coverage. Transferable audit practices do not set the event's prediction accuracy. Therefore, provider-level coverage, the freshness of coverage-based measurements, and straightforward forecasts based on historical trends should be consistently maintained in future analyses.

5.3 Limitations

There are just five dates available. They include unknown series, source attachments, deduplication by hour, and reporting timezone. The Python pipeline is created and run, but not re-created from the previously manually created pipeline. The requirement regarding the history means that most gaps in its history are not addressed, and new estimates are not independently ground-truthed. The observed-only evaluation is performed without incorporating poorly reported periods. Combining totals also does not allow for the tracking of individual journeys. Availability of Data: The authors can make the data available on reasonable request.

6 Conclusion

The resulting 194,520-record panel represents 169,932 observed counts and eight operators with 1,621 operator-site pairs, all of which are excluded from the unresolved gaps of 23,873 and the estimated 715 new operators. Contribution comprises executable preprocessing, quality findings, and reference benchmarks. The audit highlights mixed coverage, 103 sites with no observations or counts. Chronological and history-stratified results are presented to explain why observation status and temporal stability are essential, in addition to scoring, for forecasting.

Declarations

Acknowledgements. The authors gratefully acknowledge Mr. Eknath Rajaram Dawale, IAS (then Divisional Commissioner, Nashik), Mr. Deependra Singh Kushwah, IAS (then Collector, Nashik), and Mr. S. Jagannathan, IPS (then Commissioner of Police, Nashik), for their leadership and support for the joint Kumbh 2015 project. The authors also gratefully acknowledge the Telecom Regulatory Authority of India (TRAI), Nashik Municipal Corporation, Dr. Praveen Gedam, and Dr. Ramesh Raskar of the Massachusetts Institute of Technology (MIT), USA, through Kumbhathon, for their support

and contributions to the project, which provided the foundation for this research. The authors further thank the telecom operators for providing the tower-level subscriber identity module count reports.

Funding. This research received no external funding.

Author Contributions. All authors contributed to the development of the study and reviewed and approved the final manuscript.

Conflict of Interest / Competing Interests. The authors declare no competing interests.

Data Availability. The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code Availability. The audit and baseline scripts used for the reported analyses form part of the reproduction outputs and are available from the corresponding author upon reasonable request.

Ethics Approval. The study analysed aggregated telecom count data and did not involve individual subscriber trajectories. Ethics approval was therefore not required, subject to confirmation of the applicable institutional requirements.

Informed Consent. Not applicable.

Consent for Publication. Not applicable.

Generative AI and AI-Assisted Technologies. The authors declare that no generative artificial intelligence or AI-assisted technologies were used to generate the scientific content, results, or conclusions of this manuscript.

Additional Information / Supplementary Material. No supplementary material accompanies this article.

References

[1] D. Helbing, A. Johansson, and H. Z. Al-Abideen, "Dynamics of crowd disasters: An empirical study," *Phys. Rev. E*, vol. 75, no. 4, Art. no. 046109, 2007, doi: 10.1103/PhysRevE.75.046109.

[2] P. Deville *et al.*, "Dynamic population mapping using mobile phone data," *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 45, pp. 15888–15893, 2014, doi: 10.1073/pnas.1408439111.

[3] F. Ricciato, P. Widhalm, F. Pantisano, and M. Craglia, "Beyond the single-operator, CDR-only paradigm: An interoperable framework for mobile phone network data analyses and population density estimation," *Pervasive Mobile Comput.*, vol. 35, pp. 65–82, 2017, doi: 10.1016/j.pmcj.2016.04.009.

[4] M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, Art. no. 160018, 2016, doi: 10.1038/sdata.2016.18.

[5] T. Gebru *et al.*, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[6] V. D. Blondel *et al.*, "Data for development: The D4D challenge on mobile phone data," arXiv preprint arXiv:1210.0137, 2012, doi: 10.48550/arXiv.1210.0137.

[7] G. Barlacchi *et al.*, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Sci. Data*, vol. 2, Art. no. 150055, 2015, doi: 10.1038/sdata.2015.55.

[8] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, "D4D-Senegal: The second mobile phone data for development challenge," arXiv preprint arXiv:1407.4885, 2014, doi: 10.48550/arXiv.1407.4885.

[9] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The NetMob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography," arXiv preprint arXiv:2305.06933, 2023, doi: 10.48550/arXiv.2305.06933.

[10] J. Lloret, J. Tomas, A. Canovas, and L. Parra, "An integrated IoT architecture for smart metering," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 50–57, 2016, doi: 10.1109/MCOM.2016.1600647CM.

[11] G. F. Scaranti, L. F. Carvalho, S. Barbon Junior, J. Lloret, and M. L. Proença Jr., "Unsupervised online anomaly detection in software defined network environments," *Expert Syst. Appl.*, vol. 191, Art. no. 116225, 2022, doi: 10.1016/j.eswa.2021.116225.

[12] A. Cini, I. Marisca, and C. Alippi, "Filling the G_ap_s: Multivariate time series imputation by graph neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=kOu3-S3wJ7

[13] H. Wickham, "Tidy data," *J. Stat. Softw.*, vol. 59, no. 10, pp. 1–23, 2014, doi: 10.18637/jss.v059.i10.

[14] NIST/SEMATECH, "Box plot," in *e-Handbook of Statistical Methods*. Accessed: Sep. 16, 2026. [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm

[15] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: An analysis and review," *Int. J. Forecast.*, vol. 16, no. 4, pp. 437–450, 2000, doi: 10.1016/S0169-2070(00)00065-0.

[16] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.

[17] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006, doi: 10.1016/j.ijforecast.2006.03.001.

[18] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten simple rules for reproducible computational research," *PLoS Comput. Biol.*, vol. 9, no. 10, Art. no. e1003285, 2013, doi: 10.1371/journal.pcbi.1003285.

[19] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[20] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, Art. no. 1376, 2013, doi: 10.1038/srep01376.