I ISSUE BRIEF

Integrating non-conventional data sources for evidence-based policymaking and better governance in India

SPRF.IN

01
26

Pankaj Chowdhury

JANUARY 2026

Issue Brief

Integrating non-conventional data sources for evidence-based policymaking and better governance in India

II INTRODUCTION

Data is one of the fundamental pillars of evidence-based decision-making, fostering trust through transparency, which ultimately leads to better governance in a country. While developed economies have already established a robust national statistical ecosystem to ensure a steady flow of data within their countries, developing nations still struggle to generate even the most basic statistical information (Malik et al., 2025). Furthermore, even when data is available, these nations often fail to produce relevant statistical insights, primarily due to low data quality and insufficient computing capabilities (Nilashi et al., 2023).

India, a country with a population of around 1.46 billion, must establish an efficient national statistical infrastructure capable of providing timely, accurate, and comprehensive statistical data to track the country's progress on sustainable development and forecast future trajectories as part of the Viksit Bharat @2047 mission. Over the last couple of decades, India has made substantial progress not only in data gathering but also in utilising it to generate relevant insights, primarily due to major initiatives taken by the Ministry of Statistics and Programme Implementation (MoSPI), various ministries and research institutions.

However, India still faces significant challenges in six critical attributes of quality data, including accuracy, completeness, consistency, timeliness, validity, and uniqueness, which threaten to undermine the country's ambitions of implementing evidence-based policies to achieve good governance and attain the Sustainable Development Goals (SDGs) (Allen et al., 2021; Bachmann et al., 2022; Garg & Ghosh, 2025).

Table 1: Six attributes of 'Quality' data

S.No.	Attribute	It evaluates -
1	Accuracy	How correct the data is compared to a "ground truth"
2	Completeness	Whether all necessary data records and values are present, with no gaps.
3	Consistency	How uniform the data is across different systems, datasets, and formats, ensuring no contradictions.
4	Timeliness	How up-to-date the data is and whether it is available when needed
5	Validity	Whether the data conforms to the defined syntax, rules, and formats.
6	Uniqueness	The absence of duplicate records in a dataset, ensuring each entry represents a distinct entity.

Historically, India has always relied on traditional sources of data, including censuses, government-sponsored surveys and, to some extent, administrative statistical systems for decision-making and policy-making. While data sourced from the Census or Survey is excellent in terms of granularity, it often lacks timeliness due to the high costs and time-intensive nature of the data collection process. While the administrative data addresses issues related to the timeliness of the data flow to a certain degree as compared to the other two traditional sources, it lacks granularity and comprehensiveness, which constrain its effective use in policy formulation and real-time decision-making. Furthermore, the statistical ecosystem of India also suffers from issues related to the completeness and accuracy of data, which raises severe concerns about assessing ground realities and plugging data gaps (Bailey & Parsheera, 2018; Ministry of Statistics and Programme Implementation, 2025).

According to a report published in 2024 by IndiaSpend, until December 30, 2024, sixteen critical government datasets were delayed in India, and nine union ministries had failed to release their annual reports across multiple sectors, including health, environment, demography, agriculture, and criminal justice (Salve, 2024). For example, India's decennial census, originally scheduled for 2021, is now expected to be conducted in 2027 after being postponed due to the COVID-19 pandemic. Similarly, large-scale national surveys such as the National Family Health Survey (NFHS) and various MoSPI-led surveys, which serve as key instruments for monitoring emerging economic, health, and social trends in India, are conducted at intervals of three to five years. This high periodicity in India's data release cycle poses significant challenges for policymakers, as it restricts their ability to design and implement timely, evidence-based interventions in India's increasingly dynamic socio-economic environment (Savithri et al., 2022).

Beyond timeliness, the availability of relevant and accurate data is also crucial for evidence-based policy formulation and informed policy choices. It is undeniable that India's national official statistical system has made commendable progress over the years in collecting and disseminating data on a wide range of development indicators (Anand et al., 2022). However, it is still essential to analyse whether the available data is sufficient for understanding the rapidly emerging issues in India. For example, while monitoring Target 5.b of SDG 5, which aims to advocate for the use of enabling technology, particularly information and communication technology (ICTs), to promote women's empowerment, NITI Aayog relies solely on the indicator "Total Telephone Subscriptions (in millions)". This approach raises serious concerns about the relevance of the indicator, as it fails to account for gender-disaggregated information (i.e., total subscriptions by women) and overlooks other critical aspects of digital empowerment, such as internet access, digital literacy, and technology readiness (Savithri et al., 2022). Moreover, there have been several instances reported where underreporting and political sensitivities have contributed to the delay or withholding of reports that present unfavourable statistics for the ruling party (Paliwal, 2019).

As we are moving towards the next phase of our public welfare infrastructure revolution in India, data quality is no longer a peripheral technical consideration but a fundamental prerequisite for good governance and sustained public trust. When a single erroneous digit can delay a benefit, or a duplicate record inflates welfare outlays, the true cost of poor data becomes painfully apparent – impacting budgets, distorting policy, and eroding the faith citizens place in the system. The move from "scale to precision" is not just an aspiration; it is a national imperative. The aforementioned issues highlight the need for greater transparency and data integrity in the country (Garg & Ghosh, 2025; Kumar, 2025). Once systems learn to live with error, they stop trying to prevent it, and this results in labelling a database that's 80% accurate as "good enough." A mismatch between two registries is not viewed as a hindrance to service delivery, but rather as a minor technical glitch. Ultimately, quality gradually and subtly becomes optional. As an example, A large northern state declared itself open-defecation-free in 2019; however, an audit of 590 rural houses found that nearly half lacked toilets. Dashboards still showed 100 per cent coverage, and funds continued to move, indicating that data deemed "close enough" can override field reality (Patel, 2019).

The current government, under the leadership of Hon'ble Prime Minister (PM) Shri Narendra Modi is committed to the ideals of "Sabka Saath, Sabka Vikas, Sabka Vishwas, Sabka Prayaas" to ensure that basic necessities and government-sponsored welfare schemes such as Ayushman Bharat Yojana, PM Awas Yojana, Jal Jeevan Mission, PM Jan Dhan Yojana, PM Kisan Samman Nidhi or PM Ujjwala Yojana must reach all the citizens of the country, especially those who are living at remote areas or belong to poor and marginalised sections of the society. However, local governments often face challenges in efficiently collecting, maintaining, and ensuring the quality of the data due to limited infrastructure and technical capacity. Furthermore, the data stream in India typically follows a one-way flow from lower to higher administrative levels (Jeevanandam, 2023). This means that local bodies responsible for generating data often lack access to the same information when it comes to implementing data-driven policies at the grassroots level. Consequently, it becomes difficult for policymakers to assess the impact of local development initiatives, support data-driven decision-making, and analyse spatial patterns of inequality with accuracy and timeliness. In some cases, even if the data is available, it lacks granularity.

Over the past decade, India has emerged as a global leader in digital public infrastructure. At the end of the financial year (FY) 2025, the Unified Payments Interface (UPI) processed 17.89 billion transactions worth ₹ 23.9 trillion, rivalling the monthly GDP of several mid-sized economies. Aadhaar authenticated 27.07 billion identity requests in FY 2024-25, confirming its place as the common key for banking, welfare and a widening array of private-sector services. On the social-protection front, more than 369 million Ayushman cards are now in circulation, bringing cashless hospital care to roughly half a billion citizens. These are not merely statistics; they are testaments to a foundational transformation, enabling financial inclusion, empowering citizens, and streamlining welfare delivery on a scale previously unimaginable. As a whole, India has already proved it can scale data, but now the actual test is how precisely, timely and granularly India's statistical infrastructure can produce data. Statistical insights can create public value only if every record is accurate, complete, and current (Garg & Ghosh, 2025).

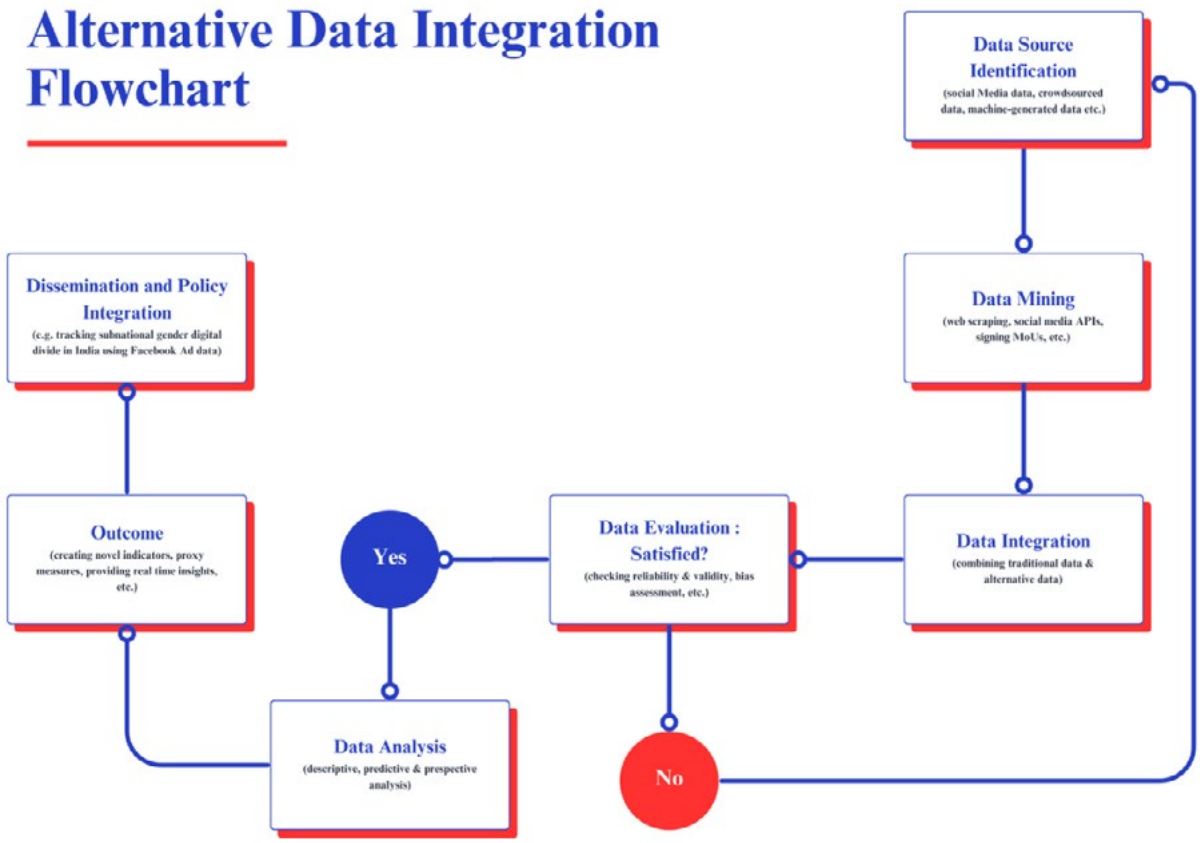
India has witnessed a sharp upward trajectory of digital expansion and adoption in the last decade. At the end of FY 2025, the total internet users in India is estimated to be around 970 million, which translates to almost 69 users per 100 people. At the same time, India's total internet consumption is expected to reach approximately 2.3 lakh petabytes of data through wireless networks, with an average consumption per subscriber of around 86 GB per month. Additionally, more than 95% of Indian villages are reported to have access to the internet with 3G/4G mobile connectivity. Furthermore, over the past decade, India has emerged as a global leader in digital public infrastructure. This unprecedented growth has been driven by the government's Digital India initiative, which aims to extend internet connectivity to even the most remote areas of the country, the widespread availability of affordable data plans and smartphones, as well as the rise of OTT platforms and social media usage in rural areas (Ministry of Communications, 2025).

This rapid digital transformation has created significant opportunities for policymakers and researchers to leverage non-traditional data sources for developing new statistical products that can complement or even substitute existing statistical indicators, thereby enhancing the granularity, timeliness, and relevance of official statistics. Non-conventional data refers to statistical information derived from diverse and often unstructured sources, such as social media platforms, sensors, and other digital channels, as opposed to traditional data sources like census records, structured surveys, and observational studies. Some of the popular non-conventional data sources that can be leveraged to drive innovation in India's official statistics include administrative data sources,

digital traces, sensor data, big data, citizen-generated data, crowdsourced data, nightlight data, satellite data, and more (Allen et al., 2021; Chandrasekaran et al., 2025).

Overall, in the current era of tech-driven modernisation and increasing digital adoption, non-conventional data presents vast potential to produce novel insights about products, services, and customer behaviours, which can directly contribute to the development of an advanced data-driven governance framework and enhance the quality of service delivery in India. A well-designed integration of non-conventional data sources with traditional data sources can provide better comprehension on emerging issues in India, leading to an improved healthcare system, strong supply chain management and logistics, convenient travel, smart farming, and a transparent FinTech ecosystem (Ministry of Statistics and Programme Implementation, 2025).

Figure 1: Framework for integrating alternative data to strengthen India's national statistical ecosystem



Non-conventional data sources offer several distinct advantages over traditional datasets. They enable real-time statistical insights, which support proactive policy formulation and the development of early warning systems. Data generated through digital footprints can serve as proxies for ground truth measures, enabling the cross-validation of official statistics and enhancing data accuracy (Allen et al., 2021). In many cases, these alternative data sources also help strengthen the granularity of the data and broaden its dissemination. Furthermore, these types of data sources can be utilised to create innovative statistical products that track development indicators more effectively than those derived from conventional sources (Weber et al., 2023). Finally, non-conventional datasets also reduce the cost of data collection, as they are typically by-products of self-generated digital activities rather than costly field surveys. Moreover, the advent

of advanced statistical techniques such as machine learning models, natural language processing (NLP), and artificial intelligence (AI) has substantially increased the potential of these alternative data sources to generate novel insights that traditional data systems often fail to capture (Abreu Lopes & Bailur, 2018).

Table 1: Case studies on how alternative data sources can be used to complement or replace traditional data systems

Traditional Data Source / Method	Alternative Data Source	Type of Use	Practical Example	Value Added
Household Surveys for Poverty Estimation	Satellite imagery + night-time lights + geospatial data	Complement (enhances granularity)	Philippines & Thailand used publicly available satellite images + ML models to produce small-area poverty estimates	Faster, cheaper updates; high-resolution poverty maps; improved targeting of social programmes
Price Collection by Field Enumerators	Web scraping, APIs, scanner data, POS data	Replacement (partial)	Canada collects 50% of CPI prices using scanner + online data; Brazil replaced manual airfare price collection with web scraping	Reduces burden, increases frequency, improves coverage, lowers cost
Census / Administrative Records for Land Use Statistics	Sentinel-2 satellite imagery + automated deep learning	Complement / Replacement for periodic updates	ISTAT uses public Sentinel-2 imagery to generate land-cover statistics	Provides near real-time land-use updates; supports SDG 11 and environmental monitoring
Travel & Tourism Surveys	Mobile phone positioning data	Replacement	Estonia has produced monthly international travel statistics for 14 years using mobile positioning data	Real-time tourism statistics; cost-effective; high volume and accuracy
Surveys for Consumer Confidence Index	Social media data (Twitter, blogs)	Replacement	Several NSOs (e.g., Netherlands) tested deriving consumer sentiment from social media for timely indexes	Near real-time sentiment tracking; broader population signal
Household Expenditure Surveys	Bank transactions, supermarket transactions, vehicle sales data	Complement	Australia's Monthly Household Spending Indicator combines multiple transactional datasets	Timely tracking of consumer spending; reduces survey frequency; high precision
Traffic Counts via Manual Observations	Road sensors, CCTV/video analytics	Replacement	Traffic statistics generated using camera data	Continuous estimation of road congestion; supports urban mobility planning

Traditional Data Source / Method	Alternative Data Source	Type of Use	Practical Example	Value Added
Agricultural Surveys	Remote sensing (EO data), rainfall sensors	Complement	Satellite-based crop monitoring used by many NSOs	Timely agricultural forecasts; climate resilience insights
Administrative Records on Electricity Access	Night-time lights (NTL) data from VIIRS / DMSP-OLS	Complement	World Bank's Light Every Night portal uses satellite images to track electricity access	Consistent, comparable electricity access indicators, especially where admin data is weak
Gender Statistics from Surveys	Social media data, mobile phone usage, radio content analytics	Complement	UN Global Pulse analysed radio content in Uganda to detect gender biases and women's issues	Real-time sentiment analysis; reveals "invisible" gender issues
On-Ground Disaster Assessment	Satellite imagery, drones, crowdsourced reports	Replacement (rapid assessments)	Satellite night-lights used to assess post-disaster impacts quickly	Faster emergency response; granular assessment
Infrastructure Access Surveys	GIS + Open geospatial data	Complement	Colombia used geospatial datasets to report SDG indicators like public transport access and public space	High-resolution spatial indicators; supports SDG 11.2, 11.3, 11.7

The larger picture of incorporating alternative data into the mainstream national statistical system is a complex process that comes with multiple challenges and limitations. First, it is pivotal to identify specific use cases where alternative data can add value, followed by a thorough assessment of their feasibility and applicability. Second, the collection, processing, and analysis of non-conventional data demand advanced technical expertise and computational capabilities. Moreover, since these data sources often emerge as by-products of digital interactions, they may suffer from coverage gaps and representational bias. It is therefore essential to validate whether proxy indicators derived from such data accurately reflect the offline population. As a result, in the Indian context, non-conventional data should be viewed as a complement to traditional data systems rather than a substitute. Third, obtaining informed consent from users poses significant ethical and legal challenges, particularly when dealing with large-scale and complex datasets, and raises critical concerns regarding data privacy, the protection of sensitive information, and compliance with regulatory frameworks (Kashyap et al., 2023).

Considering the various restrictions and difficulties associated with combining non-traditional data with traditional data sources, India must concentrate on successfully resolving these issues to enhance the calibre, thoroughness, and timeliness of official statistics in the country. To deal with this transition, a phased and strategic approach is necessary, which involves breaking down procedures into manageable, actionable steps and encouraging ongoing experimentation and methodological innovation to identify more effective frameworks for integrating novel data sources. Furthermore, India must focus on strengthening its technical readiness, ensuring the capability

to handle a massive expansion from managing millions of records each month to processing millions of datasets across diverse sources. This requires substantial investment in data storage and processing infrastructure, alongside robust data security mechanisms to safeguard sensitive information. Given that online platforms serve as major sources of non-conventional data, establishing reliable data pipelines becomes essential (Chandrasekaran et al., 2025). This can be achieved through partnerships and Memoranda of Understanding (MoUs) with e-commerce platforms, industry associations, and other stakeholders supported by legal and regulatory frameworks that enable timely and standardised data sharing. Moreover, India should set up a dedicated multidisciplinary task force comprising Statisticians, Economists, Data Scientists, and IT Specialists to ensure the scientifically sound and contextually relevant integration of alternative data sources into the official statistical system. This coordinated effort will be crucial in transforming India's data ecosystem into a more agile, data-driven foundation for evidence-based policymaking (Ministry of Statistics and Programme Implementation, 2025).

Declaration of Generative AI and AI-assisted technologies in the writing process:
During the preparation of this work, I have partially used ChatGPT for paraphrasing some sentences. After using this tool/service, I reviewed and edited the content as needed, and I take full responsibility for the content of the publication.

REFERENCES

Abreu Lopes, Claudia., & Bailur, Savita. (2018). Gender equality and big data : making gender data visible. UN-Women : Global Pulse.

Allen, C., Cameron, G., & Dahmm, H. (2021). Big Data and the Sustainable Development Goals: Innovations and Partnerships to Support National Monitoring and Reporting.

Anand, S., Lewis, V., & Gulabani, H. (2022). The need for real time and granular data to study the urban economy.

Bachmann, N., Tripathi, S., Brunner, M., & Jodlbauer, H. (2022). The Contribution of Data-Driven Technologies in Achieving the Sustainable Development Goals. Sustainability 2022, Vol. 14, Page 2497, 14(5), 2497. https://doi.org/10.3390/SU14052497

Bailey, R., & Parsheera, S. (2018). Data Localisation in India: Questioning the Means and Ends. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3356617

Chandrasekaran, N., Shah, A., Banerjee, C., Mishra, N., Madgavkar, A., Bansal, M., Matthan, R., Kaka, N., Vaishnav, C., & Ahmed, I. (2025). AI for Viksit Bharat : The opportunity for Accelerated Economic Growth.

Garg, S., & Ghosh, D. (2025). India's Data Imperative: The Pivot Towards Quality 1. https://www.niti.gov.in/sites/default/files/2025-06/FTH-Quaterly-Insight-june.pdf

Jeevanandam, N. (2023, July 24). Indian local government datasets to solve regional issues. https://indiaai.gov.in/article/indian-local-government-datasets-to-solve-regional-issues

Kashyap, R., Gordon Rinderknecht, R., Akbaritabar, A., Alburez-Gutierrez, D., Gil-Clavel, S., Grow, A., Kim, J., R. Leasure, D., Lohmann, S., V. Negraia, D., Perrotta, D., Rampazzo, F., Tsai, C.-J., D. Verhagen, M., Zagheni, E., & Zhao, X. (2023). Digital and computational demography. In Research Handbook on Digital Sociology (pp. 48–86). Edward Elgar Publishing. https://doi.org/10.4337/9781789906769.00010

Kumar, S. (2025, January 6). Enabling Data Flow for Effective Governance in India | The India Forum. https://www.theindiaforum.in/public-policy/enabling-data-flow-effective-governance-india

Malik, S. C., Gupta, R., Kanaujia, A., Sakshi, S., Gaur, A., Pal, R. K., Meena, J. K., Rathor, G., Meena, R., Sharma, P., Anju, Umar, M., & Attri, H. (2025). Sustainable Development Goals, National Indicator Framework, Progress Report 2025. https://www.mospi.gov.in/sites/default/files/publication_reports/Sustainable%20Development%20Goals%20National%20Indicator%20Framework%20Progress%20Report%2C%202025.pdf

Ministry of Communications. (2025, September). Ministry of Communications Dashboard. https://dot.dashboard.nic.in/DashboardF.aspx

Ministry of Statistics and Programme Implementation. (2025). Presentation for Using Non-conventional Data Sources in Official Statistics in India to Improve Data Centricity.

Nilashi, M., Keng Boon, O., Tan, G., Lin, B., & Abumalloh, R. (2023). Critical Data Challenges in Measuring the Performance of Sustainable Development Goals: Solutions and the Role of Big-Data Analytics. Harvard Data Science Review, 5(3). https://doi.org/10.1162/99608f92.545db2cf

Paliwal, A. (2019, March 15). 108 academicians slam Modi govt over suppression of unfavourable unemployment data - India Today. India Today. https://www.indiatoday.in/india/story/108-academicians-slam-modi-govt-over-suppression-of-unfavourable-unemployment-data-1478351-2019-03-15

Patel, S. (2019, November 27). Claim versus Reality: Has India become open defecation free? – IDEAs. https://www.networkideas.org/2019/11/27/claim-versus-reality-defecation-free/

Salve, P. (2024, December 30). Critical Data Remain Elusive, As 2024 Comes To A Close. https://www.indiaspend.com/data-gaps/critical-data-remain-elusive-as-2024-comes-to-a-close-936492#google_vignette

Savithri, R., Ojha, A., Kumar, S., Kanaujia, A., Gaur, A., Prasad, A., Khanna, S., & Kumar, A. (2022).

Guidance on Monitoring Framework for SDGs at sub national level. https://mospi.gov.in/sites/default/files/publication_reports/Guidance_on_MonitoringFramework_for_SDGs_March31_2022.pdf

Weber, I., Kashyap, R., & Zagheni, E. (2023). Using advertising audience estimates to improve global development statistics. TU Journal: ICT Discoveries, 2, 3–8. https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-2-P04-PDF-E.pdf

WWW.SPRF.IN

If you have any suggestions, or would like to contribute, please write to us at **contact@sprf.in**